Improving the measurement of poverty and social exclusion in Europe: reducing non-sampling errors

EDITED BY PETER LYNN AND LARS LYBERG

2022 edition

STATISTICAL
WORKING PAPERS

eurostat

Improving the measurement of poverty and social exclusion in Europe: reducing non-sampling errors

EDITED BY PETER LYNN AND LARS LYBERG

2022 edition

Luxembourg: Publications Office of the European Union, 2022

Theme: Population and social conditions
Collection: Statistical working papers

Print	ISBN 978-92-76-43001-8	doi:10.2785/949630	KS-01-21-385-EN-C
PDF	ISBN 978-92-76-43002-5	doi:10.2785/18803	KS-01-21-385-EN-N

Please quote this book as: Lynn, P., Lyberg, L. (eds) (2022), *Improving the measurement of poverty and social exclusion in Europe: reducing non-sampling errors*, Publications Office of the European Union, Luxembourg.

Acknowledgements

This book is an output of the Third Network for the Analysis of European Union Statistics on Income and Living Conditions (Net-SILC3), a collaboration between 17 organisations that was funded by Eurostat over a 4-year period (2016–2020) to carry out a programme of research activities and to disseminate best practice to national statistical institutes. Net-SILC3 was coordinated by the Luxembourg Institute of Socio-Economic Research. A companion book, *Improving the understanding of poverty and social exclusion in Europe* (edited by Anne-Catherine Guio, Eric Marlier and Brian Nolan), is published in the same series.

The chapter authors have provided an excellent set of chapters and responded patiently and diligently to many requests from the editors for edits, improvements and additions. The quality of the book is testament to their knowledge and commitment.

Lars passed away when we were part-way through the editing process. This book is a monument to him, but a wholly inadequate one, for he was a true giant of the survey research world. Lars contributed in many ways to the advancement of survey methods and survey quality over many years. He mentored and taught many researchers, guided the activities of statistical organisations, and disseminated good practice through books and articles. He founded the *Journal of Official Statistics* and remained editor-in-chief for 25 years. He wrote (with Paul Biemer) the seminal Wiley book *Introduction to Survey Quality* and edited several monograph books on survey methods, the most recent of which, *Big Data Meets Social Science*, was published in 2020. He was a member of the European Statistical Governance Advisory Board, chair of the Scientific Advisory Board of the European Social Survey, and co-chair of the American Association for Public Opinion Research (AAPOR) / World Association for Public Opinion Research (WAPOR) Task Force on Quality in Comparative Surveys, to mention just a few such contributions. He won the American Statistical Association's Waksberg Award (2012), the WAPOR Helen Dinerman Award (2013), the AAPOR lifetime achievement award (2018), and was Swedish Statistician of the Year in 2019.

But most of all, Lars was a good friend. He encouraged me in my professional endeavours and was always generous with his wisdom, but we also shared good food and watched football together. He always had time for the important things in life. His warmth and humanity touched all who met him. He will be sorely missed, but our lives will continue to be lit up by fond memories of him.

At the time of Lars's death, Chapter 1 was the only one that remained unwritten. We had discussed in broad terms what it should cover and the main messages we would like to convey, but we had not yet written a word, although Lars had made several pages of handwritten notes. I hope that Lars would approve of the final version, although I suspect it is only a pale shadow of the chapter it would have been had Lars been able to contribute directly to the writing.

Peter Lynn (University of Essex, Colchester, United Kingdom)

Abstract

Non-sampling error can seriously influence statistical estimates based on survey data. Almost any stage of the survey process can give rise to such statistical error, from initial decisions about the concepts to be measured by the survey through to the final stages of data editing. Two aspects of the implementation of data collection are particularly important: survey participation (or its counterpart, non-response) and survey measurement (the validity and accuracy of the answers provided by respondents). Data collection modes play an important role in determining the influence of these aspects. This book attempts to map out the influence of all possible types of non-sampling error on the European Union Statistics on Income and Living Conditions (EU-SILC) data, and to identify ways in which the error could be reduced. The majority of the chapters report research that formed part of the activities of the Third Network for the Analysis of EU-SILC (Net-SILC3), although there are also some guest chapters. The many practical conclusions include suggestions for improvements to documentation of procedures, improvements to guidance on survey procedures, capacity building in methods for dealing with error sources, and methodological studies, especially cross-national studies.

Les erreurs non dues à l'échantillonnage peuvent sérieusement influencer les estimations statistiques fondées sur les données d'enquête. Presque toutes les étapes du processus d'enquête peuvent donner lieu à une telle erreur statistique, des décisions initiales concernant les concepts à mesurer par l'enquête, en passant par les étapes finales de l'édition des données. Deux aspects de la mise en œuvre de la collecte de données sont particulièrement importants: la participation à l'enquête (ou son équivalent, la non-réponse) et la mesure de l'enquête (la validité et l'exactitude des réponses fournies par les répondants). Les modes de collecte des données jouent un rôle important dans ce qui va déterminer l'influence de ces aspects. Ce livre tente de cartographier l'impact de tous les types possibles d'erreurs non dues à l'échantillonnage dans la source de données EU-SILC, et d'identifier les moyens par lesquels ces erreurs pourraient être réduites. La majorité des chapitres font état de recherches faisant partie des activités du troisième réseau d'analyse des statistiques de l'UE sur le revenu et les conditions de vie (Net-SILC3); quelques chapitres sont des chapitres invités. Les nombreuses conclusions pratiques comprennent des suggestions pour une documentation améliorée des procédures, une meilleure guidance relative aux procédures d'enquête, un renforcement des capacités en matière de méthodes de traitement des sources d'erreur et des études méthodologiques, en particulier transnationales.

Nicht-Stichprobenfehler können statistische Schätzungen basierend auf Umfragedaten erheblich beeinflussen. Fast jede Phase des Erhebungsprozesses kann zu solchen statistischen Fehlern führen, von den ersten Entscheidungen über die in der Erhebung gemessenen Konzepte bis hin zu den letzten Phasen der Datenaufbereitung. Zwei Aspekte bei der Durchführung der Datenerhebung sind besonders wichtig: Teilnahme an der Befragung (oder ihr Gegenstück, Nichtteilnahme) und Messung (Gültigkeit und Genauigkeit der von den Befragten gegebenen Antworten). Der Modus der Datenerhebung spielt eine wichtige Rolle und bestimmt den Einfluss dieser Aspekte. Dieses Buch versucht, den Einfluss aller möglichen Arten von Nicht-Stichprobenfehlern in EU-SILC aufzuzeigen und Wege zu finden, wie Fehler reduziert werden können. Die meisten Kapitel berichten über Forschungsarbeiten, die im Rahmen des dritten Netzwerks zur Analyse der EU-Statistiken zu Einkommen und Lebensbedingungen (Net-SILC3) durchgeführt wurden, während es auch einige Gastkapitel gibt. Zu den zahlreichen praktischen Schlussfolgerungen gehören Vorschläge für eine verbesserte Dokumentation von Verfahren, eine verbesserte Anleitung zu Erhebungsverfahren, Kapazitätsaufbau bei Methoden zum Umgang mit Fehlerquellen und methodische Studien, insbesondere länderübergreifende Studien.

Contents

List of figures and tables

Figures

Tables

Introduction

1 Handling non-sampling errors in comparative surveys

Lars Lyberg and Peter Lynn ([1])

1.1. Sampling and non-sampling errors

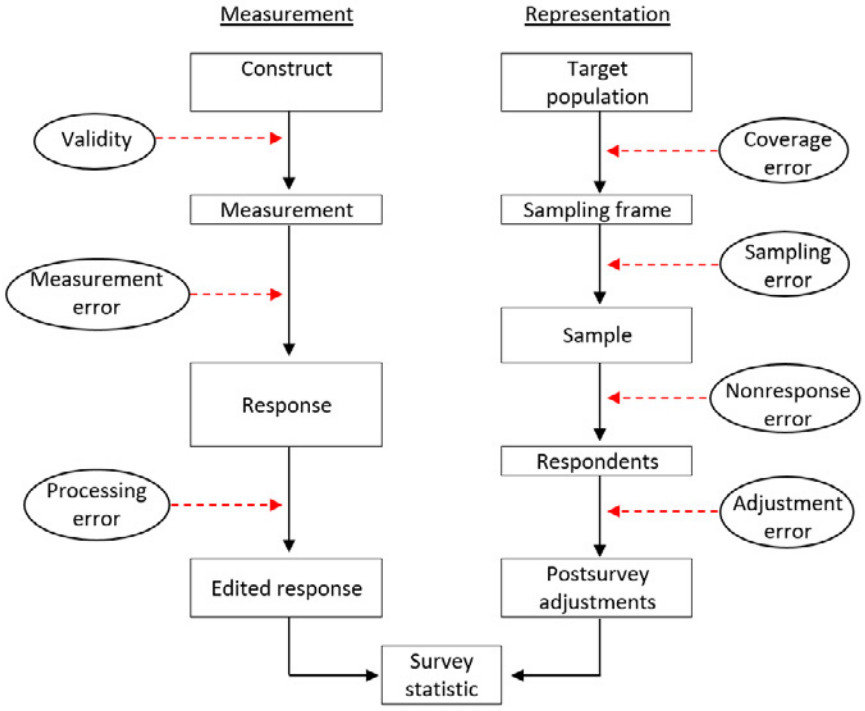
In 1895, A. N. Kjær, Chief of the Statistical Division of the Norwegian Ministry of the Interior, proposed to a meeting of the International Statistical Institute that it should be possible to obtain reasonably accurate estimates of the characteristics of the population of a country without taking a complete census (Kjær, 1897). Instead, Kjær proposed that a sample of the population could be studied. Thus, the sample survey was born. Over the following five decades, arguments raged about how a sample should be selected and whether it was reasonable or possible to make population inferences from a sample. Arthur Bowley (1926) made the case for random selection and stressed that the sampling frame should have complete coverage of the population. An influential paper by Jerzy Neyman (1934) contrasted stratified probability sampling with purposive quota sampling and introduced the concept of the confidence interval. A subsequent paper by Neyman (1938) was one of the first to jointly consider both cost functions and variance functions in determining the best sample design for a given situation. Many others contributed to the advancement of sampling theory, including, notably, Cochran (1942, 1946) , Hansen and Hurwitz (1943) and Yates (1949). Most of the core components of modern survey sampling theory were in place by the middle of the 20th century and are coherently set out in the seminal text by Hansen, Hurwitz and Madow (1953). Since then, the role of sampling variance in determining the nature of inference that can be made from sample to population has been at the core of survey design and estimation. The role of error from other sources, however, has only been recognised more recently and has taken some time to be well understood. Even now, other error sources are often ignored in the design of surveys and have very little influence on how survey data are used or how findings are interpreted.

The first typology of non-sampling errors was offered by Deming (1944). Deming's descriptions of error sources would be understood and recognised by today's survey methodologists. He discussed interviewer effects, mode effects, context effects, and measurement errors caused by questionnaire design or interview length, non-response errors and processing errors (though not always using the same terminology that would be used today). Deming also included as error sources the choice of an unrepresentative date or period for fieldwork and changes that take place in the population between the collection of data and the publication of results. In the years following the publication of Deming's typology, methods for estimating the errors arising from sources other than sampling were developed. Mahalanobis (1946) developed an interpenetrated design for the measurement of interviewer variance – a method that was subsequently extended and refined by others (Biemer

([1]) Lars Lyberg was with Demoskop, Inc., Sweden, until his death in April 2021. Peter Lynn is with the Institute for Social and Economic Research at the University of Essex, Colchester, United Kingdom. This work was supported by the Third Network for the Analysis of European Union Statistics on Income and Living Conditions (Net-SILC3), funded by Eurostat and coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The European Commission bears no responsibility for the views expressed, which are solely those of the authors. Correspondence should be addressed to Peter Lynn (plynn@essex.ac.uk).

and Stokes, 1985) – and Hansen and Hurwitz (1946) proposed follow-up sampling of non-respondents to estimate non-response error. Other typologies of non-sampling errors have since followed, including those by Kish (1965), Groves (1989) and Biemer and Lyberg (2003). Lyberg and Weisberg (2016) classified error sources into three groups: those arising from respondent selection (sampling, coverage, non-response), those relating to response accuracy (measurement errors) and those arising from survey administration. In this last category, they include mode effects, along with processing error and comparison error. All of these works stress the importance of not restricting attention to sampling error when designing a survey, but rather paying attention to the full range of features that may affect the accuracy of the survey estimates. The total survey error framework provides the means to do this (Lyberg and Stukel, 2017). The stages of the survey process at which error can be introduced are summarised in Figure 1.1. In addition, some authors (Lyberg and Stukel, 2017) have noted that inappropriate specification of the constructs to be measured can lead to problems that are sometimes referred to as specification error.

Figure 1.1: Error sources during the survey life cycle



Source: Adapted from Figure 2.5 in Groves et al. (2009).

Broadly, there are two ways in which the total survey error framework can be applied. The first is to use the framework to **evaluate** survey errors (Bailar and Dalenius, 1969), whereas the second aims to **control** the errors (Hansen and Steinberg, 1956). Both are important components of the survey manager's armoury. Evaluation provides information on the magnitude and nature of survey error components, whereas control aims to suppress the error at source by adapting the survey process. Dalenius (1967) proposed that the two approaches should be combined in what he referred to as a **total survey design**. This requires the evaluation step to go beyond mere quantification and provide an understanding of how and why the errors arise. The 1970s and 1980s saw significant progress in research into the mech-

anisms by which survey errors arise. A movement to define and understand the cognitive aspects of survey methodology set out response process models that could explain how and why measurement errors arise (Jabine et al., 1984) and could be empirically tested. For surveys of individuals, these models were further developed and described by Tourangeau, Rips and Rasinski (2000). As a result of these initiatives and related research, much is now known about the roles played by questionnaire designers, interviewers, respondents and coders and the interaction between them in shaping the bias and variance of survey estimates. Unfortunately, there remains something of a disconnect between this knowledge and survey practice, particularly within official statistics. This book aims to take some modest steps towards remedying that.

1.2. Comparative surveys

Throughout much of the history of the development of survey methodology, most surveys carried out were national or subnational. The first truly cross-national academically led surveys were the European Values Study and the World Values Survey ([2]), which were first carried out in 1981, the International Social Survey Programme (ISSP), which began in 1994, and the European Social Survey (ESS), which was first implemented in 2002. In the realm of European official statistics, the first European Union Labour Force Survey was conducted in 1960, but it was only in 1983 that it became a requirement for all EU Member States to carry out a regular Labour Force Survey with some standardised parameters. The first cross-national longitudinal survey was the European Household Panel, which was started in 1994. Cross-national student assessment surveys have been carried out since the 1960s, but it is only since 2011 that these have moved beyond the classroom to include assessment of random samples of the adult population (Kirsch and Thorn, 2019) in the form of the Programme for the International Assessment of Adult Competencies (PIAAC ([3])).

It is only since the 1990s that comparative survey methodology has grown to become a recognised subfield within the discipline of survey methodology. Comparative surveys are defined by Harkness et al. (2010a) as surveys explicitly designed to compare two or more groups. Although the aim of comparing groups could apply to almost any survey, the emphasis here is on the need to design surveys to provide comparability when the groups consist of, or contain, different cultures, countries, languages or ethnic groups. An influential study by Park and Jowell (1997) found that the data produced by the ISSP were not as comparable as researchers had thought. Although the ISSP had produced a set of methodological guidelines and standards that each participating country was meant to follow, it was discovered that adherence to the standards, and interpretation of the guidelines, varied considerably. This led to the realisation that greater control and monitoring was needed if the goals of input harmonisation were to be achieved (Jowell, 1998), and greatly influenced the design and organisation of the nascent ESS (European Science Foundation, 1998).

With the birth of the ESS and its considerable emphasis on developing cutting-edge methodology for cross-national surveys, comparative survey methodology gained a strong impetus. The Comparative Survey Design and Implementation (CSDI) Workshop was founded in 2003 through the initiative of Janet Harkness. Early meetings were held in Brussels (2003), Paris (2004), Madrid (2005) and The Hague (2006), and the attendees quickly became an active group of collaborators who set and pushed forward a research and development agenda for multinational, multiregional and multicultural context (3MC) surveys and produced a set of guidelines for carrying out cross-national surveys (Survey Research Center, 2016). The CSDI group subsequently organised a seminal conference in Berlin in 2008 (International Conference on Survey Methods in Multinational, Multiregional and Multicultural Contexts), as a result of which the field of comparative survey methodology is often referred to as 3MC. A second 3MC conference was held in Chicago in 2016. Each conference resulted in a book (Harkness et al., 2010b; Johnson et al., 2018); these volumes have served well as the most comprehensive overviews of 3MC best practice.

Achieving comparability in cross-national surveys is particularly challenging (Johnson and Braun, 2016).

([2]) https://www.worldvaluessurvey.org/
([3]) https://www.oecd.org/skills/piaac/

Lynn, Japec and Lyberg (2006) identified unique design features of cross-national surveys but concluded that the features of such surveys that most distinguished them from national or subnational surveys were the distributed nature of survey organisation and challenges in controlling the design and implementation across countries. Cross-national surveys face something of a dilemma in deciding on appropriate quality standards, as comparability between countries can be in conflict with a desire to achieve the highest possible quality in each country (Lynn, 2003). A common approach is to prescribe a minimum set of standards that must be met but allow – and indeed encourage – countries to exceed the standards.

1.3. Error sources in comparative surveys

The sources of error are, of course, the same for a comparative survey as for any other survey. Aside from sampling error, these consist of non-coverage error, non-response error, measurement error, processing error and adjustment error (Figure 1.1). Each of these sources can have multiple components. Non-response error will be partly due to non-contacts, partly due to refusals and partly due to other reasons for non-response (Lynn, 2008). Measurement error can be caused by instrument design, interviewers and respondents and the interaction between them (Biemer and Lyberg, 2003). However, in comparative surveys some error sources tend to take on greater significance than they may in a national survey, because they are particularly a threat to comparability. Specification errors are certainly in this category. It can be quite challenging to specify constructs that are equally meaningful and relevant in all countries. A major barrier to data comparability is caused by the risk that some non-sampling errors may differ systematically between countries. This risk, in turn, comes about due to differences between countries in what Lynn (2003, p. 323) refers to as:

> various constraining factors *that would perhaps be considered as 'fixed' factors in the context of a national or sub-national survey. These include:*
>
> - *the availability and coverage of sampling frames;*
> - *laws and regulations that restrict aspects of survey practice; and*
> - *the availability and abilities of survey research organisations.*

In addition, Lynn (2003, p. 323) identifies:

> other relevant factors *that are not in themselves national characteristics but which can correlate highly with nation:*
>
> - *geographical dispersal of the study population;*
> - *language(s) spoken; and*
> - *cultural and behavioural norms.*

For example, it could be that, in one country, the sampling frame tends to produce some under-coverage of people living in urban areas, whereas in another country the frame results in under-coverage of people living in rural areas. If urbanicity/rurality is associated with survey measures such as income or poverty, this will result in bias in estimates of the difference between the two countries. Similarly, the nature of social norms may result in a tendency for low-income households to over-report their income in one country, but not in another. Indeed, there is a range of evidence of differences between cultures in survey response styles (Pennell and Cibelli Hibben, 2016), a phenomenon that can lead to systematic differences between countries in measurement error. The process of translating survey questions and questionnaires into different languages can also induce systematic error differences between languages – and hence between countries in which interviews are conducted in different languages. Over the past two decades, considerable attention has been given to the development of best practice translation methods for cross-national surveys (Behr and Shishido, 2016).

The skills and practices of the survey organisation also play a role here. Interviewers may be better at probing for answers in some countries, or may make more attempts – or more effective attempts – to make contact with hard-to-contact households in some countries. Errors arising from these kinds of constraints or practices may tend to apply more or less to all sample subgroups equally in a national survey but can greatly undermine comparability in a cross-national survey. In the

presence of errors, apparent similarities and differences between countries may be just methodological artefacts. Therefore, it is necessary to prevent, detect and adjust for, *ex post*, these potential artefacts. The contribution of this volume to identifying and assessing the effect on comparability of various non-sampling error sources in a major cross-national survey – European Union Statistics on Income and Living Conditions (EU-SILC) – is to be welcomed.

1.4. Controlling error

Attempts to control error are rather more complex in the case of cross-national surveys, primarily because multiple survey agencies are involved (Lynn, Japec and Lyberg, 2006). In addition, there may be other national stakeholders, such as funders and data users, whose agendas and priorities do not necessarily coincide with those of the survey's central organisation. Without special control procedures that address this complex organisational structure, it is likely that design and implementation intentions will not be met. Surveys vary in how this is done (Lynn, Japec and Lyberg, 2006, p. 10):

> *Different models have been used for the coordination and control of survey implementation. At one extreme, it is possible to set up a large central coordination team who are able to liaise closely with, and monitor the activities of, each national team throughout the implementation process. At the other extreme, the central coordinator may simply issue some written instructions on implementation and leave each nation to follow the instructions. The first of these two models is obviously resource intensive and requires substantial funding for the central activities.*

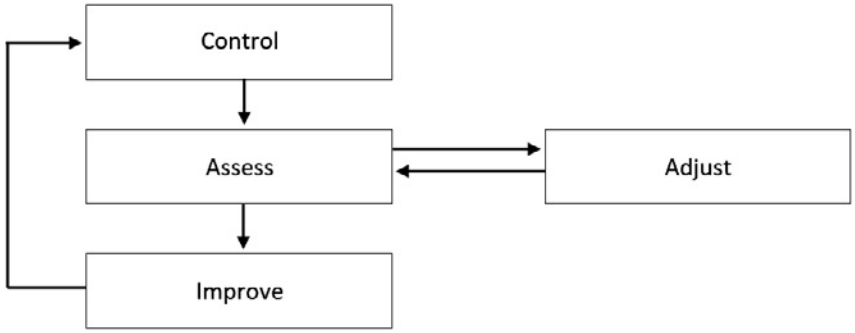
The ESS is a good example of a survey that has a sizeable central team responsible for overseeing the design and execution of the survey in each participating country ([4]). The Core Scientific Team develops a detailed specification for each round of the survey. This sets out the procedures to be followed and the standards to be implemented. Three subgroups of the Core Scientific Team are then responsible for liaising with each national coordinating team on specific issues to ensure that the specification is adhered to and that the highest possible standards are achieved. One subgroup deals with sampling and weighting, another with translation of questionnaires, and another with fieldwork. These subgroups have a remit to support the national teams in carrying out the survey: they provide advice, guidance and training, actively make suggestions regarding the details of design and implementation, and ensure that all aspects of the survey process are documented and evaluated in a standardised way. Other surveys such as the World Values Survey and the Survey of Health, Ageing and Retirement in Europe ([5]) have a somewhat less intensive model, whereby the central control is more 'light touch' but still involves regular communication with each national team. Surveys towards the other extreme of providing written instructions but little if any hands-on control include PIAAC and EU-SILC.

An important element of any survey system is continuous quality improvement. This should be closely related to procedures for control and assessment of non-sampling errors, as lessons learned should feed back into the system and result in procedural changes that are likely to reduce the impact of the errors in future. Some progress has been made in implementing survey quality improvement systems among national statistical institutes (NSIs) in Europe, although efforts have been somewhat uneven across countries (Lyberg, 2003). There are four stages within the quality improvement cycle in which non-sampling errors come to the fore (Figure 1.2). The first is the survey design stage. At this stage, many measures to control non-sampling errors can be introduced. Once data have been collected, various types of non-sampling error can then be assessed empirically. In some cases, this may lead to statistical adjustment (e.g. imputation, weighting), which, once implemented, should be followed by a further step of assessment. The final step is to learn from the experiences of implementing the survey and assessing the consequent non-sampling errors by improving the survey processes for the next survey round.

([4]) https://www.europeansocialsurvey.org/about/structure_and_governance.html

([5]) http://www.share-project.org/

Figure 1.2: Non-sampling errors in the survey quality improvement cycle



Many of the chapters in this book report the assessment of a particular aspect of non-sampling errors. Each chapter concludes with some recommendations aimed at improving the control of errors in future. In this way, we hope that this book may contribute in a modest way to improving the quality of EU-SILC. Some common themes emerge throughout the book. The following recommendations are based on these themes.

- Documentation of procedures used in EU-SILC is inconsistent and incomplete. To be able to identify and overcome weaknesses in the system, it is important to know how the survey is carried out in each country. We recommend that all NSIs are asked to compile a listing and perform an evaluation of the methods they use to handle non-sampling errors. In addition, the requirement for annual technical summaries should be strengthened.
- Knowledge of error sources and methods for dealing with them is, in many cases, lacking. A vigorous capacity-building programme should be organised. Examples of topics in which knowledge and skills are found to be lacking include questionnaire design, testing, translation and adaptation; mode effects; non-response and coverage adjustment; interviewer variance; error frameworks; social desirability bias; and coding and editing variance.
- Methodological studies are lacking. Surprisingly few NSIs are able to point to any experimental or evaluative studies of any non-sampling errors. Although the Third Network for the Analysis of EU-SILC (Net-SILC3) has provided a start in this regard, much remains to be done in order to provide the broad information base needed for informed planning going forward. One way of kick-starting this area of activity may be through the establishment of a European statistical system network ([6]). A mechanism for regular collaboration with academic researchers could also be helpful.

One rather fundamental issue for EU-SILC, as for any cross-national survey, is that the quality of the survey should be assessed with respect to its comparative objectives. In other words, it is the effect of non-sampling errors on comparability between countries that matters most. This cannot be assessed at a national level but requires evaluation of the cross-national data. This cannot therefore be left to NSIs. Cross-national methodological studies should be organised centrally. These should include comparisons of reliability and of equivalence.

If the recommendations above and those throughout the rest of the book are implemented, we believe that the quality of European statistics on poverty persistence, and other related statistics, may be improved, to the benefit of policymakers and ultimately of the citizens of Europe.

References

Bailar, B.A. and Dalenius, T. (1969), 'Estimating the response variance components of the US Bureau

([6]) https://ec.europa.eu/eurostat/cros/content/essnet-generalities_en

of the Census' survey model', *Sankhyā*, Vol. 31, pp. 341–360.

Behr, D. and Shishido, K. (2016), 'The translation of measurement instruments for cross-cultural surveys', in Wolf, C., Joye, D., Smith, T. W. and Fu, Y.-C. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London, pp. 269–287.

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality*, Wiley, New York.

Biemer, P. and Stokes, L. (1985), 'Optimal design of interviewer variance experiments in complex surveys', *Journal of the American Statistical Association*, Vol. 80, pp. 158–166.

Bowley, A. L. (1926), 'Measurement of the precision obtained in sampling', *Bulletin de l'institut International de Statistique*, Vol. 22 (supplement to book 1), pp. 1–62.

Cochran, W. G. (1942), 'Sampling theory when the sampling units are of unequal sizes', *Journal of the American Statistical Association*, Vol. 37, pp. 199–212.

Cochran, W. G. (1946), 'Relative accuracy of systematic and stratified random samples for a certain class of populations', *Annals of Mathematical Statistics*, Vol. 17, pp. 164–177.

Dalenius, T. (1967), *Nonsampling Errors in Census and Sample Surveys*, Stockholm University, Stockholm.

Deming, W. E. (1944), 'On errors in surveys', *American Sociological Review*, Vol. 9, No 4, pp. 359–369.

European Science Foundation (1998), *Blueprint: The European Social Survey (ESS) – A research instrument for the social sciences in Europe*, European Science Foundation, Strasbourg.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, Wiley, New York.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009), *Survey Methodology*, 2nd edition, Wiley, Hoboken, NJ.

Hansen, M. and Hurwitz, W. (1943), 'On the theory of sampling from finite populations', *The Annals of Mathematical Statistics*, Vol. 14, pp. 333–362.

Hansen, M. and Hurwitz, W. (1946), 'The problem of nonresponse in sample surveys', *Journal of the American Statistical Association*, Vol. 41, pp. 517–529.

Hansen, M. and Steinberg, J. (1956), 'Control of errors in surveys', *Biometrics*, Vol. 12, pp. 462–474.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953), *Sample Survey Methods and Theory*, Vols 1 and 2, Wiley, New York.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P. et al. (2010a), 'Comparative survey methodology', in Harkness, J. A. et al. (eds), *Survey Methods in Multinational, Multiregional and Multicultural Contexts*, Wiley, Hoboken, NJ, pp. 3-16.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., Pennell, B.-E. and Smith, T. W. (eds) (2010b), *Survey Methods in Multinational, Multiregional and Multicultural Contexts*, Wiley, Hoboken, NJ.

Jabine, T., Straf, M., Tanur, J. and Tourangeau, R. (1984), *Cognitive Aspects of Survey Methodology: Building a bridge between disciplines – Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology*, National Academy of Sciences Press, Washington DC.

Johnson, T. P. and Braun, M. (2016), 'Challenges of comparative survey research', in Wolf, C., Joye, D., Smith, T. W. and Fu, Y.-C. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London, pp. 41–54.

Johnson, T. P., Pennell, B.-E., Stoop, I. A. L. and Dorer, N. (2018), *Advances in Comparative Survey Methods: Multinational, multiregional and multicultural contexts (3MC)*, Wiley, Hoboken.

Jowell, R. (1998), 'How comparative is comparative research?', *American Behavioural Scientist*, Vol. 42, No 2, pp. 168–177.

Kirsch, I. and Thorn, W. (2019), 'The programme for international assessment of adult competencies: an overview', in OECD, *Technical Report of the Survey of Adult Skills (PIAAC)* (3rd edition), OECD, Paris, Preface.

Kish, L. (1965), *Survey Sampling*, Wiley, New York.

Kjær, A. (1897), 'The representative method of statistical surveys', *Kristiania Videnskapsselskabets Skrifter, Historik-filosofiske Klasse*, Vol. 4, pp. 37–56 (in Norwegian).

Lyberg, L. (2003), 'Quality improvement in European national statistical institutes', paper presented to the Washington Statistical Society, 23 November.

Lyberg, L. and Stukel, D. (2017), 'The roots and evolution of the total survey error concept', in Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C. and West, B. T. (eds), *Total Survey Error in Practice*, Wiley, Hoboken, NJ, pp. 1–22.

Lyberg, L. E. and Weisberg, H. F. (2016), 'Total survey error: a paradigm for survey methodology', in Wolf, C., Joye, D., Smith, T. W. and Fu, Y.-C. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London, pp. 27–40.

Lynn, P. (2003), 'Developing quality standards for cross-national survey research: five approaches', *International Journal of Social Research Methodology*, Vol. 6, No 4, pp. 323–336.

Lynn, P. (2008), 'The problem of non-response', in de Leeuw, E. D., Hox, J. J. and Dillman, D. A. (eds), *International Handbook of Survey Research Methods*, Lawrence Erlbaum Associates, Mahwah, pp. 35–55.

Lynn, P., Japec, L. and Lyberg, L. (2006), 'What's so special about cross-national surveys?', *ZUMA Nachrichten Spezial*, Vol. 12, pp. 7–20.

Mahalanobis, P. C. (1946), 'Recent experiments in statistical sampling in the Indian Statistical Institute', *Journal of the Royal Statistical Society*, Vol. 109, pp. 325–378.

Neyman, J. (1934), 'On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection', *Journal of the Royal Statistical Society*, Vol. 97, No 4, pp. 557–625.

Neyman, J. (1938), 'Contribution to the theory of sampling human populations', *Journal of the American Statistical Association*, Vol. 33, pp. 101–116.

Park, A. and Jowell, R. (1997), 'Consistencies and differences in a cross-national survey', *Working Papers*, SCPR, London.

Pennell, B.-E. and Cibelli Hibben, K. (2016), 'Surveying in multicultural and multinational contexts', in Wolf, C., Joye, D., Smith, T. W. and Fu, Y.-C. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London, pp. 157–177.

Survey Research Center (2016), *Guidelines for Best Practice in Cross-Cultural Surveys*, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (http://ccsg.isr.umich.edu/).

Tourangeau, R., Rips, L. J. and Rasinski, K. A. (2000), *The Psychology of Survey Response*, Cambridge University Press, Cambridge.

Yates, F. (1949), *Sampling Methods for Censuses and Surveys*, Charles Griffin, London.

2 Investing in statistics: EU-SILC

Emilio Di Meglio, Didier Dupré and Sigita Grundiza ([7])

2.1. Introduction

This chapter introduces the European Union Statistics on Income and Living Conditions (EU-SILC) instrument, which is the reference source for comparative statistics on income distribution and social inclusion in the EU. Its aim is to provide the reader of this book with a conceptual and practical insight into the background of this instrument and its main characteristics.

Reliable and timely statistics and indicators, computed from a pan-European harmonised data source and reflecting the multidimensional nature of poverty and social exclusion, are essential for monitoring the social protection and social inclusion process at national and EU levels. Furthermore, the social consequences of the global economic, financial and health crisis caused by the COVID-19 pandemic have given increased importance to data on income distribution, the social situation, and poverty and social exclusion across Europe.

2.2. EU-SILC instrument and its governance

2.2.1. Scope and geographical coverage

As with most household surveys, EU-SILC covers only people living in private households; people living in a collective household or an institution are therefore not included in the instrument. This needs to be borne in mind when carrying out statistical analyses and when interpreting indicators, both within a given country and between countries.

EU-SILC was launched in 2003 in seven countries ([8]) and later was gradually extended to all EU Member States and beyond. In 2020, the EU-SILC instrument was implemented in 37 countries, namely the 27 EU Member States, Albania, Iceland, Kosovo ([9]), Montenegro, North Macedonia, Norway, Serbia, Switzerland, Turkey and the United Kingdom ([10]). A pilot took place in Bosnia and Herzegovina in 2019, and full survey implementation was envisaged for 2021. Small areas of the national territory of some of these countries amounting to no more than 2 % of the national population are excluded from EU-SILC, as are the following national territories: the French overseas departments and territories, the Dutch West Frisian Islands with the exception of Texel, and the Isles of Scilly.

In 2019, 297 000 households and 611 000 individuals were interviewed for EU-SILC, and the complete microdata were sent to Eurostat.

([7]) Emilio Di Meglio, Didier Dupré and Sigita Grundiza are all at the statistical office of the European Union – Eurostat. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Eurostat (estat-secretariat-F4@ec.europa.eu).

([8]) Austria, Belgium, Denmark, Greece, Ireland, Luxembourg and Norway.

([9]) This designation is without prejudice to positions on status, and is in line with UNSCR 1244 and the ICJ Opinion on the Kosovo declaration of independence.

([10]) The two countries that joined EU-SILC most recently are Kosovo (in 2018) and Bosnia and Herzegovina (in 2019 (pilot)).

2.2.2. Main characteristics of EU-SILC

All EU Member States are required to implement EU-SILC, which is based on the idea of a common 'framework' as opposed to a common 'survey'. The common framework consists of common procedures, concepts and classifications, including harmonised lists of target variables to be transmitted to Eurostat.

Two types of annual data are collected through EU-SILC and provided to Eurostat.

- Cross-sectional data pertain to a given time period and include variables on income, poverty, social exclusion and other living conditions. The data for the survey of year *N* are currently to be transmitted to Eurostat by November of year (*N* + 1), although many countries manage to send the data before this deadline. In 2020, for example, 16 countries sent their 2019 data by the end of June 2020; by the end of October 2020 EU indicators were published for 27 countries.
- Longitudinal data pertaining to changes over time at the individual level are observed periodically over a 4-year period. Longitudinal data are confined to income information and a reduced set of other variables, designed to identify the incidence and dynamic processes of persistent poverty and social exclusion among subgroups of the population. The longitudinal data corresponding to the period between year (*N* – 3) and year *N* are currently to be transmitted to Eurostat by March of year (*N* + 2). Many countries manage to send longitudinal weights in advance, together with the cross-sectional transmission.

With the entry into force of the EU regulation on integrated European social statistics (IESS) in 2021 ([11]), this calendar for transmission has been modified and data have to be transmitted by December of year *N* as from 2021. Longitudinal data will have to be transmitted by October of year (*N* + 1). Eurostat proposed an integrated design with a 4-year rotational panel to those countries that had launched a new survey. A 4-year rotational panel design involves a new sample (panel) being selected each year and included in the survey for 4 years. Each new sample (replication) is similar in size and design, and representative of the whole population. Thus, in year *N*, the panels from years (*N* – 1), (*N* – 2) and (*N* – 3) are maintained, whereas the panel selected in year (*N* – 4) is dropped and replaced by a new one.

The fundamental characteristic of the integrated design is that the cross-sectional and longitudinal statistics are produced from essentially the same set of sample observations ([12]), thus avoiding the unnecessary duplications that would be involved if entirely separate cross-sectional and longitudinal surveys were used.

2.2.3. Legal basis

One of the strengths of the EU-SILC instrument is the existence of a legal basis that is binding on EU Member States, as well as a requirement for candidate and potential candidate countries. The development of the common framework, including the conception of the annual ad hoc modules, is discussed on a permanent basis with the main stakeholders, particularly within the EU Working Group for Statistics on Living Conditions chaired by Eurostat. Given that the new IESS legal basis entered into force in 2021, the focus in this chapter is on the pre-2021 EU-SILC framework.

The EU-SILC legal basis used until the implementation of the IESS regulation (and hence for the data analysed in this book) consists of three main components.

- A framework regulation ([13]) covers scope, definitions, time reference, characteristics of the data, data required, sampling rules, sample sizes, transmission of data, publication, access for scientific purposes,

([11]) Regulation (EU) 2019/1700 of the European Parliament and of the Council of 10 October 2019 establishing a common framework for European statistics relating to persons and households, based on data at individual level collected from samples, OJ L 261, 14.10.2019, p. 1 (https://eur-lex.europa.eu/eli/reg/2019/1700/oj).

([12]) Currently, only the United Kingdom derives cross-sectional and longitudinal data from two different survey instruments.

([13]) Regulation (EC) No 1177/2003 of the European Parliament and of the Council of 16 June 2003 concerning Community statistics on income and living conditions (EU-SILC), OJ L 165, 3.7.2003, p. 1.

financing, reports and studies for the EU-SILC instrument. This regulation was amended by Regulation (EC) No 1553/2005 ([14]) and Council Regulation (EC) No 1791/2006 ([15]) in order to extend the EU-SILC instrument to include the new Member States.

- Five European Commission regulations specify some technical aspects of the instrument: 'definitions' ([16]), 'fieldwork aspects and imputation procedures' ([17]) 'sampling and tracing rules' ([18]), 'list of primary [annual] target variables' ([19]) and 'quality reports' ([20]).
- The third main component is the annual Commission regulations on the list of secondary target variables, namely the ad hoc thematic modules, which cover a different topic each year and can be repeated after 5 years or less. No systematic repetition is set up.

([14]) Regulation (EC) No 1553/2005 of the European Parliament and of the Council of 7 September 2005 amending Regulation (EC) No 1177/2003 concerning Community statistics on income and living conditions (EU-SILC), OJ L 255, 30.9.2005, p. 6.

([15]) Council Regulation (EC) No 1791/2006 of 20 November 2006 adapting certain regulations and decisions in the fields of free movement of goods, freedom of movement of persons, company law, competition policy, agriculture (including veterinary and phytosanitary legislation), transport policy, taxation, statistics, energy, environment, cooperation in the fields of justice and home affairs, customs union, external relations, common foreign and security policy and institutions, by reason of the accession of Bulgaria and Romania, OJ L 363, 20.12.2006, p. 1.

([16]) Commission Regulation (EC) No 1980/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards definitions and updated definitions, OJ L 298, 17.11.2003, p. 1 (updated by Commission Regulation (EC) No 676/2006).

([17]) Commission Regulation (EC) No 1981/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the fieldwork aspects and the imputation procedures, OJ L 298, 17.11.2003, p. 23.

([18]) Commission Regulation (EC) No 1982/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the sampling and tracing rules, OJ L 298, 17.11.2003, p. 29.

([19]) Commission Regulation (EC) No 1983/2003 of 7 November 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the list of target primary variables, OJ L 298, 17.11.2003, p. 34.

([20]) Commission Regulation (EC) No 28/2004 of 5 January 2004 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the detailed content of intermediate and final quality reports, OJ L 5, 9.1.2004, p. 42.

The EU-SILC instrument is also applicable to Iceland, Norway, Switzerland and the United Kingdom. As for candidate and potential candidate countries, the implementation of EU-SILC is not compulsory until they join the EU, but it is strongly encouraged if the specific situation of a given country permits it.

2.2.4. Common guidelines

The way to implement the EU-SILC legal basis is agreed between Eurostat and the national statistical institutes (NSIs) – particularly in the EU Working Group for Statistics on Living Conditions and the task forces reporting to it ([21]). This includes guidelines on common procedures and concepts, as well as an increasing number of recommendations on how to word the underlying questions. The full set of guidelines is available to the public ([22]). The guidelines are updated yearly in order to fine-tune the data collection on particular topics or in order to further improve methodological issues with the final aim of continuously improving the comparability between countries, and are agreed by the working group.

Strategic issues regarding the development of EU-SILC are discussed in the meetings of the European Statistical System Committee (ESSC) ([23]) and the NSIs' Directors of Social Statistics.

([21]) These task forces support the work of the EU Working Group for Statistics on Living Conditions. For instance, the task force on the revision of the EU-SILC legal basis provided a major contribution to the development of the IESS regulation. The set of secondary variables included in EU-SILC modules is generally prepared by an ad hoc task force. Important work on the development of a set of material deprivation variables and of related EU social indicators was performed by the task force on material deprivation.

([22]) See, in particular, annual guidelines available in the EU-SILC dedicated interest group folder (https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d).

([23]) The ESSC is at the heart of the European Statistical System. It is chaired by Eurostat and composed of the representatives of Member States' NSIs. The European Free Trade Association (EFTA) countries Iceland, Liechtenstein, Norway and Switzerland, as well as the EFTA Statistical Office, participate as observers. Observers from the European Central Bank, the Organisation for Economic Co-operation and Development, etc., may also participate in the meetings of the ESSC. The ESSC meets three times a year.

2.3. Methodological framework

2.3.1. Contents of EU-SILC

EU-SILC is a multidimensional instrument focused on income that also covers housing, labour, health, demography, education and deprivation, to allow for the analysis of the multidimensional phenomena of poverty and social exclusion, and for the joint analysis of its different dimensions. It consists of primary (annual) and secondary (ad hoc modules) target variables, all of which are forwarded as microdata sets by Member States to Eurostat.

Given the principle of flexibility of the implementation of EU-SILC at national level, the sequence of questions needed to construct one target variable may vary from country to country. Nevertheless, recommended wording for questions is available for the ad hoc modules as well as a number of primary variables (such as health and material deprivation variables), although countries are not obliged to follow these recommendations.

The primary target variables relate to either household or individual (for people aged 16 or more) information , with each grouped into five areas:

- basic/core data, income, housing, social exclusion and labour information at household level;
- basic/demographic data, income, education, labour information and health at personal level.

The secondary target variables are introduced, and sometimes repeated after some years, only in the cross-sectional component. One ad hoc module per year has been included since 2005 ([24]):

- 2005: Intergenerational transmission of poverty,
- 2006: Social participation,
- 2007 and 2012: Housing conditions,
- 2008: Over-indebtedness and financial exclusion,
- 2009 and 2014: Material deprivation,
- 2010: Intra-household sharing of resources,
- 2011: Inter-generational transmission of disadvantages,
- 2013: Well-being,
- 2015: Social/cultural participation and material deprivation,
- 2016: Access to services,
- 2017: Health and children's health,
- 2018: Material deprivation, well-being and housing difficulties,
- 2019: Intergenerational transmission of disadvantages, household composition and evolution of income,
- 2020: Over-indebtedness, consumption and wealth as well as labour.

([24]) For detailed information on the content of these modules, see Eurostat's website (https://ec.europa.eu/eurostat/web/income-and-living-conditions/data/ad-hoc-modules).

2.3.2. Income concept

An important objective for EU-SILC is to adhere as closely as possible to the recommendations of the international Canberra Group on the definition of household income (UNECE, 2011). The income concept in the sense of the Canberra recommendations has been fully implemented since 2007 in EU-SILC.

Two main aggregates are computed from EU-SILC – total gross household income (GI) and total disposable household income (DI), which are defined as:

- GI = EI + SEI + PP + CTR + OI
- DI = GI – CTP

where:

EI = employee income (cash or near-cash employee income and non-cash employee income),

SEI = self-employment income (but not goods produced for own consumption),

PP = pensions received from individual private plans,

CTR = current transfers received (social benefits and regular inter-household cash transfers received),

OI = other sources of income received (such as capital income),

CTP = current transfers paid (tax on income and social insurance contributions, tax on wealth and regular inter-household cash transfers paid).

Employee income

In EU-SILC, employee income is covered by the collection of information on 'Gross cash or near-cash employee income', 'Gross non-cash employee income' and 'Employers' social insurance contributions'. For non-cash employee income, only company cars have been recorded since the beginning of EU-SILC and included in the income concept. Information covering all other goods and services provided free of charge or at a reduced price by employers to their employees and the compulsory component of employers' social insurance contributions are to be collected but are not yet included in the main income aggregates.

Self-employment income

Self-employment income is broken down into 'Gross cash profits or losses from self-employment' (including royalties) and the 'Value of goods produced for own consumption'. Various alternative approaches to the measurement of income from self-employment are allowed. The value of goods produced for own consumption is not included in the main income aggregates.

Private pension plans

Regular pensions from private plans – other than those covered within the 'Current transfers' item – refer to pensions and annuities received in the form of interest or dividend income from individual private insurance plans, that is, fully organised schemes in which contributions are at the discretion of the contributor independently of their employers or government.

Current transfers received

Current transfers received include social benefits and regular inter-household cash transfers received. Social benefits are broken down into family and children-related allowances, housing allowances, unemployment benefits, old-age benefits, survivors' benefits, sickness benefits, disability benefits, education-related allowances and 'other benefits not elsewhere classified'.

Other sources of income received

Three sources of income are covered under this item:

- income from rental of a property or land,
- interest, dividends and profits from capital investment in unincorporated businesses,
- income received by people aged under 16.

Current transfers paid

Current transfers paid are broken down into 'Tax on income and social insurance contributions', 'Regular taxes on wealth' and 'Regular inter-household cash transfers paid'. The 'Employers' social insurance contributions' variable is not included in the computation of the main income aggregates, even though it would be crucial for cross-country comparisons related to labour costs.

Imputed rent

Information on imputed rent has been collected from 2007 onwards for all households that do not report that they pay full rent, namely households that own the dwelling they live in (owner-occupiers) or households that enjoy subsidised rents. However, the value of imputed rent is not included in the main income aggregates. Its inclusion would have a significant impact on all income-based indicators, but a methodology for achieving comparable results for all countries is not yet available ([25]). (For a discussion on the distributional impact of imputed rent in EU-SILC and the lack of cross-country comparability of this component, see Törmälehto and Sauli (2017).)

Imputation

The EU-SILC framework requires full imputation for income components. The level of imputation of income components is reported in microdata by means of a set of detailed flags. This requirement helps to make the information delivered by the instrument more homogeneous and complete. Imputation is performed by Member States.

([25]) The position of the Indicators' Sub-Group of the Social Protection Committee is that the imputed rent component could be included in a small number of income poverty indicators that would be listed in the EU social inclusion portfolio (see below) as secondary indicators or context information.

Income reference period

In all but two countries, Ireland and the United Kingdom, the income reference period is the previous calendar year. Thus, for a survey conducted in year *N* the income information that is collected refers to the household income received between 1 January (*N* – 1) and 31 December (*N* – 1) (put differently, the 'survey year' is *N* and the 'income year' is (*N* – 1)). Ireland and the United Kingdom use a sliding reference period. In Ireland, it refers to the 12 months prior to the interview date. In the United Kingdom, it is centred on the interview date, meaning it covers 6 months before and 6 months after the interview. In addition, the respondents are asked to provide figures that relate most commonly to their current (and usual) incomes, that is, which could relate to the past week, 2 weeks or month. These figures are then annualised.

The further in time the fieldwork period is from the income reference period, the higher the risk of inconsistency between income-related variables and other socioeconomic variables (including sociodemographic variables). It is therefore essential to limit, as much as possible, the lag between the income reference period and the fieldwork by conducting the interviews preferably in the first quarter of the year.

Equivalised income

Most income-based EU social indicators are computed using an 'equivalised disposable income', which is calculated in three steps: (i) all monetary income received from any source by each member of a household is added up (this includes income from work, investments and social benefits, plus any other household income; taxes and social contributions that have been paid are then deducted from this sum); (ii) in order to reflect the differences in a household's size and composition, the total (disposable) household income is divided by the number of 'equivalent adults' using the so-called Organisation for Economic Co-operation and Development-modified (equivalence) scale, which gives a weight to all members of the household (1 to the first adult, 0.5 to the second and each subsequent person aged 14 and over, and 0.3 to each child aged under 14); (iii) finally, the resulting figure – the equivalised disposable income – is attributed to each member of the household (adults and children). This means that, for a couple and two children, income is divided by 2.1 (1 + 0.5 + 0.3 + 0.3), so that an annual income of EUR 10 500 becomes an equivalised income of EUR 5 000, which is artificially assigned to each of the four household members (i.e. including to each of the two children).

2.3.3. Sample requirements

Sampling design

EU-SILC data are to be collected from nationally representative probability samples of the population residing in private households within the country, irrespective of language, nationality or legal residence status. All private households and all people aged 16 and over within the household are eligible for the operation. Representative probability samples must be achieved both for households and for individuals in the target population. The sampling frame and methods of sample selection should ensure that every individual and household in the target population is assigned a known probability of selection that is not zero.

Sample size

The framework regulation and its updates define the minimum effective sample sizes to be achieved. The 'effective' sample size is the size that would be required if the survey were based on simple random sampling (design effect in relation to the EU at-risk-of-poverty rate indicator = 1.0). The actual sample sizes have to be larger to the extent that the design effect exceeds 1.0 because of complex sampling designs and in order to compensate for all kinds of non-response. The sample sizes for the longitudinal component refer, for any two consecutive years, to the number of households or individuals aged 16 and over that are successfully interviewed in both years. Table 2.1 gives the minimum effective sample sizes required for each EU Member State (plus Albania, Iceland, Kosovo, Montenegro, North Macedonia, Norway, Serbia, Switzerland and the United Kingdom) in terms of households and individuals aged 16 or over.

Table 2.1: Minimum effective sample size for the cross-sectional and longitudinal components

Country	Households		Individuals aged 16 or over to be interviewed	
	Cross-sectional	Longitudinal	Cross-sectional	Longitudinal
Belgium	4 750	3 500	8 750	6 500
Bulgaria	4 500	3 500	10 000	7 500
Czechia	4 750	3 500	10 000	7 500
Denmark	4 250	3 250	7 250	5 500
Germany	8 250	6 000	14 500	10 500
Estonia	3 500	2 750	7 750	5 750
Greece	4 750	3 500	10 000	7 250
Spain	6 500	5 000	16 000	12 250
France	7 250	5 500	13 500	10 250
Croatia	4 250	3 250	9 250	7 000
Ireland	3 750	2 750	8 000	6 000
Italy	7 250	5 500	15 500	11 750
Cyprus	3 250	2 500	7 500	5 500
Latvia	3 750	2 750	7 650	5 600
Lithuania	4 000	3 000	9 000	6 750
Luxembourg	3 250	2 500	6 500	5 000
Hungary	4 750	3 500	10 250	7 750
Malta	3 000	2 250	7 000	5 250
Netherlands	5 000	3 750	8 750	6 500
Austria	4 500	3 250	8 750	6 250
Poland	6 000	4 500	15 000	11 250
Portugal	4 500	3 250	10 500	7 500
Romania	5 250	4 000	12 750	9 500
Slovenia	3 750	2 750	9 000	6 750
Slovakia	4 250	3 250	11 000	8250
Finland	4 000	3 000	6 750	5 000
Sweden	4 500	3 500	7 500	5 750
Total EU	127 500	95 750	268 400	200 350
Iceland	2 250	1 700	3 750	2 800
Montenegro	3 250	2 500	8 750	6 500
North Macedonia	3 750	3 000	11 500	8 750
Norway	3 750	2 750	6 250	4 650
Serbia	4 500	3 500		
Switzerland	4 250	3 250	7 750	5 800
Turkey	7 750	5 750	21 000	
United Kingdom	7 500	5 750	13 750	10 500

Sources: Regulation (EC) No 1553/2005 and Council Regulation (EC) No 1791/2006. For candidate countries, the minimum effective sample size is not regulated.

2.3.4. Tracing rules

In order to ensure the best quality output, minimum requirements for implementation have been defined within the legal basis in addition to the definition of the minimum sample size. These rules concern, for instance, the use of proxy interviews, the use of substitutions, fieldwork duration, non-response procedures and tracing (or 'following') rules.

In each country, the longitudinal component of EU-SILC consists of one or more panels or subsamples (four subsamples in the recommended 4-year rotational design). For each panel/subsample, the initial households representing the target population at the time of its selection are followed for a minimum period of 3 years on the basis of specific tracing rules. The objective of the tracing rules is to follow up individuals over time.

In order to study changes over time at the individual level, all sample individuals (members of the panel/subsample at the time of their selection) should be followed up over time, despite the fact that they may move to a new location during the life of the panel/subsample. However, in the EU-SILC implementation some restrictions are applied for cost and other practical reasons. Only those people staying in one private household or moving from one to another in the national territory are followed up. Sample individuals moving to a collective household, an institution or national territories not covered in the survey, or moving abroad (to a private household, collective household or institution, within or outside the EU), would normally not be traced. The only exception would be the continued tracing of those moving temporarily (for an actual or intended duration of less than 6 months) to a collective household or institution within the national territory covered, as they are still considered household members. Tracing rules have changed with the entry into force of the IESS regulation.

2.4. Information on quality

2.4.1. Some comparability issues

The flexibility of the EU-SILC instrument may be seen as both its main strength and its main weakness. Although flexibility allows EU-SILC to be embedded into the national systems of social surveys, it can create problems of harmonisation and comparability across countries. This section addresses some of these comparability issues.

Different sampling designs

Almost all countries have used the integrated design proposed by Eurostat.

The EU-SILC framework encourages the use of existing sources and/or administrative data. However, in practice, not all EU-SILC variables can be obtained from registers and administrative data. Hence, it is possible to establish two groups of countries on the basis of the data source used in EU-SILC.

- In the countries referred to as 'register' countries (Denmark, Finland, Iceland, the Netherlands, Norway, Slovenia and Sweden), most income components and some items of demographic information are obtained through administrative registers. Other personal variables are obtained by means of interview from a sample of individuals according to the 'selected respondent model' (see below for more details), whereby only one member of the household answers the detailed questionnaire, whereas the income information is derived from register data for all household members. More and more countries are moving towards retrieving income information from registers but without moving to the selected respondent model. This is the case for Belgium, Estonia, Spain, France, Italy, Cyprus, Latvia, Malta and Austria, which use registers and/or a combination of register and survey data to construct some income variables (see Zardo Trindade and Goedemé, 2020).
- In other countries, the full information is obtained by means of a survey of households and interviews with household members.

All the national sampling designs ensure strict cross-sectional representativeness and enable a significant number of individuals to be followed over a period of at least 4 years. In line with the legal requirements, all samples are probabilistic ([26]), with updated sampling frames and stochastic al-

([26]) Germany used quota sample by derogation until 2008.

gorithms used to select statistical units. The sampling designs used in 2018 by country were as follows:

- sampling of dwellings or addresses: Albania, Austria, Croatia, Czechia, France, Hungary, Latvia, the Netherlands, Poland, Portugal, Romania, Spain and the United Kingdom;
- sampling of households: Belgium, Bulgaria, Cyprus, Denmark, Germany, Greece, Ireland, Italy, Kosovo, Luxembourg, Malta, Serbia, Slovakia and Switzerland;
- sampling of individuals: Estonia, Finland, Iceland, Lithuania, Norway, Slovenia and Sweden (all these countries are 'register' countries except for Lithuania).

In all cases, sample unbiased estimates can be produced on firm theoretical grounds. In almost all countries, the coverage bias is controlled with frequent updates of the frame.

Countries have designed their sample so as to achieve a good trade-off between reporting needs at subnational level and the cost-effectiveness of the data collection. Significant increases in the sample size, driven by subnational reporting requirements in view of the new framework regulation concerning EU-SILC adopted in October 2019 (see Alaminos et al., 2021), were recorded in Greece and Portugal and are planned in other countries.

Differences in the method of data collection

In most countries (i.e. the non-register countries), all members aged 16 or over in selected households are asked to respond to a personal questionnaire, whereas in the register countries only one selected respondent per household receives a personal questionnaire. These two different rules have different impacts on the tracing of individuals over time (longitudinal dimensions), depending on whether only one or all household members are interviewed over time. The selected respondent model needs some adaptation in order to avoid bias in the follow-up of children. The two different rules lead to different weighting schemes. In particular, when the selected respondent type is used, the weights of the household and of the selected respondent are obviously different.

In 2018, the most frequent mode of data collection was computer-assisted personal interviews, which were used as the primary mode of data collection in 16 countries (Belgium, Bulgaria, Croatia, Cyprus, Estonia, Spain, France, Hungary, Ireland, Italy, Latvia, Luxembourg, Malta, Austria, Poland and Portugal). This was followed by paper and pencil interviews, which were used as the primary mode in four countries (Czechia, Greece, Romania and Slovakia), computer-assisted telephone interviews, which were used in four countries (Lithuania, Slovenia, Finland and Sweden) and computer-assisted web interviews, which were used in two countries (Denmark and the Netherlands). Self-administered paper questionnaires, used in some countries as a residual mode, are used as the primary mode in Germany. Some other countries are testing web questionnaires, and some are testing mixed modes.

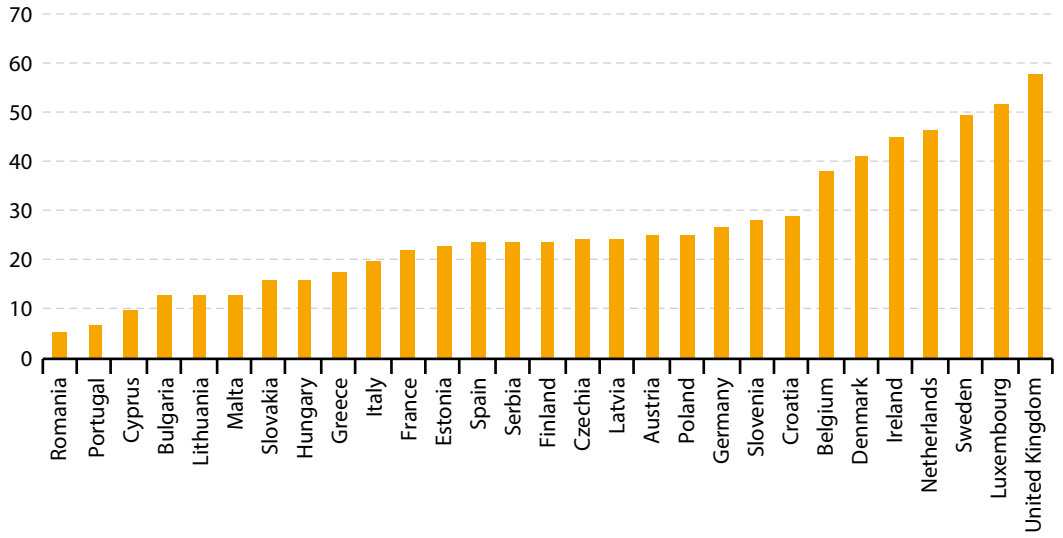
Different non-response rates

Non-response is measured in EU-SILC at three stages: address contact, household interview and personal interview. Figure 2.1 presents the overall non-response rates for individuals for the whole sample broken down by country.

Total non-response of the selected households and individuals had to be less than 40 %, which was seen as a challenge for a non-mandatory survey. The overall non-response rate in the personal interview for the whole sample of 27 Member States was equal to or below 10 % in 2018 in three countries: Romania (6 %), Portugal (7 %) and Cyprus (10 %). At the other extreme, non-response rates exceeded 30 % in six countries and even 50 % in Luxembourg (51.7 %); for non-EU countries for which information is currently available, the non-response rate was 58 % in the United Kingdom and 24 % in Serbia. Further details on response rates are presented in Chapter 5 of this book.

The creation of models using external variables in order to correct non-response is highly desirable. Most of the countries apply either a standard post-stratification, based on homogeneous response groups, or a more sophisticated logistic regression model.

Figure 2.1: Overall personal non-response rates, 2018
(%)



NB: Countries are presented in order of their non-response rate.
Sources: EU-SILC country quality reports.

2.4.2. Quality reports

Adopted in 2005, the European Statistics Code of Practice sets common standards for the independence, integrity and accountability of the national and EU statistical authorities. The EU statistical authorities have undertaken to adopt a comprehensive approach to high-quality statistics that builds on a common definition of quality in statistics, in which the following dimensions are addressed.

- **Relevance.** European statistics must meet the needs of users.
- **Accuracy and reliability.** European statistics must accurately and reliably portray reality.
- **Timeliness and punctuality.** European statistics must be disseminated in a timely and punctual manner.
- **Coherence and comparability.** European statistics should be consistent internally and over time, and comparable between regions and countries; it should be possible to combine and make joint use of related data from different sources.
- **Accessibility and clarity.** European statistics should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, and should be available and accessible on an impartial basis with supporting metadata and guidance.

This European definition of quality is monitored in EU-SILC by annual quality reports prepared by both the countries and Eurostat for the EU level, and managed through an integrated IT system.

The national quality reports provide a useful insight into national implementation practices as well as substantive information from which to draw preliminary conclusions regarding the quality of the instrument. This material is complemented by the information that Eurostat collects through its frequent contact with national statistical authorities, in particular as regards data validation, which is an integrated process with tools shared with Member States.

The purpose of the EU quality reports is to summarise the information contained in the national quality reports. Their objective is to evaluate the quality of the instrument from a European perspective,

that is, by establishing cross-country comparisons of some of its key quality characteristics.

The EU quality reports, as well as most of the national country reports, are publicly available ([27]).

2.5. Data and indicators

2.5.1. Data access

EU-SILC data are disseminated either as aggregated data or as microdata sets. Individual EU-SILC records are considered confidential data within the meaning of Article 23 of Regulation (EC) No 223/2009 (statistical law), because they allow indirect identification of statistical units (individuals and households). In this context, they should be used only for statistical purposes or for scientific research.

Aggregated results relate to indicators and statistics on income distribution and monetary poverty, living conditions, material deprivation and childcare arrangements. They are presented as predefined tables or as multidimensional data sets and may be extracted in a variety of formats ([28]).

Commission Regulation (EU) No 557/2013 granted the European Commission permission to release anonymised microdata to researchers. Anonymised microdata are defined as individual statistical records that have been modified in order to control, in accordance with best practices, the risk of identification of the statistical units to which they relate. Both EU and national rules are applied for anonymisation, and are described in full with each release. The modifications involve variable suppression, global recoding and the randomisation of some variables.

Twice a year, Eurostat releases anonymised microdata to researchers (files are available through the secure Communication and Information Resource Centre for Administrations, Businesses and Citizens). Each release contains data from the latest available operation as well as revisions from any previous data sets. A detailed description of the full procedure for accessing microdata is provided on the Eurostat website ([29]).

2.5.2. Indicators computation

In order to monitor progress towards the Europe 2020 strategy, an analytical tool has been put in place – the joint assessment framework (JAF). The JAF underpins evidence-based policymaking in the social domain. In particular, it is used as an analytical tool in the dialogue between the Commission and the Member States to support the identification of key challenges and help Member States establish their priorities. In each policy area, progress in the implementation of policies and towards the related EU social objectives is assessed quantitatively on the basis of a limited number of commonly agreed indicators. A large number of indicators are computed on the basis of EU-SILC, which has become the second pillar of household social survey statistics at EU level, complementing the EU Labour Force Survey, which focuses on labour market information.

The use of commonly agreed indicators (not only in the context of the JAF but also, more widely, to analyse the social situation across the EU and monitor progress towards the commonly agreed EU social objectives) is an essential component of EU cooperation in the social field. The development of EU social indicators is a dynamic process and is the responsibility of the Social Protection Committee and its Indicators' Sub-Group. The work of the national delegations of experts who make up the group and the secretariat provided by the European Commission's Directorate-General for Employment, Social Affairs and Inclusion (in close cooperation with Eurostat) has enabled the set of indicators (and breakdowns of these) to be considerably enriched.

EU social indicators are grouped in four portfolios: an 'overarching' portfolio and a portfolio for each of the three main social areas in which Member States cooperate (poverty and social exclusion;

([27]) See the EU-SILC interest group quality folder (https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/d2628591-af98-4060-9376-03a45337d7b3?p=1&n=10&sort=modified_DESC).

([28]) Data and publications can be accessed on the European Commission's website (https://ec.europa.eu/eurostat/web/income-and-living-conditions).

([29]) http://ec.europa.eu/eurostat/web/microdata/overview

pensions; and healthcare and long-term care) ([30]). The indicators are permanently updated and disseminated on the Eurostat website ([31]).

2.6. Way forward

Although EU-SILC has become the EU reference for data on income and living conditions, Eurostat and a number of stakeholders are reflecting on ways to further improve the tool and its (potential) uses. This book, and more generally the analysis and activities of the Third Network for the Analysis of EU-SILC – which prepared it – are part of an effort to improve EU-SILC and the development and analysis of social indicators based on it.

As mentioned above, a revision of the legal basis of EU-SILC is now being implemented. The main objectives of the revision are as follows.

- In the context of the modernisation of social statistics, integrate EU-SILC with other data collection operations, implement the standardisation of variables and modules, use administrative data sources more widely and improve statistical frames.
- Increase the responsiveness of the instrument to new policy needs, currently and for the future.
- Deliver EU-SILC data faster.
- Maintain the stability of the main indicators, with adapted frequency and keeping a cross-cutting approach.
- Maintain and if possible slightly decrease the current burden and costs.
- Allow sufficient regional breakdown.
- Ensure adequate accuracy and quality of measurements.
- Adapt to multimode and multisource data collection operations.
- Ensure a general consistency of the different elements of the tool (e.g. frequency of non-annual modules and length of the longitudinal component).

References

Alaminos, E., Di Meglio, E., Dupré, D., Grundiza, S. and Kaczmarek-Firth, A. (2021), 'Planned future developments of EU-SILC', in Guio, A.-C., Marlier, E. and Nolan, B. (eds), *Improving the understanding of poverty and social exclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 343–352.

Social Protection Committee (2015), *Portfolio of EU social indicators for the monitoring of progress towards the EU objectives for social protection and social inclusion – 2015 update*, European Commission, Brussels.

Törmälehto, V.-M. and Sauli, H. (2017), 'The distributional impact of imputed rent in EU-SILC 2007–2012', in Atkinson, A. B., Guio, A.-C. and Marlier, E. (eds), *Monitoring Social Inclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 141–157.

UNECE (United Nations Economic Commission for Europe) (2011), *Canberra Group – Handbook on household income statistics*, UN, Geneva (http://www.unece.org/fileadmin/DAM/stats/groups/cgh/Canbera_Handbook_2011_WEB.pdf).

Zardo Trindade, L. and Goedemé, T. (2020), 'The comparability of the EU-SILC income variables: review and recommendations', *Eurostat Statistical Working Papers*, Publications Office of the European Union, Luxembourg.

([30]) More information on the EU social indicators can be found on the European Commission's website (http://ec.europa.eu/social/main.jsp?catId=830&langId=en). See also Social Protection Committee (2015).

([31]) https://ec.europa.eu/eurostat/web/main/data/database

Population coverage and survey non-response

3 The effect of exclusions from the target population on EU-SILC

Tara Junes ([32])

3.1. Introduction

This chapter addresses the extent to which the European Union Statistics on Income and Living Conditions (EU-SILC) target population represents the entire population of a country. To address this question, we must first consider what is meant by the 'entire population'. The definitions usually adopted by population censuses are useful in this regard. Censuses are conducted on either a de facto or a *de jure* residency rule basis. The total of all usual residents is referred to as the *de jure* population (all usual residents, whether or not they are present at the time of the enumeration) and the total of all individuals is referred to as the de facto population (people who are physically present in the country or area at the reference date, whether or not they are usual residents). Usually, the census is based on a mix of the *de jure* and de facto population enumeration, because inclusions and exclusions of population groups depend on the national circumstances. In a general sense, national governments are responsible for the health and well-being of the union of these populations, so it may be reasonably expected that policy-relevant data analysis should relate to these populations.

However, the target population of EU-SILC is 'all private households and their current members residing in the territory of the Member States at the time of data collection' (Commission Regulation (EC) No 1982/2003). A private household is defined as 'a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of living essentials' (Regulation (EC) No 1177/2003). People not living in private households belong to the non-private household population. This consists of people living in collective or institutional households ([33]), the homeless and itinerants.

The target population comprises the group of units about which survey information is sought. In this chapter, we study the effect of excluding the non-private household population from the target population, that is, survey under-coverage due to restrictions in the definition of the target population. We will concentrate on the proportion and composition of the non-private household population in EU-SILC countries. For more information about coverage problems and special sampling methods for multinational surveys, see Gabler and Häder (2016), Heeringa and O'Muircheartaigh (2010) and Lynn et al. (2007).

A list of target population units used for drawing the sample is defined as the sampling frame. The sampling frame determines how well the target population is covered and affects the choice of the data collection method. Ideally, the frame would contain every unit of the target population and auxiliary data taken from various sources (for example from a population census or from an administrative

([32]) Tara Junes is with Statistics Finland. The author wishes to thank Anne-Catherine Guio, Tarja Hatakka, Lars E. Lyberg, Peter Lynn, Eric Marlier and Veli-Matti Törmälehto for their valuable comments and suggestions. All errors are strictly the author's responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Tara Junes (tara.junes@stat.fi).

([33]) Collective households or institutional households (as opposed to private households) are, for instance, hospitals, old people's homes, residential homes, prisons, military barracks, religious institutions, boarding houses and workers' hostels.

register) to be used to develop an efficient sampling strategy. The sampling frame can be a simple list of population elements; for example, a number of countries use registries of addresses or of people as sampling frames (Groves et al., 2009, chapter 3). The process used to construct the sampling frame affects the causes and scale of coverage problems. For example, most of the EU-SILC countries use a population census or register as a sampling frame. It is possible that some target population units are totally missing from the sampling frame. Sampling frame under-coverage of the target population is dealt with in Chapter 4. These two under-coverage aspects – namely under-coverage related to restrictions of the target population (this chapter) and population units missing from the frame population (Chapter 4) – together constitute the total under-coverage of EU-SILC.

This chapter is organised as follows. In Section 3.2, we introduce the requirements for the EU-SILC sampling frames. Section 3.3 is about the under-coverage induced by the adopted target population definition. Section 3.4 describes a case study carried out with Finnish EU-SILC data. Section 3.5 concludes by discussing the findings of this chapter and their implications.

3.2. General requirements for EU-SILC sampling frames

The implementing regulation concerning sampling and tracing rules (Commission Regulation (EC) No 1982/2003) clearly states that the reference population of EU-SILC is all private households and their current members residing in the territory of the EU Member States at the time of data collection. People living in collective households and in institutions are excluded from the target population. Small parts of the national territory may also be excluded (see Section 3.3.1).

The EU-SILC data should be based on a nationally representative probability sample of the private household population within the country, irrespective of language, nationality or legal residence status. This applies to both the cross-sectional and the longitudinal components. All individuals aged 16 and over within the private households are eligible for the survey. Every individual and household in the target population has a known and non-zero probability of selection. The sampling frame and methods of sample selection should ensure this.

The ultimate units used in the sample selection are not strictly regulated. They may be addresses, households or people, provided that they are selected with a known probability. However, from the ultimate selection units it is always necessary to construct a sample of households, the probability of each household in the sample being determined through its association with units in the sample selected. The analysis units can be households, all members, adult members or possibly a subsample of adult members; these are the units to which the information collected pertains. Their probabilities of selection (or the corresponding sample weights) are determined through their association with the sample household. The collection unit refers to the person or source providing the information.

3.3. Under-coverage induced by the adopted target population definition

The sampling frame under-coverage induced by the adopted target population definition can be studied using the 2011 Census Hub data on non-private and private household populations. According to Commission Regulation (EC) No 1982/2003, the target population of EU-SILC does not have to cover (i) some small parts of the national territories and (ii) the non-private household population. The 2011 Census Hub data provide an overview of the second alternative concerning the non-private household population. The exclusion of certain geographical areas is discussed only briefly in the following subsection.

Table 3.1: National territories that may be excluded from EU-SILC

Country	Territories
France	French overseas departments and territories
Ireland	All offshore islands, with the exception of Achill, Bull, Cruit, Gorumna, Inishnee, Lettermore, Lettermullan and Valentia
Netherlands	The West Frisian Islands, with the exception of Texel
United Kingdom	Scotland north of the Caledonian Canal, the Isles of Scilly

Source: Commission Regulation (EC) No 1982/2003.

3.3.1. Exclusion of geographical areas

According to the implementing regulation, small parts of the national territory amounting to no more than 2 % of the national population and the national territories listed in Table 3.1 may be excluded from EU-SILC ([34]). The territories listed in Table 3.1 are of different sizes. For example, there were about 1.9 million people living in the French overseas departments and territories in 2019 (this constituted about 2.9 % of the total population of France ([35])). The West Frisian Islands population comprises about 21 000 people, amounting to a share of about 0.1 % of the total population of the Netherlands (CBS, 2012).

It is, of course, permitted to include national territories amounting to less than 2 % in EU-SILC. For example, Finland includes Åland in EU-SILC, even though it accounted for around 0.5 % of the total population in 2019 ([36]). According to the national EU-SILC quality reports, it seems to be quite rare that national territories are excluded. Hence, we focus here on the effect of excluding the non-private household population.

3.3.2. The non-private household population

According to Regulation (EC) No 1177/2003, a private household is defined as 'a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of living essentials'. Commission Regulation (EC) No 1201/2009 on population and housing censuses has a somewhat similar definition for private households.

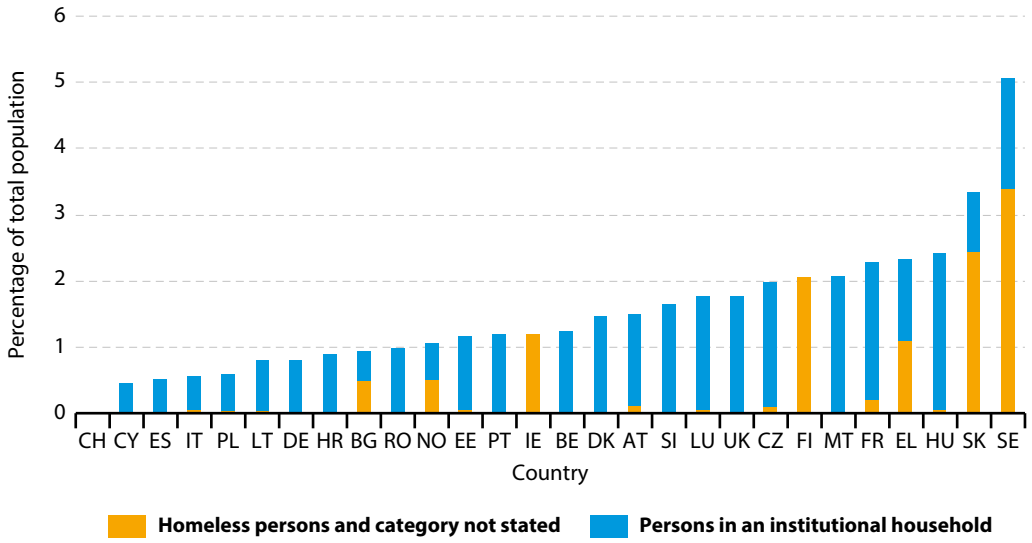
To identify private households for population censuses, Member States apply either the housekeeping concept or, if that is not possible, the household-dwelling concept. According to the housekeeping concept, a private household is either a one-person household or a multiperson household. The multiperson household is defined as a group of two or more individuals who combine to occupy the whole or part of a housing unit and to provide themselves with food and possibly other essentials for living. Members of the group may pool their incomes to a greater or lesser extent. The definition seems to be quite close to the private household definition of EU-SILC.

The household-dwelling concept of the population census (Commission Regulation (EC) No 1201/2009) is clearly more register based than the housekeeping concept. The household-dwelling concept considers all individuals living in a housing unit to be members of the same household, such that there is one household per occupied unit. It does not include any information about the shared expenditures between the household-dwelling unit members. In the household-dwelling concept, the number of occupied housing units and the number of households occupying them are equal.

To get an idea about the structure of the frame under-coverage induced by the adopted target population definition, we next study the demographics of the non-private population using 2011 Census Hub data. We acknowledge that the EU-SILC definition of the non-private household population

([34]) This is an agreement between the Member States concerned and the European Commission.
([35]) Population on 1 January by Nomenclature of Territorial Units for Statistics 2 region (TGS00096) (https://ec.europa.eu/eurostat/databrowser/view/tgs00096/default/table?lang=en).
([36]) Statistics Finland 2020 population structure statistics (http://www.stat.fi/til/vaerak/tau_en.html).

Figure 3.1: Share of the non-private household population among the total population, 2011



NB: Countries are ranked according to their share of the non-private household population. The share of people living in a non-private household in Sweden was about 5 % of the total population in 2011. The share of people living in an institutional household was about 2 % of the total population. The share of primarily homeless people and people belonging to a category not stated was about 3 % of the total population.

Source: 2011 Census Hub data (accessed 25 January 2019).

slightly differs from the corresponding definition of the population and housing census (EU-SILC does not apply the household-dwelling concept). The data are also somewhat out of date and do not fully take into account the recent changes in immigration ([37]). However, these data are a reliable and European-wide source for our study.

The 2011 Census Hub classifies the non-private household population into three subcategories:

1. people in an institutional household,
2. primarily homeless people,
3. people not living in a private household, but category not stated.

According to Commission Regulation (EC) No 1201/2009, primarily homeless people are people living on the streets without a shelter that would fall within the scope of a living quarter. A living quarter is housing that is the usual residence of one or more individuals. For the definition of subcategories 1 and 3, there seems to be no explicit guidance in the regulation.

Figure 3.1 represents the share of the non-private household population among the total population in 2011 ([38]). In most countries, the share of the non-private household population was less than 2 % – that is, quite modest. Slovakia and Sweden seem to have the highest proportions. They both apply the household-dwelling concept, but this cannot be the only reason for the high proportions (there are other countries using the same concept but with a lower share of the non-private household population, e.g. Finland). In the metadata of the 2011 Census Hub, it is said that, in Sweden, people who cannot be linked to a dwelling cannot form a household and are classified as 'people not living in a private household, but category not stated'. This category consists of 322 001 people in Sweden, which is the highest value in the cen-

([37]) The number of asylum seekers increased sharply in many European countries in 2015 because of the crisis in Syria.

([38]) The number of people living in non-private households in Switzerland is zero, according to the 2011 Census Hub. According to its notes, the values in this hypercube refer not to the total Swiss population but rather to the resident permanent population aged 15 years or older in private households.

sus data (the second highest is 116 294 people in Greece). However, the metadata of the census do not give any reason for this high figure.

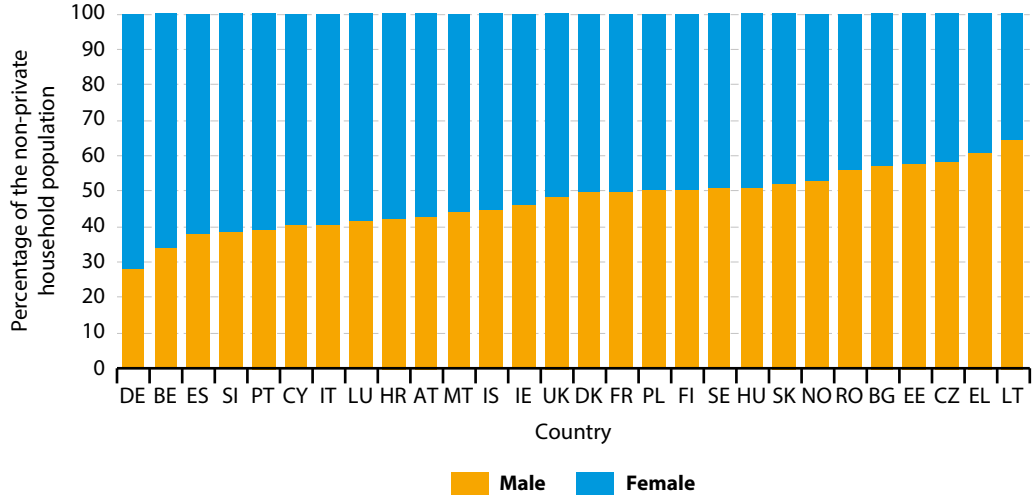
For most of the countries, the non-private household population seems to mainly consist of people living in an institutional household. However, in Ireland and Finland, the number of people living in an institutional household seems to be zero. In these two countries, almost all the people belonging to the non-private household population were classified into the subcategory 'people not living in a private household, but category not stated'. At least in the case of Finland, which applies the household-dwelling concept, the institutional household population was classified into this third subcategory. The total figure for the non-private household population in Finland includes the institutional household population.

3.3.3. Demographics of the non-private household population

Figure 3.2 presents the gender distribution of the non-private household population. In eight countries, the gender distribution is quite uniform, that is, the difference between the share of males and the share of females is less than 5 percentage points (countries from the United Kingdom to Slovakia in Figure 3.2). Some countries in southern Europe, namely Spain, Portugal, Cyprus, Italy and Croatia, seem to belong to a group in which the share of females is about 20 percentage points higher than the share of males. However, in Romania, Bulgaria and Greece, the situation is the opposite: the share of males is about 20 percentage points higher than the share of females. The countries in northern Europe, except for Iceland, seem to belong to the group in which the distribution between the genders is quite uniform.

The gender distribution differences between countries are interesting, but it is important to study the differences between the private and the non-private household populations within the countries. If there are significant differences between those populations, we may conclude that the definition of the target population creates bias towards the survey results compared with the total population (including private and non-private households). If the differences are small, we may suppose that the possible bias is not significant.

Figure 3.2: Gender distribution of people belonging to the non-private household population, 2011



NB: Countries are ranked according to their share of males in the non-private household population. The share of males in Germany was 28 % of the non-private household population in 2011.

Source: 2011 Census Hub data (accessed 25 January 2019).

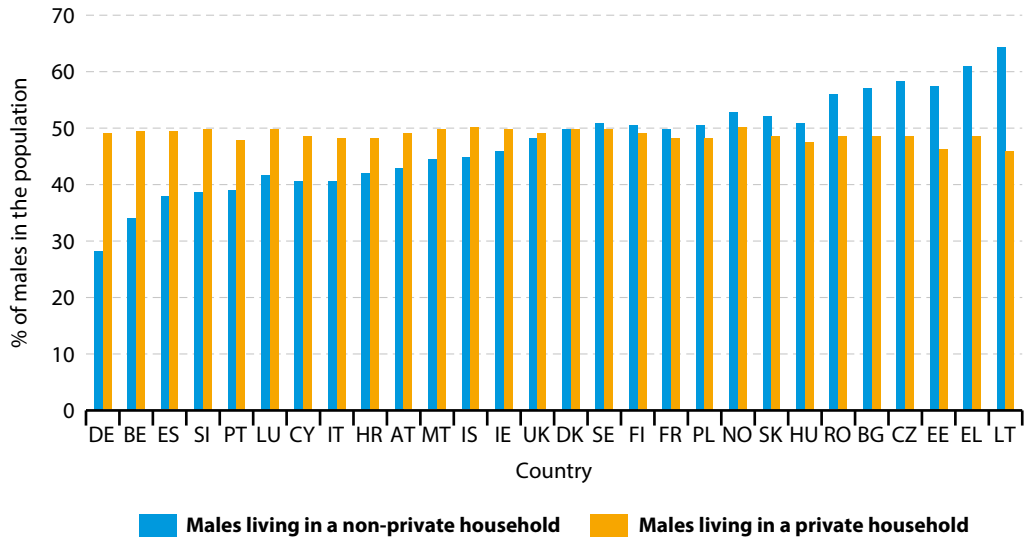
Shares of males in the non-private and private household populations are presented in Figure 3.3. In 10 countries, the absolute value of the difference between the shares of males in the non-private and private household populations is less than 5 percentage points (countries from Ireland to Hungary). In 21 countries, the absolute value of the difference between the non-private and private household populations is less than 10 percentage points (countries from Portugal to Czechia).

There are some countries with substantial differences in the gender distribution between the non-private and private household populations. In Germany, there are fewer males in the non-private household population than in the private household population by about 20 percentage points. In Lithuania, the situation is the opposite, that is, there are more males in the non-private household population than in the private household population by about 20 percentage points.

However, even if there seems to be some differences in the gender distribution, we may suppose that these are not significant enough to bias the survey results. For example, in Germany in 2011 the share of males in the private household population was 48.9 %, and in the total population, including non-private and private households, it was 48.7 %. In Lithuania, the corresponding figure for males in the private household population was 45.9 %, and in the total population it was 46.1 %. Thus, in Germany the inclusion of the non-private household population in the target population would result in a decrease of 0.2 percentage points in the share of males. The corresponding figure for Lithuania would be an increase of 0.2 percentage points. Thus, even though according to Figure 3.3 differences in the distribution of males were quite large, the consequent coverage bias is relatively small. This stems from the fact that, in almost all the countries, the share of the non-private household population was less than or equal to 2 % of the total population in 2011.

The mean age of people living in a non-private household was 53 years in 2011. The corresponding figure for the private household population was 40 years. The classified age distribution of the non-private household population is presented in

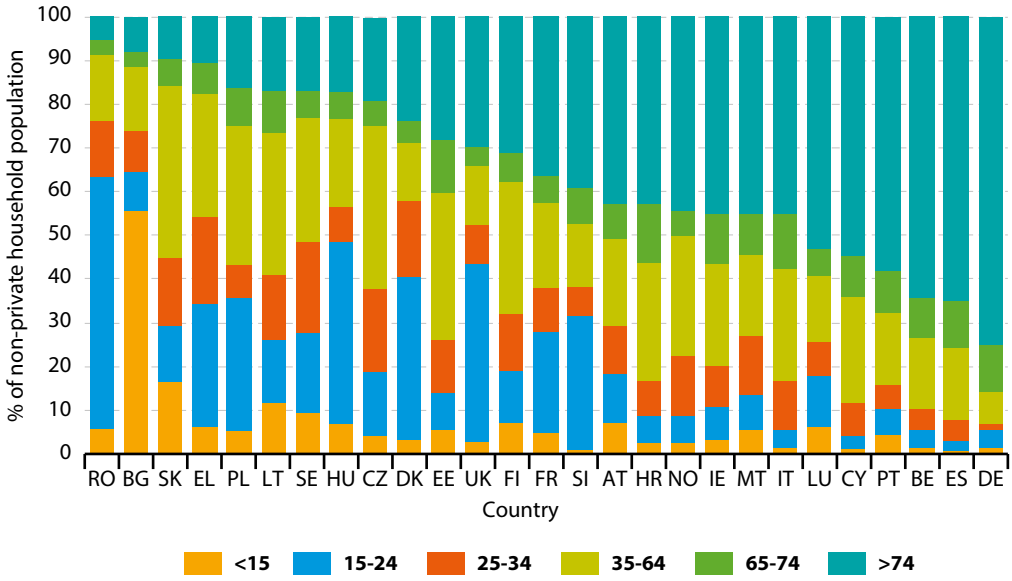
Figure 3.3: Shares of males in the private and non-private household populations, 2011



NB: Countries are ranked according to their difference between the share of males in the non-private household population and the share of males in the private household population. In non-private households in Germany in 2011, 28 % of persons were male, whereas in private households in Germany in 2011, 49 % were male.

Source: 2011 Census Hub data (accessed 25 January 2019).

Figure 3.4: Classified age distribution of the non-private household population, 2011



NB: Countries are ranked according to their total share of people aged 75 years or over in the non-private household population. In Romania, the share of people aged 75 years or over was 5 % of the total of non-private household population in 2011.

Source: Author's own computation using 2011 Census Hub data (accessed 25 January 2019).

Figure 3.4. There are 15 countries in which the share of people aged under 65 years in the non-private household population is greater than 50 % of the non-private household population, that is, the share of children and younger people is significantly higher than the share of elderly people ([39]). In 12 countries, the situation is the opposite, that is, the share of elderly people is significantly higher (50 % or more) than the share of children and younger people. We may conclude that there are significant differences in the age distribution of the non-private household population between countries.

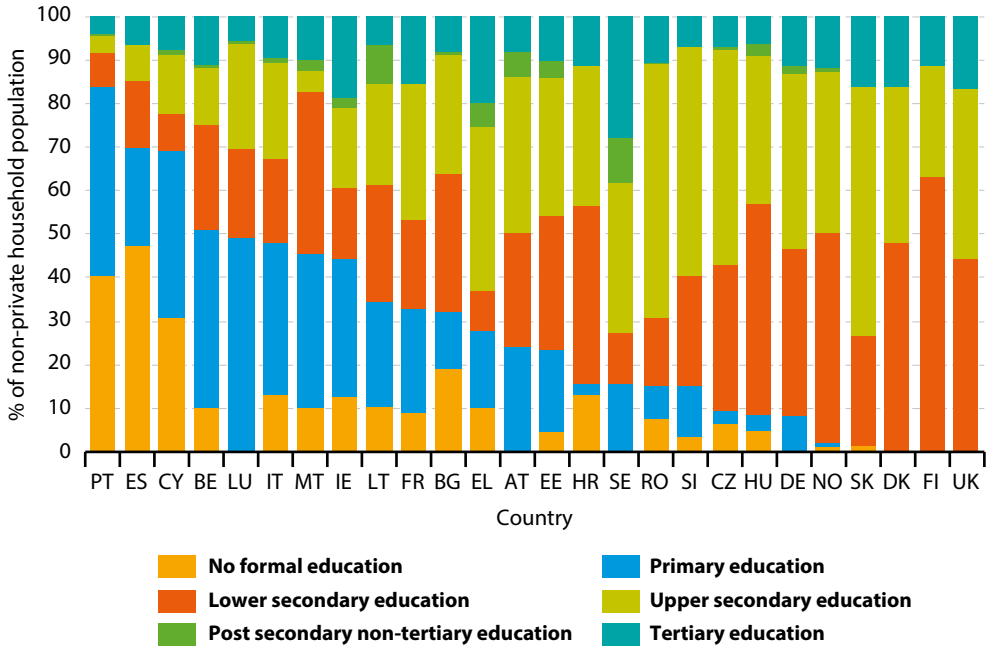
We next turn our attention to the education level of the private and non-private household populations. In most of the European countries, the education level of the private household population was upper secondary education in 2011 ([40]). In the non-private household population, the European-wide education level distribution seems to vary more than in the private household population. In 10 countries, the education level was lower secondary education, and in 10 countries the education level was upper secondary education. In five countries, the education level was primary education or no formal education at all. Hence, it seems that the education level of the non-private household population could be somewhat lower than that of the private household population.

The education level distribution of the non-private household population is presented in Figure 3.5. There are eight countries in which the share of people having no formal education or having primary education is above 40 % of the total non-private household population. On the contrary, in Sweden the share of people having tertiary education is 28 % of the non-private household population.

([39]) In Bulgaria and Romania, the share of people aged under 25 years is over 60 % of the total non-private household population. It could be that this phenomenon is related to military conscription (Romania) or living conditions among children (Bulgaria). For example, in Bulgaria the share of children at risk of poverty or social exclusion aged less than 16 years was 33.4 % in 2018 according to Eurostat [table ilc_peps01]. The corresponding figure for the EU was 23.8 %.

([40]) The computations have been made using the 2011 Census Hub data. The 'not applicable' category for individuals aged under 15 years was not included in the computations.

Figure 3.5: Education level distribution of the non-private household population, 2011



NB: Countries are ranked according to their total share of people classified into the groups 'no formal education' and 'primary education'. In Portugal, the share of people having no formal education was 40 % of the total non-private household population in 2011.

Source: Author's own computation using 2011 Census Hub data (accessed 25 January 2019).

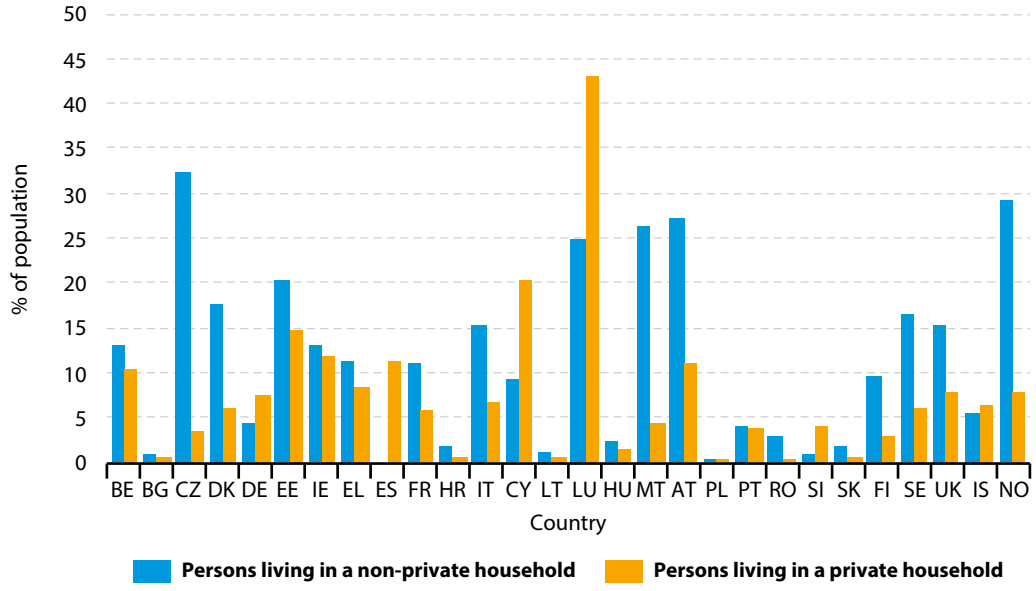
As in the abovementioned discussion concerning age distribution, we may conclude that there are differences in the education level distribution of the non-private household population between the countries.

Thus far, we have concluded that there are some significant differences in the gender, age and education level distributions between the private and the non-private household populations. The final subject for our study is the citizenship of people living in non-private and private households. In general, most of the people living in a private household have citizenship of the reporting country. There are only two countries in which the proportion of people who are not citizens of the reporting country is greater than or equal to 20 % of the private household population, namely Cyprus and Luxembourg (see Figure 3.6).

In the non-private household population, there are more people with foreign citizenship. In six countries, the share of people having citizenship of a foreign country is greater than or equal to 20 %. This suggests that immigration may have an effect on the differences between the private and the non-private household populations. However, there are 14 countries in Figure 3.6 in which the absolute difference between the private and the non-private household populations is less than 5 percentage points, that is, not significant.

The descriptive analysis shows differences in gender, age distribution, education level and citizenship between and within private and non-private households across countries. We will continue the discussion about the effects of including the non-private household population in the EU-SILC target population in Section 3.5, which concludes this chapter. However, it is already evident that more detailed and up-to-date information is needed before making any recommendations.

Figure 3.6: People with foreign citizenship in the non-private and private household populations, 2011



NB: In Belgium, 10 % of people living in a private household had foreign citizenship in 2011.

Source: 2011 Census Hub (accessed 25 January 2019).

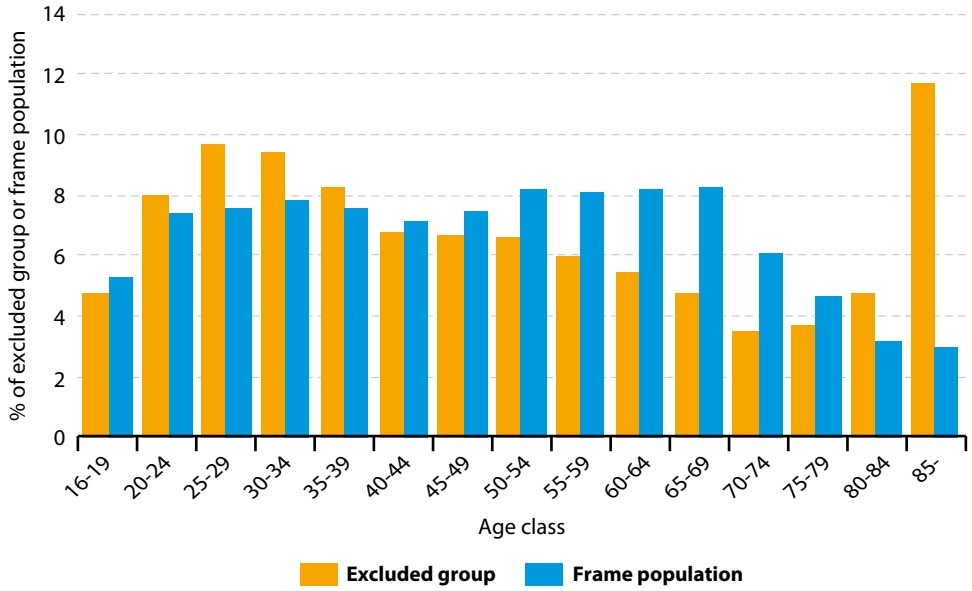
3.4. Case study: under-coverage induced by the non-private household population in the Finnish EU-SILC

This section presents a case study that (i) illustrates an approach that countries could use to assess the effect of the target population definition on coverage, and (ii) provides an analysis of the effects of the definition on more detailed and pertinent variables than those available from the Census Hub and using more up-to-date data, including on income and risk of poverty.

The sampling frame for the Finnish EU-SILC is the population register. The population register includes information about the household-dwelling units based on the household-dwelling concept of the population census. Using the population register information, we can compare the non-private household population in Finland with the private household population. The analysis is done purely with register data. The only weakness in the analysis is that the household-dwelling concept may, in practice, differ slightly from the private household definition of EU-SILC. However, with register data we do not have to worry about small sample sizes, and we are able to study the demographics in more detail.

The target population of the Finnish EU-SILC consists of private households permanently living in Finland at the end of the income reference year (31 December). At the end of 2016, there were 4 550 794 residents in Finland aged 16 years or older according to Statistics Finland's population structure statistics. This group is referred to as the total population. The frame population from which the EU-SILC sample was selected consisted of 4 474 994 people. Thus, there were 75 800 people (1.67 % of the total population) excluded because

Figure 3.7: Age distribution of people belonging to the frame population and to the excluded group in the Finnish EU-SILC 2017 sampling frame



NB: In the group excluded from the frame population, about 5 % of the people are aged between 16 and 19 years.
Source: Author's own computation using data from Statistics Finland.

of the target population definition ([41]). This group is referred to as the excluded group ([42]).

The age distributions of the frame population and the excluded group are presented in Figure 3.7. The share of people aged 80 years or more is clearly bigger among the excluded group than among the frame population. The total number of people aged 80 years or more among the excluded group is 12 443, that is, about 17 % of the total. This is likely to be the result of residency in institutions for care and/or for the elderly.

Another minor peak in the age distribution of the excluded group seems to be between 25 and 34 years. About 19 % of the excluded group are aged between 25 and 34 years, whereas the corresponding share in the frame population is about 16 %. About 43 % of the people belonging to the excluded group are outside the labour force or students. These people could be asylum seekers or perhaps students who are temporarily resident in Finland. About 58 % of the people belonging to the excluded group are male. The corresponding share in the frame population is significantly lower, at 49 %.

The share of immigrants is clearly higher in the excluded group. Most (93 %) of the people belonging to the frame population are born in Finland and have parents who are also born in Finland. In the excluded group, about 77 % have this kind of Finnish origin. About 21 % of the people belonging to the excluded group have foreign parents and are also born abroad. In the frame population, the corresponding share is only about 6 %.

There are also major differences in the distribution of education level between the excluded group and the frame population. About 31 % of people belonging to the frame population have a university-level degree, compared with about 15 % in

([41]) The excluded group consisted of those without a permanent address, the institutional population (for example those living in old people's homes, care institutions, prisons or hospitals in the long term), asylum seekers and those temporarily resident in Finland.
([42]) Because the first sampling phase of the Finnish EU-SILC is restricted to people aged 16 years or older, we study here only the demographics of the corresponding total population.

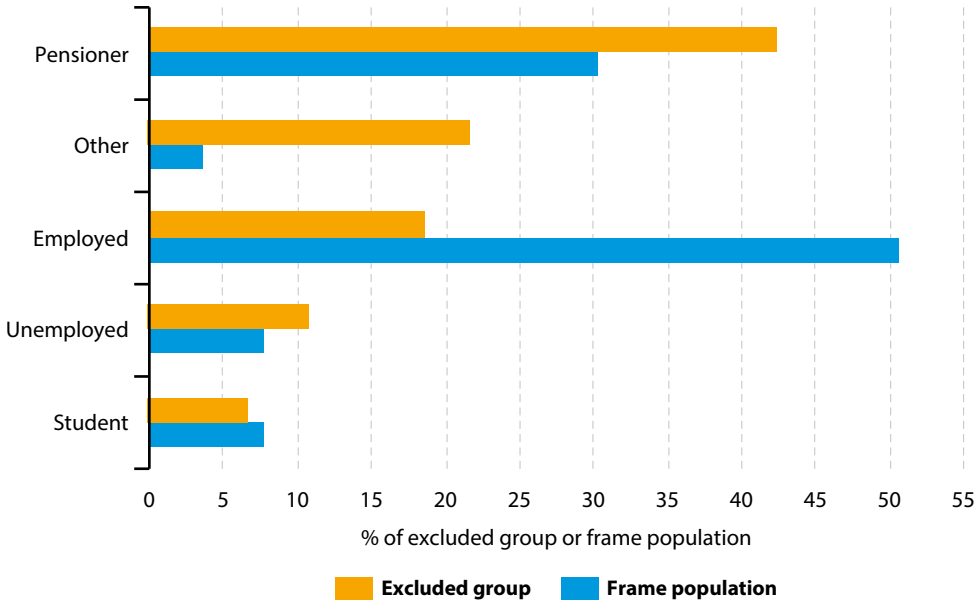
the excluded group. However, the lower share of university-level degrees may be a result of missing education information. Unfortunately, it is not possible to distinguish between missing education information and the lowest level of education because of a lack of information in the source data set. In the excluded group, the share of people having either missing education information or the lowest education level is about 56 %. In the frame population, the corresponding share is much smaller – about 27 %.

It seems that the excluded group consists mainly of pensioners and other people outside the labour force (see Figure 3.8). About 42 % of people belonging to the excluded group are classified as pensioners and 22 % as being outside the labour force. In the frame population, only 30 % of people are pensioners and 4 % belong to the mixed group outside the labour force. Thus, the share of employed people is much smaller in the group of excluded people (19 %) than in the frame population (51 %).

Next, we study the income distribution of the people belonging to the excluded group. To do this, we have linked register data on personal income from administrative sources to the frame population with the help of personal identity codes. About 16 % of the people belonging to the excluded group had zero personal disposable net monetary income. This is a much higher share than in the frame population, in which about 2 % of the people had zero personal disposable net monetary income.

According to Table 3.2, the median and mean personal disposable net monetary incomes are both smaller in the excluded group than in the frame population. In fact, all the statistical figures presented in Table 3.2 are smaller for the excluded group than for the frame population. However, if we look at the income components at a more detailed level, it can be seen that the excluded group has, on average, higher current transfers received than the people belonging to the frame population. For example, the median value of current transfers re-

Figure 3.8: Main activity status of the people belonging to the frame population and to the excluded group in the Finnish EU-SILC 2017 sampling frame



NB: In the group excluded from the frame population, about 42 % of the people are pensioners.
Source: Author's own computation using data from Statistics Finland.

Table 3.2: Distribution of personal disposable net monetary income in 2016 (EUR)

Population	p10	p25	Median	Mean	p75	p90
Excluded group	0	5 694	14 182	14 884	19 308	26 182
Frame population	8 013	13 750	20 663	23 370	28 774	38 415
Total population	7 731	13 653	20 504	23 229	28 653	38 724

NB: p10 indicates 10th percentile, etc.

Source: Author's own computation using data from Statistics Finland.

ceived for the excluded group was EUR 9 202 compared with a median value of about EUR 4 678 for people belonging to the frame population in 2016.

There is some evidence that the higher share of current transfers received among the non-private household sector is not related only to the year studied here (2016). Törmälehto (2019) found in his study that the share of current transfers received and property income were higher for non-private households than for private households at the end of 2014. The result could be explained by the fact that these components are more age dependent and received relatively more by the elderly.

Table 3.2 also includes income information for the total population, namely the population including the frame population and the excluded group. It seems that the exclusion of the non-household (institutional etc.) population from the total population does not affect the income level significantly. For example, the median value of net monetary income was only EUR 159 larger for the frame population than for the total population. It seems that the coverage bias created by the exclusion of the non-private household population is quite small, at least in the case of personal disposable net monetary income.

For EU-SILC analysts, the concept of equivalised income – which takes household composition into account – may be of more interest than personal disposable income. Here, we use the equivalised income of the register-based dwelling units as a proxy for household equivalised income. For the excluded group, we have no dwelling size or income information available. Hence, we presume these dwelling units to be one-person dwellings having equivalised income that is equal to personal disposable net monetary income. In the frame population, the at-risk-of-poverty rate is 0.32 percentage points lower than in the total population including the excluded group (Table 3.3). There are about 25 000 more people at risk of poverty in the total population than in the frame population.

3.5. Conclusions

In this chapter, we studied the possible bias caused by the exclusion of the non-private household population from the EU-SILC target population using 2011 Census Hub data and a more detailed case study for Finland. The non-private household population seems to consist mostly of people living in institutional households. In most of the Member States, the people living in a non-private household were older and had a lower education level than the people living in a private household.

Table 3.3: At-risk-of-poverty indicator for people aged 16 or over and disposable net monetary income per consumption unit in 2016

Population	Median (EUR)	Mean (EUR)	People at risk of poverty	At-risk-of-poverty rate (%)
Frame population	24 537	27 677	601 886	13.45
Total population	24 369	27 464	626 533	13.77

Source: Author's own computation using data from Statistics Finland.

In general, the proportion of the population living outside private households was quite small: in most countries, it was less than or equal to 2.5 % of the total population. This implies that, even though there may be significant differences between the demographic structures of the private and non-private household populations, the coverage bias created by the exclusion of non-private households from the EU-SILC target population is likely to be modest.

A more detailed analysis of the income structure of the non-private household population could be carried out only with the Finnish EU-SILC 2017 sampling frame. The income level was significantly lower for the excluded non-private household population than for the frame private household population. However, the excluded group received, on average, more current transfers than the frame population. This could relate to the fact the excluded people are older, and a large proportion of them are classified as pensioners.

For social statistics measuring income and living conditions-related concepts, for example at risk of poverty, the most difficult groups to survey may have an income distribution different from the total population. This is also highlighted in the United Nations Economic Commission for Europe (UNECE) *Guide on Poverty Measurement*. The guide points out that poverty is usually more prevalent among hard-to-reach groups, and hence all of the population or subpopulation of interest should be included in poverty statistics (UNECE, 2017).

The UNECE *Guide on Poverty Measurement* also gives a recommendation to national statistical institutes to explore the feasibility of extending the coverage of poverty statistics from private households to the total population (UNECE, 2017). In the case study based on the Finnish population register, the inclusion of the non-private household population in the current frame population resulted in an increase of about 25 000 people at risk of poverty and a decrease of EUR 168 in the median equivalised disposable income of the dwelling units. There is a clear need to repeat this exercise for other EU-SILC countries in order to establish the desirability of extending EU-SILC coverage to the non-household population, bearing in mind that this extended coverage would come at a cost.

There are already some surveys collecting information from specific subpopulations. For example, the Survey of Health, Ageing and Retirement in Europe (SHARE) includes information about the health, socioeconomic status and social and family networks of individuals aged 50 or older, covering 28 European countries and Israel (Bergmann et al., 2017). The Second European Union Minorities and Discrimination Survey (EU-MIDIS II) collected comparable data in all 27 EU Member States and the United Kingdom on experiences of discrimination in different areas of life (labour market, education, housing, health and other services) and social inclusion (FRA, 2017).

SHARE does not include people living in prisons or hospitals or who are out of the country during the entire survey period; that is, the non-private population is at least partly excluded from the frame population (Bergmann et al., 2017). In EU-MIDIS II, the sampled individuals had to be living in private households in the EU Member State surveyed for at least the past 12 months. However, in a small number of countries people living in a non-private household were also included (e.g. Malta) in order to completely cover the target population (FRA, 2017). Thus, SHARE and EU-MIDIS II are examples of surveys having a specific target population, but neither of them includes the whole of the non-private population in its target population.

The inclusion of the relatively small group of people not in private households in the current EU-SILC target population definition may require Member States to apply a targeted sampling strategy to obtain enough observations from this group of households. For example, it is highly probable that the non-response rate among the non-private household group would be significantly larger than among the private household population. The fieldwork material and survey questionnaire should also be designed more carefully to better take into account this special group.

It should be noted that EU-SILC is an output harmonised instrument, with current data collection modes ranging from paper and pencil interviewing to web questionnaires. Hard-to-reach subpopulations may require use of specific survey tools, which may not be possible in all EU-SILC countries. Collecting comparable data in a decentralised cross-national survey is more challenging than

in centralised input harmonised surveys such as SHARE and EU-MIDIS II.

To be able to include the non-private household population in the EU-SILC target population, we would need to identify this group among the total population. Currently, some Member States exclude the non-private household population before the fieldwork, when they create the actual sampling frame. These countries could, in theory, include the non-private household population in their sampling frames. However, the countries explicitly reporting exclusion of the non-private household population from the frame were in the minority. It seems that, at European level, many sampling frames lack sufficient information or coverage to be able to include the non-private household population.

The findings reported in this chapter demonstrate that we cannot assume that coverage bias due to the EU-SILC target population definition is negligible, nor that it is consistent across countries. However, we still know relatively little about the nature and extent of population under-coverage in different countries. Further research using detailed and up-to-date data is needed to confirm the magnitude and significance of the bias created by the exclusion of the non-private household population from the EU-SILC target population. The census data used in this chapter are somewhat out of date, but the census is the only European-wide source for studies concerning the non-private household population. However, when data from the next census round in 2021 are available, they should be used to repeat the demographic studies in Section 3.3.3 to see to what extent the situation has changed or remained the same. This is a task that should be quite easy to perform, and we recommend that it is carried out as soon as possible. Additional case studies from other countries, using register data that provide more detail than the census, would also be informative.

References

Bergmann, M., Kneip, T., De Luca, G. and Scherpenzeel, A. (2017), 'Survey participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), wave 1–6', *SHARE Working Paper Series*, No 31-2017 (http://www.share-project.org/fileadmin/pdf_documentation/Working_Paper_Series/WP_Series_31_2017_BergmannKneip.pdf).

CBS (2012), 'Dutch Caribbean population exceeds 21 thousand', 3 January (https://www.cbs.nl/en-gb/news/2012/01/dutch-caribbean-population-exceeds-21-thousand).

FRA (European Union Agency for Fundamental Rights) (2017), *Second European Union Minorities and Discrimination Survey – Technical report*, FRA, Vienna (https://fra.europa.eu/en/publication/2017/eumidis-ii-technical-report).

Gabler, S. and Häder, S. (2016), 'Special challenges of sampling for comparative surveys', in Wolf, C., Dominique, J., Smith, T. W. and Fu, Y. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London, pp. 346–356.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau R. (2009), *Survey Methodology*, 2nd edition, Wiley, Hoboken, NJ.

Heeringa, S. G. and O'Muircheartaigh, C. (2010), 'Sampling designs for cross-cultural and cross-national survey programs', in Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. Ph. et al. (eds), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, Wiley, Hoboken, NJ, pp. 251–268.

Lynn, P., Häder, S., Gabler, S. and Laaksonen, S. (2007), 'Methods for achieving equivalence of samples in cross-national surveys: the European social survey experience', *Journal of Official Statistics*, Vol. 23, No 1, pp. 107–124.

Törmälehto, V.-M. (2019), 'Reconciliation of EU statistics on income and living conditions (EU-SILC) data and national accounts', *Eurostat Statistical Working Papers*, Publications Office of the European Union, Luxembourg (https://ec.europa.eu/eurostat/documents/3888793/9959642/KS-TC-19-004-EN-N.pdf/cd90cd0f-ebcf-43b6-ab4b-c28bcdda4f46).

UNECE (United Nations Economic Commission for Europe) (2017), *Guide on Poverty Measurement* (https://unece.org/DAM/stats/publications/2018/ECECESSTAT20174.pdf).

4 Frame errors in EU-SILC: under-coverage

Tara Junes ([43])

4.1. Introduction

The previous chapter discussed under-coverage of the de facto population due to the restrictive European Union Statistics on Income and Living Conditions (EU-SILC) target population definition. This chapter discusses a separate source of under-coverage, namely under-coverage of the target population due to units being missing from the sampling frame.

EU-SILC is designed to produce statistics about income, social inclusion and living conditions, covering topics such as poverty and social exclusion. In this kind of European-wide survey, the coverage and quality of the sampling frame is important from the policy monitoring point of view. Sampling frames should represent comparable target populations, and indicators based on samples from those sampling frames should not reflect different populations in different countries. Some types of population elements missing from the sampling frame may be particularly likely to induce bias for the topics covered in EU-SILC. For example, missing information about the homeless and illegal immigrants could have an effect on the social indicators of EU-SILC.

In this chapter, Section 4.2 first provides a short introduction to the terminology of sampling frames. Section 4.3 discusses sampling frame types, methods used to construct the sampling frames and update frequency of the sampling frames in the EU-SILC countries. It seems that most of the countries use population census or population register data in the construction of EU-SILC sampling frames, but the frame update frequency varies somewhat between countries.

The sampling frame coverage rates and possible reasons for the underlying under-coverage are discussed in Section 4.3.2. The analysis and main findings are based on data collected for the ESSnet KOMUSO project on quality of multisource statistics under work package 2 (Quality measures and indicators of frames for social statistics) and information collected from the *Comparative EU Quality Report 2016* (Eurostat, 2018). Most of the reported frame under-coverage rates are quite modest. The main underlying reason for frame under-coverage seems to be related to immigration. However, the information collected covers only a few EU-SILC countries. More detailed information is needed to draw European-wide conclusions regarding the extent and structure of frame under-coverage.

The under-covered target population units are quite hard to study. If we knew their characteristics, we would include them in the target population. However, in some cases the sampling procedure may create frame under-coverage of a kind that we are able to study. Typically, samples are selected some months before the start of data collection in order to allow time for fieldwork planning and organisation. In Section 4.4, the Finnish EU-SILC 2017 sampling frame under-coverage due to the time of the sample selection is discussed. The Finnish EU-SILC sample is selected in November, although the reference tar-

([43]) Tara Junes is with Statistics Finland. The author is obliged to Lionel Viglino for sharing the ESSnet KOMUSO project data with the Net-SILC3 project. The author wishes to thank Rudi van Dam, Anne-Catherine Guio, Tarja Hatakka, Lars E. Lyberg, Peter Lynn, Eric Marlier and Veli-Matti Törmälehto for their valuable comments and suggestions. All errors are strictly the author's responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Tara Junes (tara.junes@stat.fi).

get population is the private household population at the end of the year. This section finds that about 12 000 target population members were excluded from the sampling frame at the time of the sample selection. Section 4.5 concludes by discussing the findings of this chapter and their implications.

4.2. Definition of sampling frames and under-coverage

One of the main concerns in statistics is making conclusions about a well-defined population using a sample survey. Following the terminology presented by Groves et al. (2009, p. 69), this kind of population consists of fundamental units referred as 'elements'. Elements may be different kinds of units, but in most household populations they are people living in the households.

The group of elements about which survey information is wanted forms the 'target population' of finite size. As stated by Groves et al. (2009, p. 69), the elements of the target population have to exist within a specified time frame, and they have to be observable. Sometimes, it is not possible to collect the information from the intended target population because of, for example, data collection operation limits. The actual more restricted population from which the survey information is collected is called the 'survey population'.

A list of target population units used for drawing the sample is defined as the 'sampling frame'. The sampling frame determines how well the target population is covered and affects the choice of the data collection method. Ideally, the frame would contain every unit of the target population and some auxiliary information. As listed by Groves et al. (2009, chapter 3), frames may consist of very different types of population units, from maps of areas to time periods during which target events would occur. The sampling frame can also be a simple list of population elements; for example, a number of countries use registries of addresses or of people as sampling frames (Groves et al., 2009, chapter 3).

The types of sample design and estimation procedures that are possible are defined by the structure of the sampling frame, the information it contains and the quality of that information. A more complex sampling design requires a complex sampling frame with good auxiliary information. From the simplest type of frame list just clearly identifying each element of the target population, a simple random sample may be selected. Lessler and Kalsbeek (1992) point out that the nature of the target population should determine the type of frame. In reality, it is the other way round; that is, the nature of the frame available determines the survey population and even the target population.

The frame has to identify the sampling units and distinguish them clearly from each other. To create a proper sampling frame for a survey, we also need clear rules of association linking each population element to a sampling unit. A target population unit is labelled as 'covered' by the frame if it is included in the frame. Sampling frame under-coverage is an error created by the missing 'under-covered' target population units. The source of the coverage error is the sampling frame itself. It is therefore important to assess the quality of the sampling frame and its completeness for the target population. As noted by Groves (2004, chapter 3), coverage error also has to be considered in the case of a census of the target population, that is, not only when discussing sample surveys. Missing units from the frame materials will be missing from the census-based and survey-based statistics alike.

As mentioned already in the introductory section, frame under-coverage problems are difficult to identify and solve. The coverage problems are, however, well known, and some suggestions have been made for how to reduce them, at least to some extent. As Groves et al. (2009, pp. 88–90) point out, missing population units could be identified using additional frames or by supplementing the frame with different techniques. Groves et al. (2009, pp. 90–91) describe, for example, 'multiplicity sampling' in which population elements are added to the population through network sampling.

Sampling frames are important topics to study, because the inference made from survey data depends heavily on the quality of the sampling frame. In the next section, the construction of EU-SILC sampling frames and their known under-coverage related to the target population are discussed.

4.3. Under-coverage in the EU-SILC sampling frames

This section discusses the coverage of the EU-SILC sampling frames. It is an important but quite difficult topic to study using the available EU-SILC quality report information. To be able to present any detailed conclusions, more information was required from the EU Member States. Fortunately, this task had already been carried out in the ESSnet KOMUSO project. In Sections 4.3.1 and 4.3.2, data collected in the ESSnet KOMUSO project are supplemented with information collected from the *Comparative EU Quality Report 2016* (Eurostat, 2018) ([44]).

An important topic to study is how the overall coverage of the sampling frames is assessed. For instance, the ESSnet KOMUSO questionnaire included a question about whether the statistical quality of the sampling frame source is assessed or not. However, there were no follow-up questions about the methods used for the possible statistical evaluation. Hence, we end up in a situation in which countries may say that their sampling frame coverage is good, but we seem to have no information about the criteria used to justify this kind of argument. The following subsections try to give an overview of the reported under-coverage issues, but more harmonised indicators are needed to properly analyse the quality and coverage of the EU-SILC sampling frames.

4.3.1. Sampling frames in the EU-SILC countries

Sampling frames of household units may consist of very different types of lists, ranging from population registers to address- or area-based lists. Each of them has its own implications for sampling frame coverage. For example, population registers generally exclude illegal immigrants, whereas address lists do not (at least when the illegal immigrants are not living under the open sky). However, address lists may exclude residential addresses at which a business is also run or at which the residents collect mail from a postal box rather than have it delivered.

There is no strict regulation about the sampling scheme to be applied in EU-SILC. However, some countries have access to registers and use them as a source for the collection of income and other data. These are referred to as register countries ([45]). For them, Regulation (EC) No 1177/2003 allows the use of a sample of people rather than a sample of complete households in the interview survey.

All the EU-SILC countries reported using either a population register or the population and housing census in the construction of their sampling frames. However, many countries reported also using other administrative registers (e.g. income tax register, building register, social security register) or other information in the sampling frame construction phase (Table 4.1). The number of registers used varied greatly; for example, some countries reported using more than three register sources plus the census or population register data.

One typical source of other information was a list of dwelling addresses, which were matched or linked with the population information. However, no countries reported using an address list as the sole source for the construction of the sampling frame (this also applies to the United Kingdom, which reported using a population census and postcode address file to construct the sampling frame). Households and dwellings were sampled using both dwelling information and addresses. In most of the countries, the list of sampling units also included geographical variables.

Table 4.1: The number of EU-SILC countries tabulated by sources used in construction of the sampling frame

Population information	Use of other administrative registers or other information	
	No	Yes
Population and housing census	8	7
Population register	10	6

Source: ESSnet KOMUSO project material and *Comparative EU Quality Report 2016* (Eurostat, 2018).

([44]) A total of 28 EU-SILC countries provided answers concerning the sampling frames in the ESSnet Komuso project. With the help of the Eurostat quality documentation, the data were extended to cover 31 countries. The countries included in EU-SILC but excluded from the study in Section 4.3 are Iceland, Montenegro, North Macedonia and Turkey.

([45]) The register countries are Denmark, Finland, Iceland, the Netherlands, Norway, Slovenia and Sweden.

One important issue related to the coverage of the frame is the frame revision or update frequency. A total of 28 countries reported, in the ESSnet KOMUSO project, how often they update or revise the sampling frame (Table 4.2). Three quarters of them revise or update their sampling frame continuously, monthly, quarterly or annually ([46]). However, a significant number of countries also reported updating the frame less frequently than once a year. A sampling frame can become out of date quite quickly without regular updates. In particular, regular immigration between countries makes it important for sampling frames to be kept up to date. Otherwise, we end up in a situation in which some people are duplicated in the sampling frames of different countries or are missing from the country where they currently live.

Almost all countries reported that the sampling frame covered the entire national territory of the country.

Some countries reported clearly that they excluded the non-private household population from their sampling frame. It seems that the exclusion is often based on the type of address information. If the address of the sampling unit was recognised as belonging to a prison, hospital, nursing home, collective household, etc., the unit would be excluded from the final sampling frame. For this sort of exclusion procedure, it is important to keep the sampling frame up to date. The effect of excluding the non-private household population was studied in Chapter 3.

Table 4.2: Sampling frame revision or update frequency in EU-SILC countries

How often is the sampling frame revised or updated?	Number of countries
Continuously	8
Monthly	2
Quarterly	6
Annually	5
Other	4
No updates	3

Source: ESSnet KOMUSO project material.

([46]) The update or revision process is not necessarily done by the national statistical institute.

4.3.2. Under-coverage of EU-SILC sampling frames

Only eight countries gave an estimate of the under-coverage rate of their EU-SILC sampling frame, namely an estimate of the proportion of people missing from the frame relative to the total population. The frame under-coverage rates are presented in Table 4.3. Three countries reported having a frame under-coverage rate of less than 1 %. Germany reported the highest frame under-coverage rate, that is, having 10–15 % of the population under-covered by the sampling frame (in Germany the sampling frame is already a result of sampling, complicating studies of the frame under-coverage).

There is obviously a lot of variation in the estimated under-coverage rates. Unfortunately, we have no detailed information about the methods countries used to estimate their frame under-coverage rate. Most of the countries described the distribution of the under-coverage very briefly, that is, with one sentence stating that the frame under-coverage is or is not randomly distributed. Only one country reported that its frame under-coverage estimate was a result of a specific enumeration survey carried out after the 2011 census.

In total, 10 countries reported studying the under-coverage of their sampling frame. The most common reason reported for under-coverage was under-coverage of people with a foreign background, namely foreign students, diplomats, asylum seekers and illegal immigrants. Some countries also reported that some people with a low income or with a low education level were missing from the sampling frame.

Table 4.3: The frame under-coverage rates for EU-SILC

Country	Percentage
Belgium	1.00
Cyprus	1.97
Finland	0.50
Germany	10–15
Italy	1.50
Portugal	3.70
Sweden	0.10
Switzerland	0.47

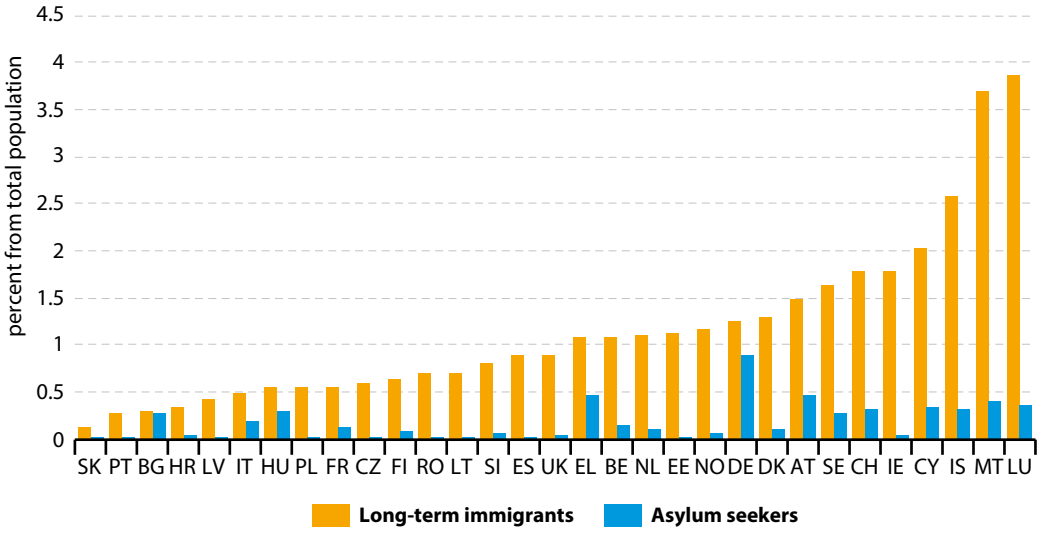
Source: ESSnet KOMUSO project material.

The lack of people with a foreign background implies a systematic under-coverage of the EU-SILC population, as the target population is all residents, regardless of nationality. Some countries reported that delays in obtaining information or lack of precise information about the immigration process were the main reasons for the under-coverage. This usually concerns countries using a population register as a sampling frame and not countries using address-based sampling frames.

We can get an idea about the magnitude of the under-coverage created by the immigration process by looking at the number of immigrants and asylum seekers arriving in the EU-SILC countries in a particular year. The share of immigrants and asylum seekers arriving in 2016 varied greatly between countries (Figure 4.1). However, the share of immigrants seems to be quite modest for every country. This suggests that frame under-coverage caused by the delayed immigration process affects only a small part of the population and is therefore likely to have only a modest effect on estimations concerning the total population. This source of under-coverage could be reduced by making the time gap between sample selection and the end of fieldwork shorter. With a shorter time gap, the frame would be more up to date, and consequently there would be less under-coverage of recent immigrants.

Having an out-of-date frame can also introduce other forms of under-coverage. For example, people aged 16 tend to be under-represented in register-based samples. This is because a sample of people aged 16 or over is selected. Even if the sample includes people aged 16 on the date when the fieldwork is planned to start, rather than the date the sample is selected, this will lead to under-coverage in proportion to the length of the fieldwork period, as, during the course of the fieldwork, some sampled 16-year-olds will turn 17, whereas some 15-year-olds (not included in the sample) will turn 16.

Figure 4.1: Share of recent immigrants in the total population in the reference year 2016



NB: The countries are ranked according to their share of recent immigrants (arrivals in 2016) in the population.

Source: Eurostat database tables tps00191, tps00176 and tps00001 (accessed 25 January 2019).

4.4. Case study: under-coverage in the Finnish EU-SILC sampling frame

In Finland, the EU-SILC sampling frame includes people defined as usual residents ([47]) and having a registered municipality of residence. The sampling frame is based on the population information system maintained by Finland's Population Register Centre. The population information system includes basic data on all Finnish citizens and foreign people permanently resident in Finland. If a person is not registered in the population information system, no information about him or her will be available (e.g. asylum seekers with unfinished arrival procedures) (Statistics Finland, 2019). Hence, we cannot study the frame under-coverage by using register information.

However, it is possible to study the frame under-coverage caused by the time of the selection of the sample. The Finnish EU-SILC target population consists of private households permanently living in Finland at the end of the statistical year. The sampling design is two-phase stratified sampling. In the first phase, a so-called master sample is formed by selecting 50 000 target people who are aged 16 or over by means of systematic sampling (Statistics Finland, 2019).

After the selection of the first-phase sample, dwelling unit information is merged with the selected target person information (Statistics Finland, 2019). This is a very laborious phase in which some of the dwelling units are processed manually ([48]). The fieldwork period of the Finnish EU-SILC usually begins at the beginning of January. Because the time-consuming phase of the dwelling unit construction has to be done before the beginning of the fieldwork period, the sample for the Finnish EU-SILC is selected in November. This is done even though the reference period for the target population is the last day of the statistical year (31 December). Hence, there is some under-coverage in the sampling frame caused by the time of the selection of the sample. As the fieldwork period has to begin at the start of January, there are potential frame under-coverage problems arising from the early sample selection time point. In the following discussion, the consequences of this can be seen.

According to Table 4.4 the amount of frame under-coverage created by the time of the selection of the sample is quite small. The total number of people missing from the sampling frame is 12 181 (about 0.3 % of the frame population). It seems that there are slightly more males missing from the frame than females ([49]).

According to Figure 4.2, a peak in the age distribution of the under-covered people occurs between 20 and 34 years. This may be related to immigration and the slow registration process for immigrants.

Table 4.4: Under-covered population versus frame population in the Finnish EU-SILC 2017 sampling frame

Gender	People under-covered by the time of the sample selection		Frame population in November	
	N	Percentage	*N*	Percentage
Male	7 080	58	2 175 085	49
Female	5 101	42	2 288 741	51

NB: The frame population covers sampling units registered in the population information system in November at the time of sampling frame construction. All people born in 2000 and later are included in the sampling frame (i.e. all people aged 16 or over in the income reference year 2016). 'People under-covered' is the group of sampling units registered in the population information system on 31 December but missing from the register at the time of sampling frame construction. There were 7 080 males registered in the population information system on 31 December but not registered at the time of sampling frame construction. The total number of males in the sampling frame in November was about 2.2 million.

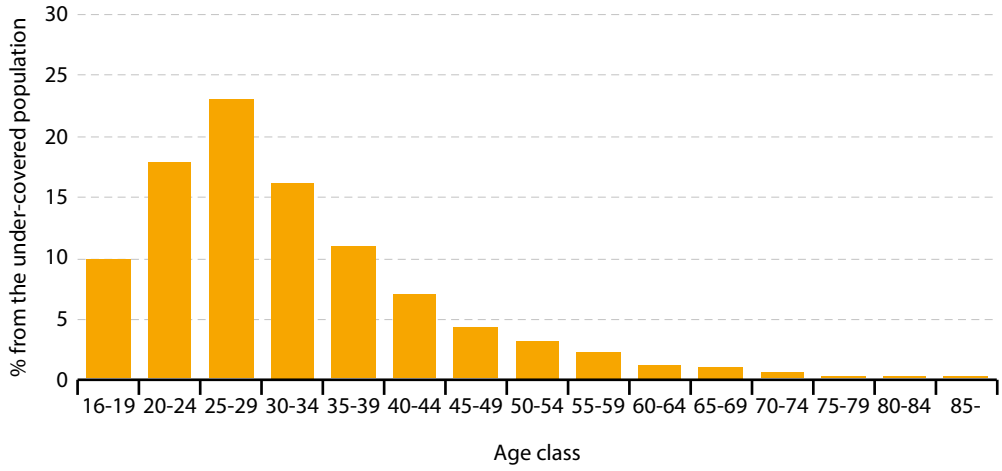
Source: Author's computation using data from Statistics Finland.

([47]) Regulation (EU) No 1260/2013 of the European Parliament and of the Council of 20 November 2013 on European demographic statistics, OJ L 330, 10.12.2013, p. 39: usual residents are defined as (i) those who have lived in their place of usual residence for a continuous period of at least 12 months before the reference time; or (ii) those who arrived in their place of usual residence during the 12 months before the reference time with the intention of staying there for at least 1 year.

([48]) The amount of manual work has substantially decreased in recent years. However, the composition of the biggest dwelling units is still manually processed to make sure the final dwelling household composition equals the household composition as closely as possible.

([49]) The sample frame includes only people aged 16 or over. That is why individuals aged less than 16 years and belonging to the frame population (e.g. newborn babies) are excluded from our study.

Figure 4.2: Age distribution of under-coverage due to early sample selection in the Finnish EU-SILC 2017



Source: Author's computation using data from Statistics Finland.

In fact, in the age classes from 16 years to 49 years more than 90 % of the under-covered people have immigrant parents and are born abroad. The education level of the under-covered people is not very high. About 90 % of the under-covered people have either basic education or belong to a group for which no education information is available.

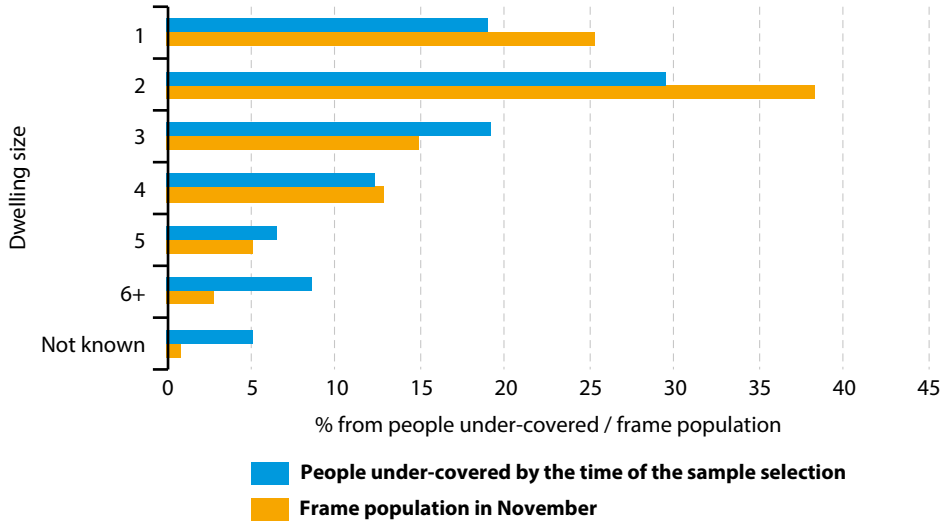
Figure 4.3 shows that the under-covered people live in larger dwelling units than the frame population in November. Most of the people in the frame population, namely 64 %, live in a dwelling of one or two people. The corresponding share among the under-covered people is just 48 %. However, about 9 % of the under-covered people live in a dwelling unit with six or more people, compared with 3 % of the frame population.

The people excluded from the frame are very distinct in terms of the proportion of people living in dwelling units containing people without family ties. About 25 % of the under-covered people live in a dwelling with at least two people without family ties, whereas only 3 % of the frame population are in that situation. Part of this difference may be explained by newly arrived foreign students. About 51 % of under-covered people belonging to this category were aged between 20 and 29 years.

According to Table 4.5, the median and mean personal disposable net monetary incomes are both substantially smaller in the under-covered group than in the frame population. The same is true of the other distributional statistics presented in Table 4.5. The under-covered group clearly has a much lower income level than the frame population when measured by personal disposable net monetary income. The share of people having zero personal disposable net monetary income was 36 % in the under-covered group, whereas the corresponding share in the frame population was around 2 %.

If we look at specific income components, it can be seen that the under-covered group seems to have a significantly lower income level than the frame population for every income component (e.g. factor income, current transfers paid and received). This is not surprising, as most of the people belonging to the excluded group were young and had a low education level compared with the frame population.

Figure 4.3: Dwelling size distribution of excluded people and of the total population in the Finnish EU-SILC 2017 sampling frame



NB: About 19 % of the under-covered people are living in a one-person dwelling. The corresponding share in the frame population is about 25 %.

Source: Author's computation using data from Statistics Finland.

The most important question to ask is how likely it is that the frame under-coverage caused by the early sample selection will create significant bias in sample-based estimates. A comparison between the sampling frame used in the sample selection and the actual targeted end-of-the-year population reveals that the amount of bias seems to be quite small. For example, the median value of disposable net monetary income is only EUR 31 larger for the frame population in November than for the actual end-of-the-year population (see Table 4.5).

The case study presented here of the Finnish EU-SILC sampling frame materials could perhaps be replicated in other EU-SILC countries with a time gap between sample selection and the reference population definition. With the available register information, it was not difficult to compare the sampling frame distribution with the known reference population distribution. However, it may be that this kind of time gap between the reference population definition and the sample selection is quite rare.

Similarly, the effect of the length of the fieldwork period on under-coverage of 16-year-olds could be studied. The prime concern here is that the effect is likely to cause differences between register and

Table 4.5: Distribution of personal disposable net monetary income in the Finnish EU-SILC 2017 sampling frame (EUR)

Population	p10	p25	Median	Mean	p75	p90
People under-covered by the time of the sample selection	0	0	1 493	6 699	10 122	20 913
Frame population in November	8 143	13 785	20 694	23 412	28 797	38 440
Target population on 31 December	8 013	13 750	20 663	23 370	28 774	38 415

NB: p10 indicates the 10th percentile, etc.

Source: Author's computation using data from Statistics Finland.

non-register countries. It could also cause differences between register countries if fieldwork periods differ greatly.

4.5. Conclusions

In this chapter, the frame under-coverage of EU-SILC sampling frames was studied. According to the ESSnet KOMUSO project data, most Member States use the population and housing census or a population register in the construction of their sampling frame. Some countries also supplement the census or register data with additional information (e.g. using other administrative registers). Almost all countries reported that their sampling frame covered the entire national territory of the country. A majority of the countries revise or update their sampling frame at least once a year. Hence, the overall starting point for the sampling seems to be fairly good.

However, only a few countries reported studying the under-coverage of their EU-SILC sampling frame. Eight countries gave an estimate of the possible extent of people not covered by their sampling frame. The one common reason for sampling frame under-coverage was related to immigration. There are, for example, foreign students and asylum seekers who should be included in the private household population but are missing because registers are slow to be updated with information on the immigration process or the information is missing completely.

Existing information about the under-coverage of the EU-SILC sampling frames seems to be scarce. It seems that the material collected in the ESSnet KOMUSO project is the only cross-national source that contains information on the under-coverage of EU-SILC sampling frames at the moment – and this information is rather limited. Without more detailed information about sampling frame coverage, we cannot really conclude whether or not frame under-coverage makes a significant contribution to estimation bias. The possibility certainly cannot be ruled out.

The fact that only eight countries could give an estimate of the extent of frame under-coverage shows the paucity of information about frame quality. To improve the quality of the existing information, more detailed questions about the coverage of the sampling frame could perhaps be included in the EU-SILC quality reporting. At a minimum, all countries should, from time to time, analyse the coverage of their sampling frame using whichever methods are most suitable given the nature of the sampling frame.

Practical strategies for obtaining probability samples of under-covered or hard-to-reach groups seem to be rare. Till-Tentschert, Reichel and Latcheva (2018) recognise the non-existence or poor quality of the sampling frames for certain target groups as a major challenge that survey researchers face. They have listed some suggestions for how traditional sampling methods could be adapted to obtain high-quality samples of special hard-to-reach groups of the population. One suggestion is to use alternative sampling approaches, such as so-called respondent-driven sampling or location sampling.

In respondent-driven sampling, a group of initial respondents is first selected non-randomly (so-called seeds). Each of the initial respondents is then asked to refer one or more members of the target group (Till-Tentschert, Reichel and Latcheva, 2018). For instance, if the objective is to achieve good coverage of newly arrived immigrants, a certain group of them could be approached and asked to refer their associates, who in turn would be asked to refer their associates, until no more new references were obtained. The approach relies on the subpopulation in question being well connected. In location sampling, we first have to collect information on all possible locations where the target group may gather. A random sample of locations is then selected from the complete list of gathering places and surveyed. At each location, the respondents are randomly selected (Till-Tentschert, Reichel and Latcheva, 2018). If we think about newly arrived immigrants, listing all possible gathering places could be quite difficult. However, for a smaller subgroup, such as asylum seekers, this may be a method to at least consider.

Studies concerning sampling frame under-coverage are not easy to perform. In the previous chapter, the frame under-coverage related to the adopted target population definition was discussed and

it was noted that collecting information from the non-private household population would require targeted sampling strategies and redesign of the questionnaires and fieldwork materials. However, the non-private household population relates to a classification made by the national authorities. It is, at least in theory, accessible, whereas the missing population units are not.

The major challenge for future studies is how to collect information about those population units that are completely missing from the sampling frame. It is quite hard to imagine any other way than carrying out a survey using the face-to-face interview method and some specific sampling techniques. Perhaps this could be done after a census round as a post-enumeration survey. The coverage issues should be given more attention, because the inference made from survey data depends heavily on the quality of the sampling frame. As noted by Groves et al. (2009, chapter 3), samples can be no better than the frames from which they are drawn.

References

Eurostat (2018), *Comparative EU Quality Report 2016* (https://circabc.europa.eu/sd/a/6f7191df-c72d-4537-aee4-81972188497d/2016%20EU%20SILC%20ESQRS.zip).

Groves, R. M. (2004), *Survey Errors and Survey Costs,* Wiley, Hoboken, NJ.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009), *Survey Methodology*, 2nd edition, Wiley, Hoboken, NJ.

Lessler, J. T. and Kalsbeek, W. D. (1992), *Nonsampling Error in Surveys*, Wiley, New York.

Statistics Finland (2019), *Income Inequality (International Comparison) 2017 – Quality report: income distribution statistics* (http://www.stat.fi/til/tjt/2017/01/tjt_2017_01_2018-12-19_laa_001_en.html).

Till-Tentschert, U., Reichel, D. and Latcheva, R. (2018), 'Surveying hard to reach groups in cross-country research', *Working Papers,* No 20, Expert meeting on measuring poverty and inequality, Vienna (https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.15/2018/mtg1/F._FRA.pdf).

5 Representativeness of 2011 EU-SILC responses and response rates over time

Natalie Shlomo, Annemieke Luiten and Barry Schouten ([50])

5.1. Introduction

This chapter investigates the representativeness of the 2011 European Union Statistics on Income and Living Conditions (EU-SILC) responding samples across European countries, compared with population totals obtained from the 2011 census, for a number of variables that feature in both EU-SILC and the census. These variables are age, sex, economic activity, education level and citizenship. The chapter also examines EU-SILC response rates over time, from 2006 to 2017.

Representativeness of the responding EU-SILC panel is crucial for comparability and accuracy of EU-SILC survey statistics. In the first part of this chapter, we study the general population aged 16 years and older from which EU-SILC draws samples and how the EU-SILC response compares with this population. We define representativeness as a feature that depends on a specified set of variables. Response is representative when response propensities are constant for the selected variables. In other words, when we cannot find that the variables provide any explanation of response, then it is representative relative to this variable set.

To assess whether a survey is representative of a target population, we need sample or population distributions of the selected variables and the absence of context measurement effects in observing these variables. The latter means that measurement of the variables is not related to being a respondent or non-respondent. In practice, the set of variables for which these conditions hold may be limited, leaving more room for non-response to impact representativeness on key survey variables without signalling this on selected variables. In an evaluation of 14 survey data sets, Schouten et al. (2016) conclude that surveys that are less representative on selected variables also tend to be less representative on non-selected variables, even after non-response adjustment on the selected variables. This provides empirical evidence that, on average, weaker representation is a signal of a bigger problem.

The assessment of 2011 representativeness is difficult, as for most countries only a limited set of variables against which to judge representativeness is available on the sample frame. To circumvent this problem, Shlomo et al. (2009) and Bianchi et al. (2019) introduced population-based R-indicators, as described in Section 5.2. The evaluation is performed for the 2011 survey, to be as close as possible in time to the 2011 European censuses, from which our benchmark population distributions are drawn. The census distributions are available on the Eurostat Census Hub ([51]).

We perform the evaluation of representativeness for the 2011 EU-SILC data sets for the following 26 countries: Austria, Belgium, Bulgaria, Croatia, Czechia, Denmark, Finland, France, Germany,

([50]) Natalie Shlomo (natalie.shlomo@manchester.ac.uk) is a professor of social statistics at the University of Manchester; Annemieke Luiten (a.luiten@cbs.nl) is employed by Statistics Netherlands; and Barry Schouten (sg.schouten@cbs.nl) is employed by Statistics Netherlands and is a professor by special appointment at Utrecht University. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors.

([51]) https://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census

Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Malta, the Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland and the United Kingdom. The following countries were not included in the analysis: Cyprus and Luxembourg did not have census data available on the Eurostat Census Hub; and Romania and Slovakia did not converge to the correct variance of the response propensities. This can occur when response rates are very high. Note that a possible reason for apparently very high response rates is that some of the EU-SILC data sets may be subset components of a larger survey and the response rate is recorded only for the EU-SILC survey, assuming that the larger survey has a 100 % response rate.

In the second part of this chapter, we assess the (individual) response rates of EU-SILC over time, from 2006 to 2017. The response rates are derived from the overall survey quality reports, which are available on the website of the Communication and Information Resource Centre for Administrations, Businesses and Citizens ([52]). Data for 2007 and 2012 were not available at the time of access (May 2020).

In Section 5.2, we elaborate on our methodology for assessing representativeness through R-indicators and coefficients of variation (CVs) for the 2011 EU-SILC surveys based on auxiliary information available at the population level from the 2011 census round. In Section 5.3, we describe the data sets and variable selection. In Section 5.4 and in the appendix, we provide results. Section 5.5 contains the analyses of response rates over time. We end with conclusions in Section 5.6.

5.2. Assessment of representativeness

Schouten, Cobben and Bethlehem (2009) introduced the concept of representative response. A response to a survey is said to be representative with respect to X when response propensities are constant for X, namely $\rho_i \equiv \rho_X(x_i) = \overline{\rho},\ \forall x_i$; here, $\overline{\rho}$ denotes the average response propensity in the population. The overall measure of representative response is the R-indicator based on the set of population response propensities $\{\rho_i : i \in U\}$ and defined as $R_\rho = 1 - 2S_\rho$, where S_ρ denotes the standard deviation of the individual response propensities, and $S_\rho^2 = \frac{1}{N-1}\sum_U (\rho_i - \bar{\rho}_U)^2$, where $\bar{\rho}_U = \sum_U \rho_i / N$. The R-indicator takes values in the interval $\left[1 - \sqrt{\frac{N}{N-1}}, 1\right]$, with the upper value 1 indicating the most representative response (the ρ_i's display no variation) and the lower value $1 - \sqrt{\frac{N}{N-1}}$ (which is close to 0 for large surveys) indicating the least representative response (the ρ_i's display maximum variation). An important related measure of representativeness is the CV of the response propensities $CV_\rho = \frac{S_\rho}{\bar{\rho}_U}$. This is a relevant measure when considering population means or totals as parameters of interest. In those cases, it may be used instead of the R-indicator, as it standardises the measure of representativeness to the response rate. The CV bounds the absolute non-response bias of unadjusted response means for a variable Y divided by its standard deviation.

The original definition for sample-based R-indicators is based on the assumption that auxiliary variables are available for both the responding and the non-responding sample units. Denote the sample survey by s selected from a finite population U. The sizes of s and U are denoted by n and N, respectively. The sample is assumed to be drawn using a probability sampling design $p(.)$, where the sample s is selected with probability $p(s)$. The first-order inclusion probability of unit i is denoted π_i, and $d_i = \pi_i^{-1}$ is the design weight. The survey is subject to unit non-response. For available auxiliary information, it is possible to estimate response propensities for all sampled units by means of regression models: $g(\rho_i) = x_i^T \beta$, where $g(.)$ is a link function, r_i is the dependent variable, where $r_i = 1$ if unit i responds and $r_i = 0$ otherwise, and $x_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$ is the vector of K explanatory variables. Ideally, it is desirable to select the auxiliary variables x_i in such a way that the missing at random assumption (Lit-

([52]) https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp

tle and Rubin, 2002) holds as closely as possible. The response propensities are typically modelled by generalised linear models. Shlomo, Skinner and Schouten (2012) use a logistic link function. Let $\hat{\rho}_i$ be an estimator for ρ_i. The sample-based estimator for the R-indicator is $\hat{R}_{\hat{\rho}} = 1 - 2\hat{S}_{\hat{\rho}}$, where $\hat{S}^2_{\hat{\rho}}$ is the design-weighted sample variance of the estimated response propensities, and $\hat{S}^2_{\hat{\rho}} = \frac{1}{N-1}\sum_s d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$, where $\hat{\bar{\rho}}_U = \sum_s d_i \hat{\rho}_i / N$. The estimator for the CV is defined as $\frac{\hat{S}_{\hat{\rho}}}{\hat{\bar{\rho}}_U}$. R-indicators and partial R-indicators have been shown to be good-quality measures for assessing representativeness of the sample and target populations (Schouten, Shlomo and Skinner, 2011; Schouten and Shlomo, 2017).

In the EU-SILC data sets, there is no information on non-responding units. Therefore, we assess representativeness of the 2011 EU-SILC responding data sets using population-based auxiliary information whereby the population distributions are obtained from the 2011 European censuses (Eurostat Census Hub).

5.2.1. Population-based R-indicators

The set of responding units is denoted by r, so $r \subset s \subset U$. As above, let r_i be the response indicator variable so that $r_i = 1$ if unit i responds and $r_i = 0$ otherwise. Hence, $r = \{i \in s; r_i = 1\}$. Let us suppose that the typical target of inference is a population mean $\bar{Y} = N^{-1}\sum_U y_i$ of a survey variable.

We assume that the data available for estimation purposes consist first of the values $\{y_i; i \in r\}$ of the survey variable, observed only for respondents. Then, we suppose that information is available on the values $x_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$ of a vector of auxiliary variables X. We usually suppose each $x_{k,i}$ is a binary indicator variable, where x_i represents one or more categorical variables. We assume that values of x_i are observed for all respondents so that $\{y_i, x_i; i \in r\}$ is observed.

We assume that x_i is known at the aggregate level: the population total $\sum_U x_i$ and population cross-products $\sum_U x_i x_i^T$. We refer to this type of information as population-based auxiliary information. Here, the variables X are taken from the 2011 censuses in the different countries.

Response propensities are defined as the conditional expectation of the response indicator variable r_i given the values of specified variables and survey conditions: $\rho_i \equiv \rho_X(x_i) = E_m(r_i \mid x_i)$ and $E_m(.)$ is the expectation with respect to the model underlying the response mechanism. In the population-based setting, we model the response propensities under an identity link function whereby the true response propensities satisfy $\rho_i = x_i^T \beta$, $i \in U$. The identity link function is a good approximation to the more widely used logistic link function when response rates are mid-range, between 30 % and 70 %, which is the typical response rate obtained in national surveys such as EU-SILC.

For the linear probability model, the estimate of ρ_i is given by $\hat{\rho}_i^{OLS} = x_i^T \left(\sum_s d_i x_i x_i^T \right)^{-1} \sum_s d_i x_i r_i$, $i \in s$. In the case of population-based auxiliary information in which we know both population totals and cross-products, we note that $\sum_s d_i x_i$ and $\sum_s d_i x_i x_i^T$ are unbiased estimates for $\sum_U x_i$ and $\sum_U x_i x_i^T$, respectively, and that in large samples we may expect that $\sum_s d_i x_i \approx \sum_U x_i$ and $\sum_s d_i x_i x_i^T \approx \sum_U x_i x_i^T$. It follows that, in the population-based setting, we may approximate $\hat{\rho}_i^{OLS}$ by $\hat{\rho}_i^P = x_i^T \left(\sum_U x_i x_i^T \right)^{-1} \sum_r d_i x_i$, $i \in r$. Note that $\hat{\rho}_i^P$ is computed only on the set of responding units.

In the population-based setting, an estimator for the R-indicator is then $\hat{R}_{\hat{\rho}^P} = 1 - 2\hat{S}_{\hat{\rho}^P}$, where $\hat{S}^2_{\hat{\rho}^P} = \frac{N}{N-1}\left\{ \frac{1}{N}\sum_r d_i \hat{\rho}_i^P - [\frac{1}{N}\sum_r d_i]^2 \right\}$ and $\hat{\rho}_i^P$ is esti-

mated as above. The estimator of the R-indicator makes the estimator $\hat{S}^2_{\hat{\rho}}$ linear in $\hat{\rho}_i^P$, which provides an advantage for bias adjustment computations. Furthermore, we use propensity weighting by $\hat{\rho}_i^{P-1}$ to adjust for non-response bias. The estimation of the CV in the population-based setting is straightforward: $CV_{\hat{\rho}^P} = \frac{\hat{S}_{\hat{\rho}^P}}{\hat{\bar{\rho}}_U}$ where $\hat{\bar{\rho}}_U = \sum_r d_i / N$.

Shlomo, Skinner and Schouten (2012) derive analytic approximations for the bias and standard errors of the sample-based estimate of the R-indicator. The bias in the sample-based R-indicator arises mostly from 'plugging in' estimated response propensities in the sample variances. This source of bias is referred to as small sample bias. A much smaller and usually negligible contribution to the bias originates from using sample means rather than population means. For the estimated population-based R-indicators, the statistical properties are different from their sample-based counterparts. As these estimators use less information, the standard errors are larger. The bias of the population-based R-indicators may also be larger, since, in addition to the bias that was evident for small sample sizes in the sample-based estimators, the population-based estimators will likely have bias arising from the estimation of the sample means and covariances and from the restriction to (propensity-weighted) response means. To reduce the bias of the population-based estimators, we adjust $\hat{S}^2_{\hat{\rho}^P}$ for bias. This leads to the adjusted version of the estimator for the R-indicator $\hat{R}^{Adj}_{\hat{\rho}^P} = 1 - 2\left[\hat{S}^2_{\hat{\rho}^P} - \hat{B}_{\hat{\rho}^P}\left(\hat{S}^2_{\hat{\rho}^P}\right)\right]^{1/2}$ and similarly for the CV. The expressions for the bias $\hat{B}_{\hat{\rho}^P}\left(\hat{S}^2_{\hat{\rho}^P}\right)$ under a complex survey design appear in Bianchi et al. (2019). To estimate the variance $V\left(\hat{R}^{Adj}_{\hat{\rho}^P}\right)$, we use the bootstrap method (Efron and Tibshirani, 1993; Wolter, 2007).

5.2.2. Unconditional partial R-indicators and coefficients of variation

The unconditional partial R-indicator measures the amount of variation of the response probabilities between the categories of a variable. The larger the between-category variation is, the stronger the relationship is and the stronger the impact of the variable on response. As earlier, let x_k be one of the components of the vector $\boldsymbol{x}$. Suppose x_k is categorical and has H categories. Let n_{hr} denote the weighted respondent sample size in category h, for $h = 1, 2 \ldots, H$. That means $n_{hr} = \sum_{i=1}^{r} d_i \Delta_{h,i}$, where $\Delta_{h,i}$ is the 0–1 indicator for responding unit i being a member of category h: $\sum_{h=1}^{H} n_{hr} = \hat{N}r$, where $\hat{N}r$ is the estimated total responding population. Define $\hat{\bar{\rho}}_h$ as the average of the response probabilities in category h of x_k for the responding units and $\hat{\bar{\rho}}$ as the overall average response probability based on the estimated population-based response probabilities $\hat{\rho}_i^P$ calculated in Section 5.2.1. The estimate for the unconditional partial R-indicator for variable x_k is $R_U(x_k) = \sqrt{\frac{1}{N}\sum_{h=1}^{H} n_{hr}\left(\hat{\bar{\rho}}_h - \hat{\bar{\rho}}\right)^2}$. The upper bound of the unconditional partial R-indicator is 0.5. The larger the value of the partial R-indicator, the stronger the association of the variable with non-response. By computing and comparing the unconditional partial indicators for a set of variables, it can be established for which variables the relationships are strongest. The unconditional partial R-indicator at the category level h for variable x_k is $R_U\left(x_k^h\right) = \sqrt{\frac{n_{hr}}{N}}\left(\hat{\bar{\rho}}_h - \hat{\bar{\rho}}\right)$ and can assume positive and negative values. The variable-level unconditional CV is defined as $\frac{R_U(x_k)}{\hat{\bar{\rho}}}$, and similarly the category-level version of the CV is defined as $\frac{R_U\left(x_k^h\right)}{\hat{\bar{\rho}}}$. Note that, at the category level, a negative sign represents under-representation and a plus sign represents over-representation.

5.3. Data

As mentioned, we used the 2011 cross-sectional EU-SILC data sets. We selected variables that exist in both the census and EU-SILC, and that were

also available for complete cross-tabulation for all selected variables in the Eurostat Census Hub. The five variables selected initially were age group (15–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75 and over), sex (male, female), economic activity status (employed, unemployed, inactive), educational level (secondary and post-secondary, university, all other) and citizenship (reporting country, foreign born). We could not use urban density, as the definitions in EU-SILC and the census proved to be too different to use (size of the locality in the census versus degree of urbanisation (DB100) in EU-SILC). Citizenship is available for the census and also featured in the variable list for EU-SILC except for two countries included in the analysis – Poland and Slovenia.

Another problem arose in the age group boundaries. Some of the census tables make use of 5-year age groups, which made any comparison with the EU-SILC population aged 16 and over problematic. For most of the tables, it proved possible to construct 2 × 2 tables with age in 1-year groups, thus circumventing this problem. For those tables for which this was not possible, the number of 15-year-olds in the various categories had to be estimated and removed from the counts to obtain tables with counts of those aged 16 years and over.

A third issue is the definition of activity status. In the census, this is based on the International Labour Organization (ILO) definitions, whereas EU-SILC uses a self-declared current 'main activity status' that captures the person's own perception of their main activity. It differs from the ILO concept to the extent that people's own perception of their main activity status differs from the strict definitions used by the ILO. For instance, many people who would regard themselves as full-time students or homemakers may be classified as ILO employed if they have a part-time job. Similarly, some people who consider themselves 'unemployed' may not meet the strict ILO criteria of taking active steps to find work and being immediately available.

Other issues (e.g. five International Standard Classification of Education levels in 2011 EU-SILC and six in the census, differences in the definition of 'missing' or 'other' for activity and education) are solved by collapsing categories and redefining the 'other' category. However, all of these small differences and adaptations introduce noise to the measurements. We presume that the level of noise is similar for all countries. As a final procedure, all census tables were calibrated to both univariate and bivariate counts on all five variables using a raking procedure to ensure consistent totals and subtotals.

Another problem was that the design weights were not available in the EU-SILC data sets. Therefore, design weights were approximated as follows.

(1) The total of the final survey weights in the EU-SILC data set were benchmarked to the population total according to the 2011 census count N^*. This was done by multiplying each final survey wei0ght w_i (the PB040 variable) in the EU-SILC data set by the ratio $N^* / \sum_r w_i$ to obtain w_i^*.

(2) Individual response rates are available in overall quality reports (see the EU-SILC hub ([53]). In these reports, the final individual response rate denoted R_p is defined as $(R_a \times R_{h|a} \times R_{p|ha})$, where R_a is the address contact rate, $R_{h|a}$ is the household response rate given an address contact, and $R_{p|ha}$ is the individual response rate given both contact and household response. Based on the final response rate R_p and the population total N^*, we calculate what would have been the total of the responding population as $M^* = R_p \times N^*$. We then pro-rated all of the modified survey weights w_i^* from (1) by multiplying by the ratio M^* / N^*.

Although this approach for approximating design weights does not mitigate the corrections that were made to compensate for non-response bias in the final survey weights w_i^*, it is envisaged that it will still capture unequal inclusion probabilities that may have been used in the sampling design of a country's EU-SILC. Therefore, we implemented this simple pro-rating in the responding population to obtain the pseudo-design weights: $d_i^* = w_i^* \left(\frac{M^*}{N^*} \right)$.

([53]) https://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports

5.4. Results

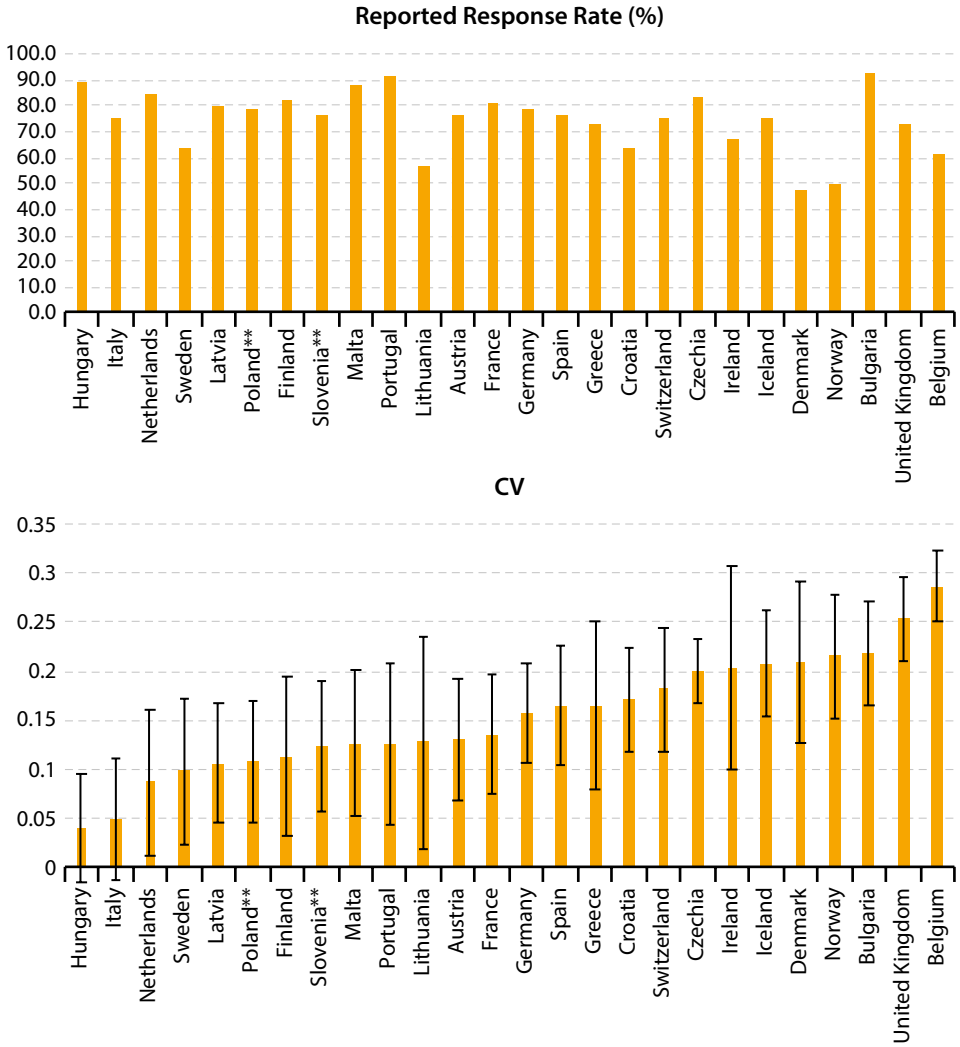
We report the CVs, as these are standardised to response rates and therefore provide a more accurate comparison measure across countries. Note that high CVs represent lower representativeness. We provide the 95 % confidence interval based on the bootstrap variance estimates.

Note that for two countries, Poland and Slovenia, the citizenship variable had to be excluded. Therefore, the results for these countries are not strictly comparable with those of other countries. We denote these countries with (**) in the figures.

5.4.1. Population-based coefficients of variation and response rates for 2011

Response rates for 2011 EU-SILC data sets and their CVs (with confidence intervals) are presented in Figure 5.1; the order of the countries is according to the magnitude of the CV. The standard errors

Figure 5.1: Response rate and population-based CV for 2011 EU-SILC



for the confidence intervals were calculated using a bootstrap with 300 repetitions. Aside from Hungary and Italy, the confidence intervals do not cross the horizontal line at zero, and hence there is a significant lack of representativeness for each country. It is not always the case that high response rates result in smaller CVs, as one would expect. This can be seen in the case of Bulgaria, which had a high response rate yet achieved a high CV. Lithuania had a low response rate yet achieved a CV similar to that for other countries with higher response rates. The Pearson correlation coefficient between the CV and response rate is – 0.4 ($p < 0.001$).

There are caveats to Figure 5.1. It is possible that the lack of representativeness results from measurement differences between the 2011 census data and the 2011 EU-SILC. In addition, for some countries where EU-SILC is a subsample from a larger survey, such as the Labour Force Survey, we do not know the true response rates, as they may be reported assuming 100 % response in the larger survey. Therefore, care should be taken when evaluating and interpreting the results comparatively across countries. We also note that the confidence intervals can be very large, especially for those countries with smaller sample sizes.

5.4.2. Unconditional partial variable-level coefficients of variation

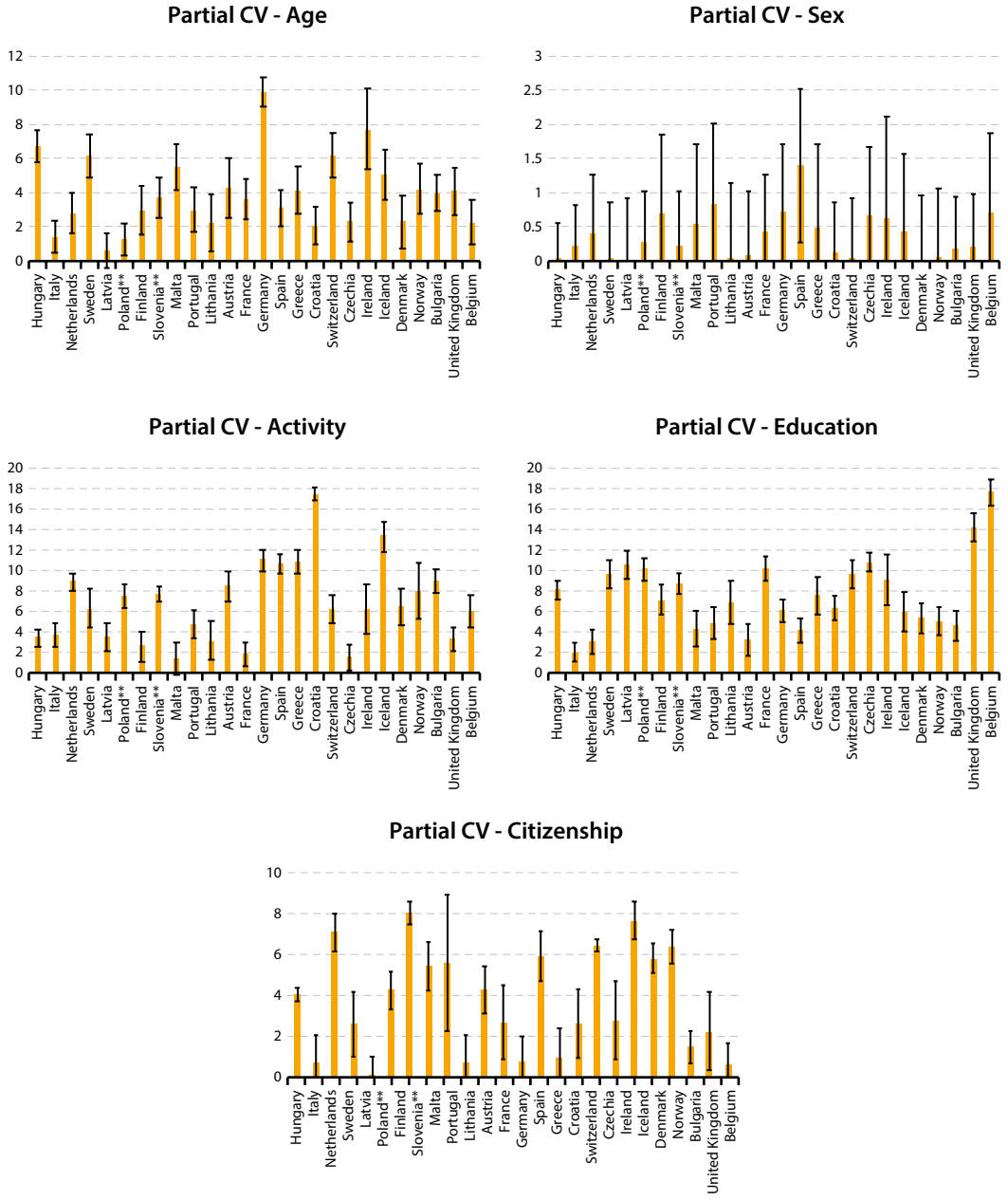
Figure 5.2 presents the population-based unconditional partial variable-level CVs (multiplied by 100). Note that each plot in Figure 5.2 has a different *y*-axis scale, since the aim is to compare across countries and not across variables, so care should be taken when making comparisons across the variables. It should be noted that economic activity and education level have a scale on the *y*-axis of up to 20 and have larger variations in the partial variable-level CVs than the other variables of age, citizenship and sex. The order of the countries in each of the plots is fixed according to the magnitude of the overall CV, as shown in Figure 5.1. It is interesting to note that the unconditional partial variable-level CVs do not exhibit the same monotonic pattern as Figure 5.1.

The confidence intervals can be very large. Confidence intervals that overlap with the horizontal line at zero mean that the CV is not significantly different from zero and the variable does not contribute to the lack of overall representativeness. This is the case for sex, for example, which, except for Spain, has zero effect on the lack of representativeness. This result, however, may be due to the way that the design weights were approximated based on the final survey weights, since the final survey weights would clearly have been adjusted for non-response according to sex. Age has more variation and largely significant CVs across countries, although we would expect here a similar effect of the approximated design weights. Economic activity and education level show the largest partial variable-level CVs and highly contribute to the lack of representativeness, although as mentioned this could also be an artefact of measurement differences between the 2011 census data and the 2011 EU-SILC. Although they have a smaller range, the partial variable-level CVs for citizenship are significant across all countries, and there is also evidence of measurement differences; for example, Poland and Slovenia did not have these variables in their 2011 EU-SILC data sets.

5.4.3. Unconditional partial category-level coefficients of variation

We present the unconditional partial category-level CVs (multiplied by 100) in the appendix. Figures 5.5–5.9 show, respectively, the plots of the categories of age, sex, economic activity, education level and citizenship. For each category within a variable shown in Figures 5.5–5.9, the y-axes for the partial category-level CVs have the same scale to ease comparisons across categories. In addition, countries are always ordered by the value of their overall CV (as in Figure 5.1). Across the categories of the variables, there are no consistent patterns of high (positive or negative) CVs compared with the overall CV, as reflected in the order of the countries. Note that a positive CV denotes over-representation and a negative CV denotes under-representation. A CV in which the confidence interval includes zero means that the category does not contribute to the lack of representativeness.

Figure 5.2: Unconditional partial variable-level CV for the variables age, sex, economic activity, education and citizenship (multiplied by 100)



For the age categories in Figure 5.5, we find significant partial category-level CVs for ages 15–24, with under-representation in Finland and over-representation in Malta, Austria and Iceland; for ages 25–34, there is under-representation in Ireland and the United Kingdom; for ages 35–44, there is under-representation in Hungary, but for ages 45–54 there is over-representation; for ages 55–64, there

is over-representation in France and Ireland; and for ages 65–74, there is over-representation in Sweden, Germany and Switzerland. For ages 75 and over, we see mainly under-representation in all countries, with large and significant CVs in Sweden, Slovenia, Malta, Germany, Greece and Switzerland.

For the sex categories in Figure 5.6, partial category-level CVs are small and mostly not significant, with the exception of Spain, which shows a significant under-representation of males and significant over-representation of females. Again, as seen for the partial variable-level CVs, these results for age and sex can be affected by non-response adjustments that may be still embedded in the approximated pseudo- design weights.

Economic activity and education level categories show higher magnitudes of unconditional partial category-level CVs and more variability, particularly for the unemployed and other education categories. For the economic activity categories in Figure 5.7, we find a smaller range of partial category-level CVs for the employed than for the unemployed and not active categories. The CVs for the unemployed largely show under-representation, with high and significant CVs in Germany and Iceland, but there is large over-representation in Bulgaria. There are several countries that do not have significant partial category-level CVs for the employed, showing that this category does not contribute to the lack of representativeness, as seen in Slovenia, Malta, Lithuania, France, Greece, the United Kingdom and Belgium. The CVs for the unemployed have a larger range, with under-representation in the Netherlands and Spain, and over-representation in Slovenia, Greece, Croatia and Norway. Countries where the unemployed category is not contributing to the lack of representativeness are Finland, Malta, Czechia, Ireland, Iceland and Denmark. The partial category-level CVs for the not active category have smaller variation across countries than those for the unemployed and are largely significant, with over-representation in Germany and Iceland. Countries where the not active category is not contributing to the lack of representativeness are Italy, Latvia, Malta, Lithuania and France.

For education level categories in Figure 5.8, we find high and significant partial category-level CVs for the other education category, with over-representation in Latvia and Poland and large under-representation in Hungary, Sweden, Slovenia, France, Switzerland, Czechia and Belgium. Those countries where the other education category is not affecting the lack of representativeness are Germany, Ireland, Bulgaria and the United Kingdom. For the secondary and post-secondary category, those countries where the category is not affecting the lack of representativeness are Italy, the Netherlands, Finland, Portugal, Lithuania, Austria, Norway and the United Kingdom, although France has large and significant over-representation. For the university category, those countries where the category is not affecting the lack of representativeness are Hungary, Czechia and the United Kingdom, although there is large and significant over-representation in Ireland and Belgium.

Finally, for the citizenship categories in Figure 5.9, we find small (and many non-significant) partial category-level CVs for the reporting countries category, but many significant CVs for the foreign-born category, with under-representation particularly in Hungary, the Netherlands, Finland, Malta, Portugal, France, Greece, Czechia, Iceland, Denmark and Norway, and over-representation in Lithuania.

5.5. Response rates over time

The response rates for 2011 EU-SILC cross-sectional data are shown in Figure 5.1. In this section, we provide an overview of the (individual) wave 1 response rates over time, from 2006 to 2017. The response rates are derived from the overall survey quality reports, which are available on Circabc's website ([54]). Data for 2007 and 2012 were not available at the time of access (May 2020).

As described above, response rates are an important aspect in the creation of representativeness. National statistical institutes therefore strive for high response rates. Nevertheless, de Leeuw, Hox and Luiten (2018) and Luiten, de Leeuw and Hox (2020) show that survey response rates decreased steadily and continually from 1980 to 2015, both

([54]) https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp

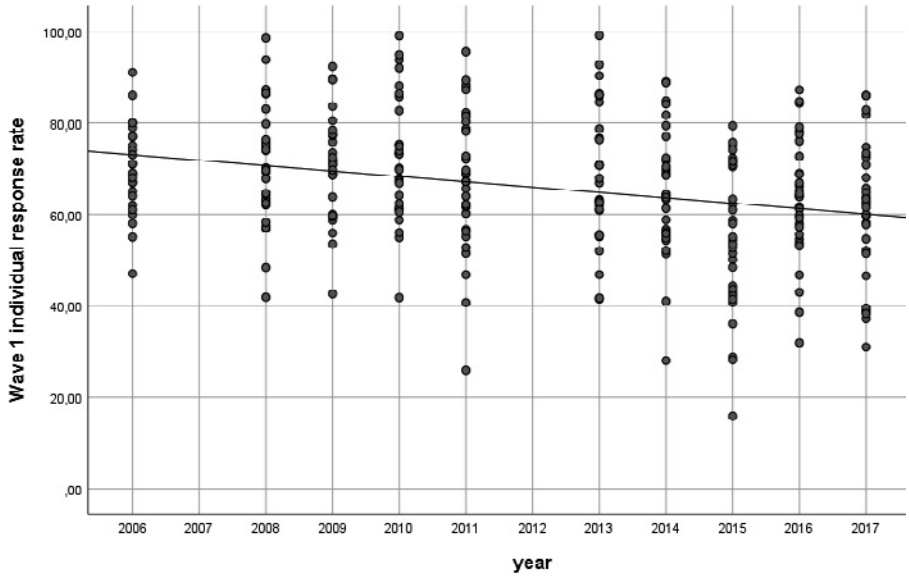
in the (often mandatory) Labour Force Survey and in other social surveys. Although there are large differences between countries in response rates, the decrease tends to be uniform across countries, with a decrease of about 0.73 percentage points per year. Figure 5.3 shows a scatterplot for the individual wave 1 response rates for EU-SILC, from 2006 to 2017. As above, individual response rates R_p are defined as ($R_a \times R_{h|a} \times R_{p|ha}$), where R_a is the address contact rate, $R_{h|a}$ is the household response rate given an address contact, and $R_{p|ha}$ is the individual response rate given both contact and household response. Response is cumulative at the three stages (address contact, household interview and personal interview). Trend analyses show that a significant downward response trend exists for the response rates ($\beta = -0.27$, $p < 0.001$).

Separate regression analyses per country specify that all countries but four (Bulgaria, Spain, Malta and Portugal) show a negative slope, although this is not always significant. Significant downward trends ($p < 0.05$) are observed for Belgium, Denmark, Finland, Luxembourg, the Netherlands, Slovakia, Slovenia, Sweden and the United Kingdom, whereas the downward trends in Ireland and Poland are close to significance ($p < 0.10$).

Table 5.1 gives an overview of the mean response rates. The data show a large dip of more than 10 percentage points in the response rates in 2015. A regression analysis to investigate whether the downward trend was caused by this phenomenon showed that the negative trend remained highly significant when the 2015 data were excluded.

Figure 5.3 and Table 5.1 show that there is a large disparity in response rates between countries: response rates may differ by as much as 70 percentage points. Differences between countries are influenced by design aspects such as the mandatory or voluntary nature of EU-SILC, by the modes used (web, paper, computer-assisted telephone interviewing or computer-assisted personal interviewing, or a mix), by the availability of register information, by the use of incentives and the kind of incentives, and by various other design aspects. Luiten, de Leeuw and Hox (2020) show, however, that, even after controlling for fieldwork design, substantial differences between countries may still exist.

Figure 5.3: Total individual wave 1 response rates from 2006 to 2017, plus trend line



NB: First-wave response rates for EU Member States and non-EU EU-SILC countries including the United Kingdom. The sloped line indicates the overall trend in these response rates.

Source: 2006–2017 comparative quality reports (https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp).

Table 5.1: Mean individual response rates, 2006–2017

Year	*n*	Mean	Standard deviation	Minimum	Maximum
2006	25	69.8	9.9	47.0	91.0
2008	28	70.2	13.3	41.9	98.5
2009	29	69.8	11.5	42.6	92.3
2010	30	72.2	13.5	41.8	99.0
2011	31	66.8	15.1	25.9	95.5
2013	29	68.5	15.1	41.5	99.1
2014	30	65.1	13.9	28.0	89.0
2015	31	54.6	15.9	15.9	79.4
2016	30	63.4	13.4	31.9	87.2
2017	30	60.1	15.2	31.0	86.0
Total	293	65.9	14.6	15.9	99.1

NB: These are the mean individual response rates (R_p) over countries for wave 1 panels. In addition, the standard deviation – a measure of the variance between countries – is shown, as well as the minimum and maximum rate within each available year. *n* indicates the number of countries that participated in EU-SILC in a given year.

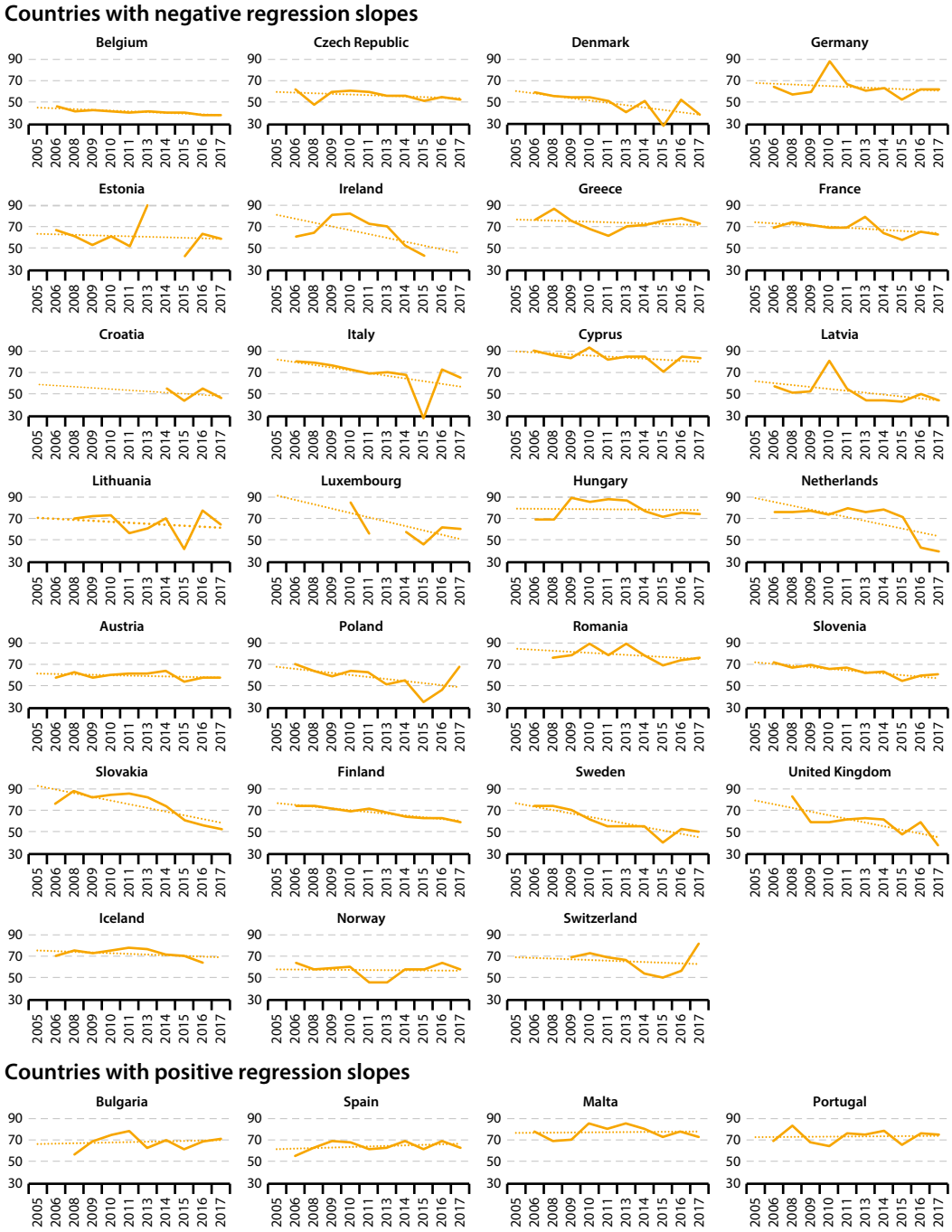
In contrast to the findings of de Leeuw, Hox and Luiten (2018) for the Labour Force Survey and Luiten, de Leeuw and Hox (2020) for various other social surveys, the decline within each country for EU-SILC is far less homogeneous. Figure 5.4 shows that there may be large fluctuations within one country in the response rates over time. The fluctuations may be sudden dips in response, like in a number of countries in 2015 and 2016, but also sudden highs. Marked examples are Germany in 2010, Estonia in 2013 and Latvia in 2010. In the time period studied here, there may be very large differences in response between years within one country: extreme examples are almost 52 percentage points in Italy between the highest response rate in the series and the lowest, 51 percentage points in Luxembourg, 48 percentage points in Estonia and 46 percentage points in Ireland. It would be worthwhile to study the causes of these sudden highs and lows and to understand the lessons to be learned for avoiding non-response. As seen in Section 5.4, in which we looked at the single response rate in 2011 and its correlation to the CV as the representativeness indicator, we see examples in Figure 5.4 of consistently high response rates over time with high CVs, and vice versa.

5.6. Conclusions

In this chapter, we have examined the representativeness of the 2011 EU-SILC data sets using overall indicators and unconditional partial variable-level and category-level indicators in the form of the CV of the response propensities, whereby the response propensities are estimated from population-based auxiliary data obtained from the 2011 European censuses. We also examined trends in response rates over time. We found that high CVs and lack of representativeness are not necessarily associated with response rates. The representativeness indicators show the contrast between those responding and those not responding to the survey. Therefore, for those countries with high CVs, the unconditional partial variable-level and category-level indicators should be examined to determine which variables and categories of variables are under-represented, particularly for the categories in economic activity, education level and citizenship. These findings can then be used to target sample units for non-response follow-up to obtain a more representative data collection.

It is important to keep in mind the limitations of this study. It was not possible to separate out wave 1 respondents from the 2011 EU-SILC data sets with meaningful population benchmarks from the 2011 censuses, and we assessed the representativeness of the full 2011 samples. Reported response rates tend to be conditional on the sample issued to the field: for some countries, these can be very different from the true unconditional response rates, for example if the EU-SILC sample is selected from the respondents to a previous survey. In addition, the

Figure 5.4: Individual response rates per country from 2006 to 2017, plus trend lines



NB: Response rates and trend lines per country. The upper part of the figure shows countries with a negative trend line (although sometimes it is very slight); the lower part shows countries with a positive trend line.

Source: 2006–2017 comparative quality reports (https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp).

representativeness indicators can be influenced by different measurements between the 2011 census and EU-SILC data, and each country should evaluate those country-specific measurement differences to assess to what extent they caused high CVs. Another limitation is the lack of design weights on the EU-SILC data sets. The design weights had to be approximated by pro-rating the final survey weights according to the published response rates, and this may affect the findings of this study. Firstly, response rates may not be accurately reported when EU-SILC is carried out on a subsample from a larger survey. Secondly, pro-rating the survey weights to derive approximate design weights means that the effects of non-response adjustments that are particularly likely to affect the age and sex distributions will remain. We assume that these limitations are similar across all countries, and hence the comparative nature of the study is not affected, though it should be noted that the set of auxiliary variables used for non-response adjustment differs between countries (see Chapter 12 of this book).

To continue with the legacy of this analysis, we recommend that design weights should be included in the EU-SILC data sets to allow for more meaningful assessments of representativeness on the socioeconomic and demographic variables, and that a standardised indicator of panel membership should be included so that analysts can separate out effects of initial non-response from attrition. We also recommend another round of assessment of representativeness for a more current EU-SILC in which each country is responsible for producing population statistics for the bivariate distributions of these variables, and is able to separate out the effects of first-wave non-response from attrition and to provide more detailed information about relevant aspects of response calculation and design. Software code is available from the authors.

References

Bianchi, A., Shlomo, N., Schouten, B., Da Silva, D. and Skinner, C. (2019), 'Estimation of response propensities and indicators of representative response using population-level information', *Survey Methodology*, Vol. 45, No 2, pp. 217–247.

de Leeuw, E., Hox, J. and Luiten, A. (2018), 'International nonresponse trends across countries and years: an analysis of 36 years of labour force survey data', *Survey Methods: Insights from the Field*, 20 December (https://surveyinsights.org/?p=10452).

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley, Hoboken, NJ.

Luiten, A., de Leeuw, E. and Hox, J. (2020), 'Survey nonresponse trends and fieldwork efforts in the 21st century: results of an international study across countries and surveys', *Journal of Official Statistics*, Vol. 36, No 3, pp. 469–487.

Schouten, B. and Shlomo, N. (2017), 'Selecting adaptive survey design strata with partial R-indicators', *International Statistical Review*, Vol. 85, No 1, pp. 143–163.

Schouten, B., Cobben, F. and Bethlehem, J. (2009), 'Indicators for the representativeness of survey response', *Survey Methodology*, Vol. 35, No 1, pp. 101–113.

Schouten, B., Shlomo, N. and Skinner, C. (2011), 'Indicators for monitoring and improving representativeness of response', *Journal of Official Statistics*, Vol. 27, No 2, pp. 231–253.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016), 'Does balancing survey response reduce nonresponse bias?', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 179, No 3, pp. 727–748.

Shlomo, N., Skinner, C., Schouten, B. de Heij, V., Bethlehem, J. and Ouwehand, P. (2009), *Indicators for Representative Response Based on Population Totals – Deliverable 2.2, work package 3 of the 7th framework programme (FP7) of the European Union RISQ project* (http://hummedia.manchester.ac.uk/institutes/cmist/risq/RISQ-Deliverable-2-2-V1.pdf).

Shlomo, N., Skinner, C. and Schouten, B. (2012), 'Estimation of an indicator of the representativeness of survey response', *Journal of Statistical Planning and Inference*, Vol. 142, No 1, pp. 201–211.

Wolter, K. M. (2007), *Introduction to Variance Estimation*, 2nd edition, Springer, New York.

Appendix: Unconditional partial category-level coefficients of variation

Figure 5.5: Unconditional partial category-level CV for categories of age (multiplied by 100)

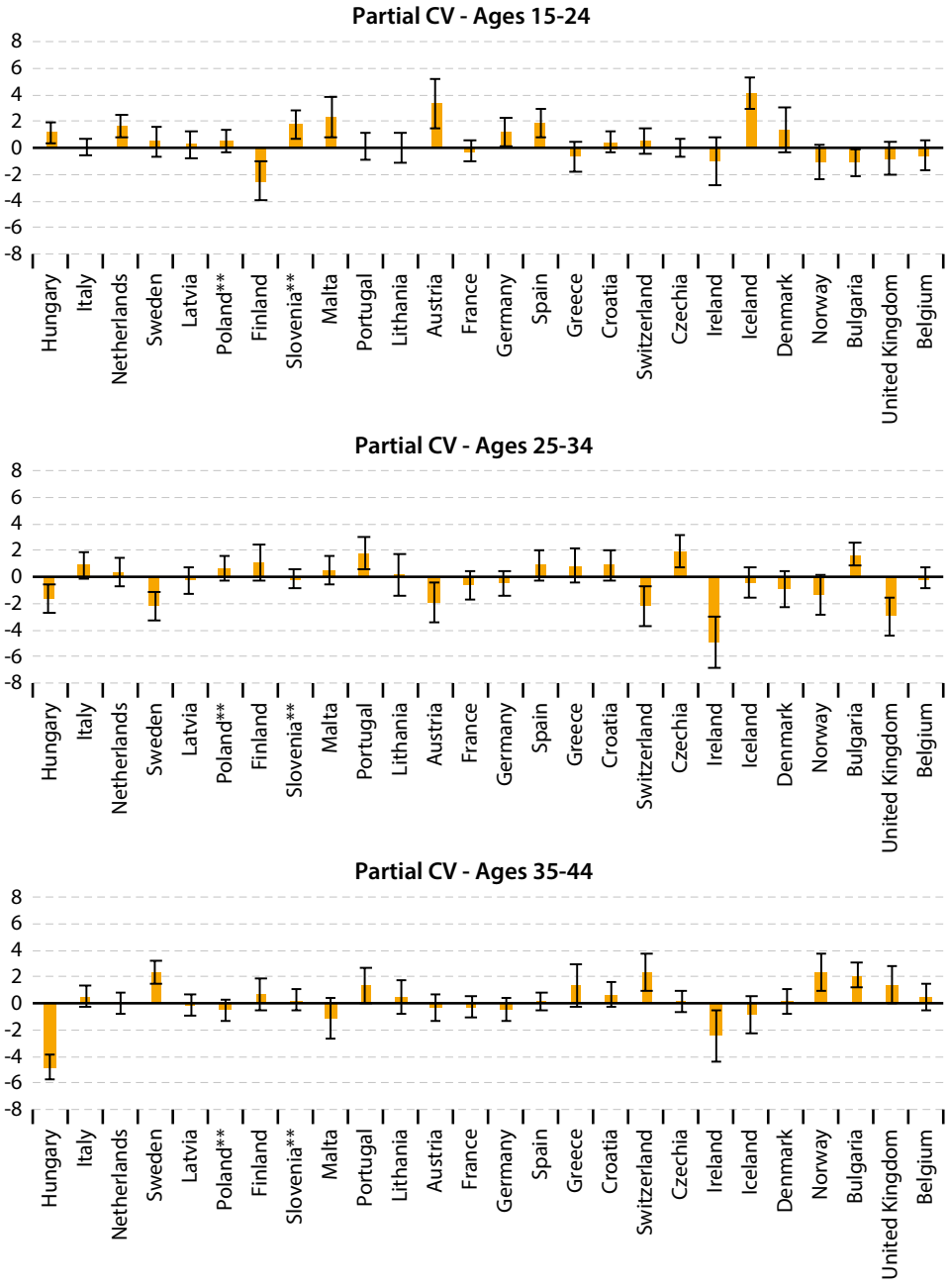


Figure 5.5 continued

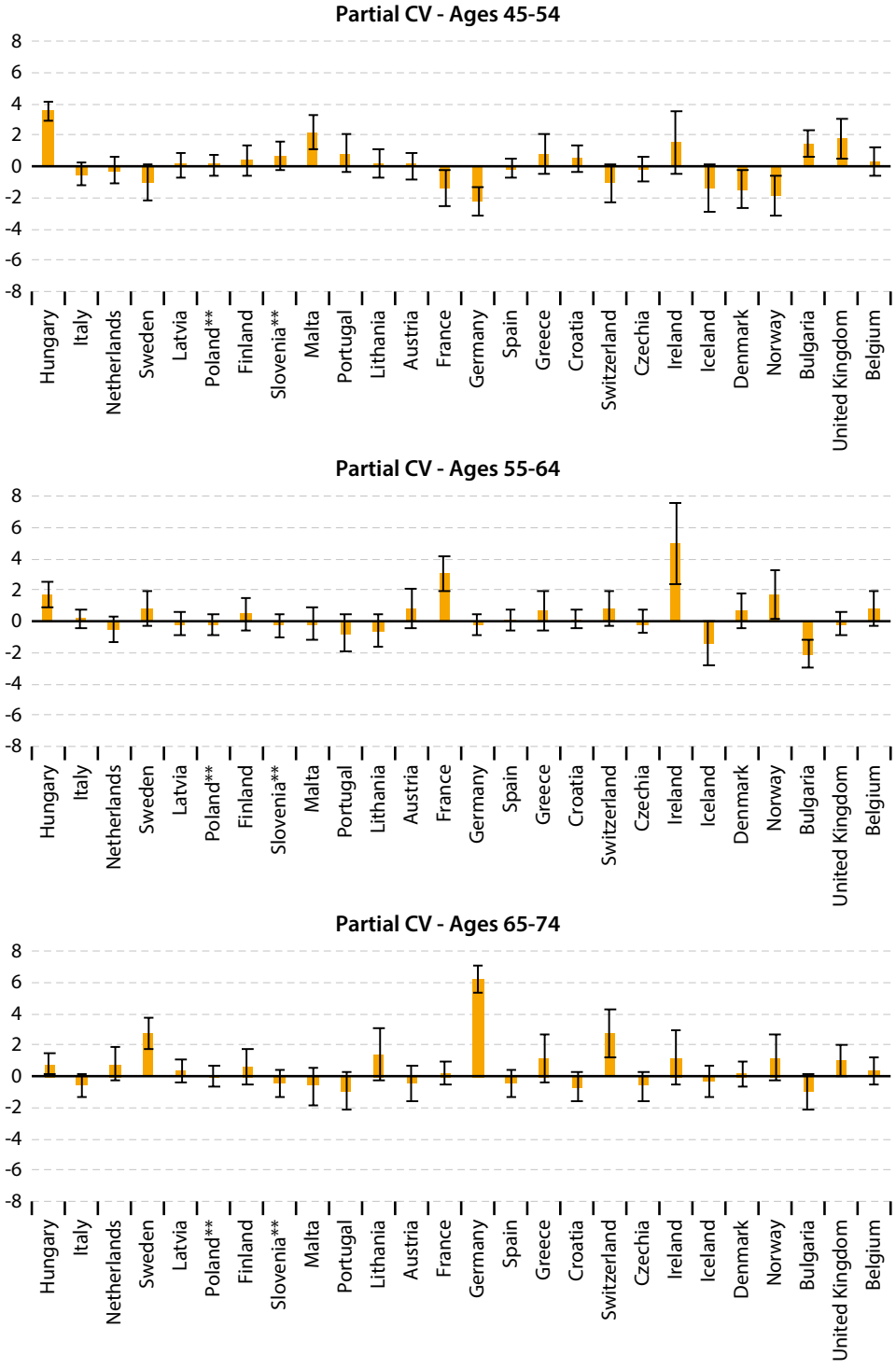


Figure 5.5 continued

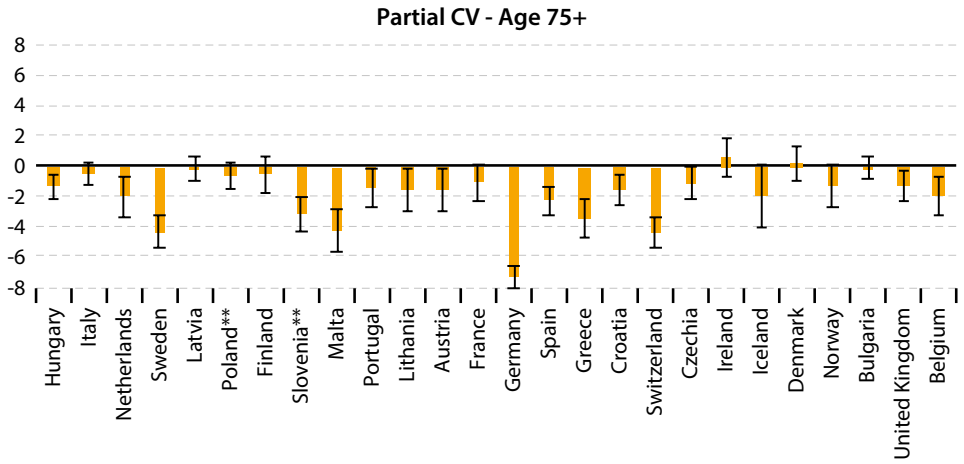


Figure 5.6: Unconditional partial category-level CV for categories of sex (multiplied by 100)

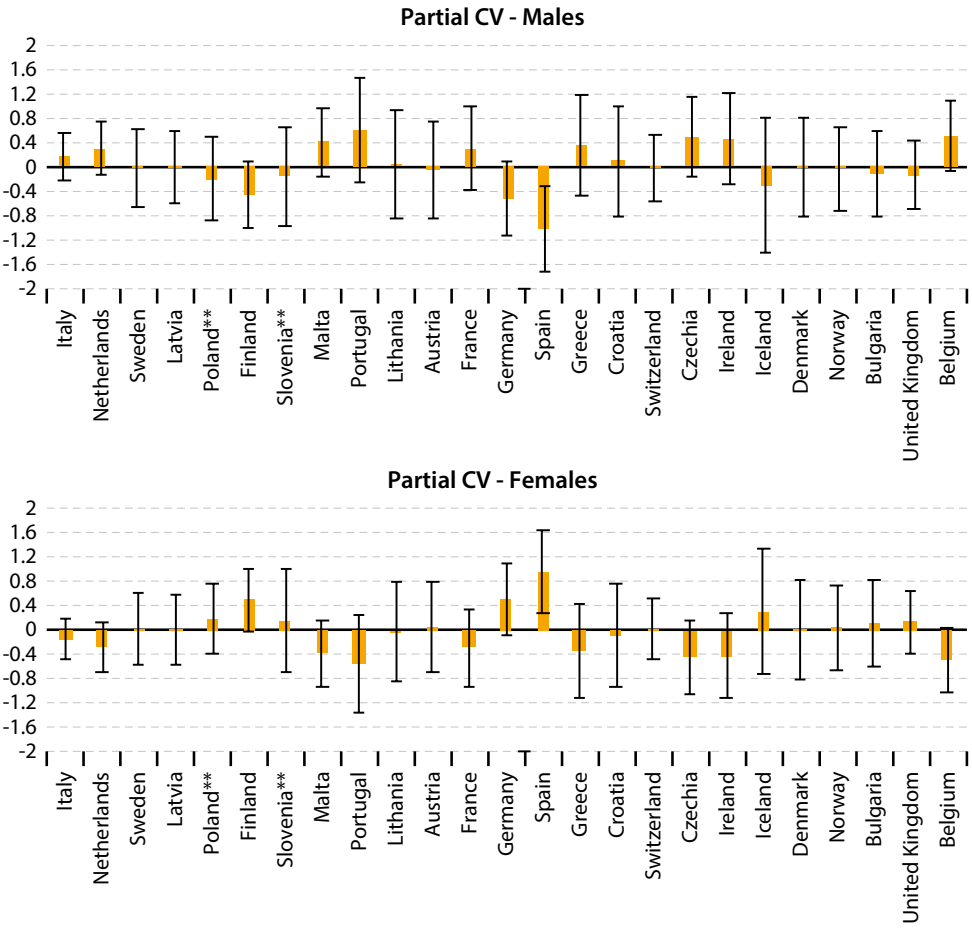


Figure 5.7: Unconditional partial category-level CV for categories of economic activity (multiplied by 100)

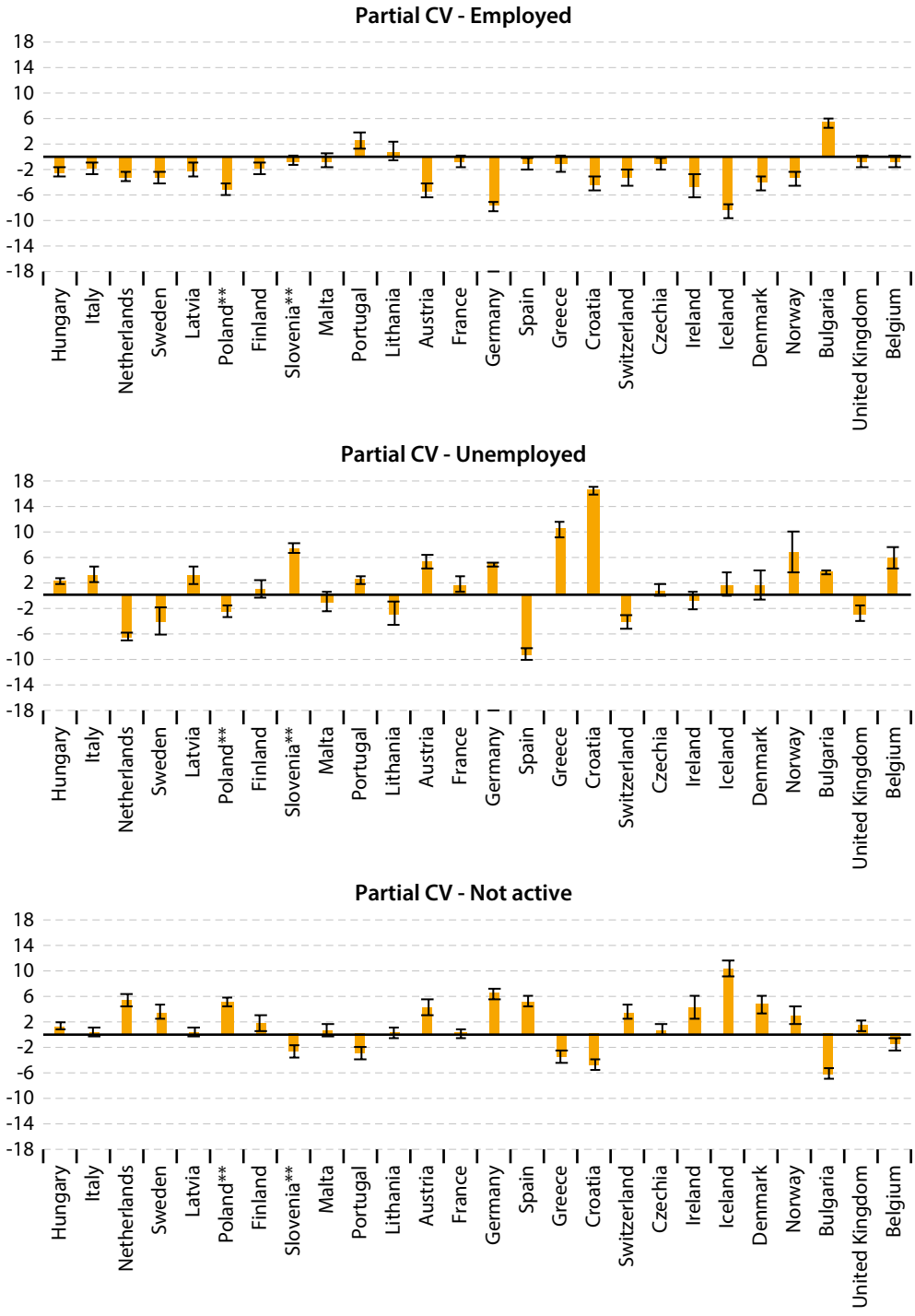


Figure 5.8: Unconditional partial category-level CV for categories of education (multiplied by 100)

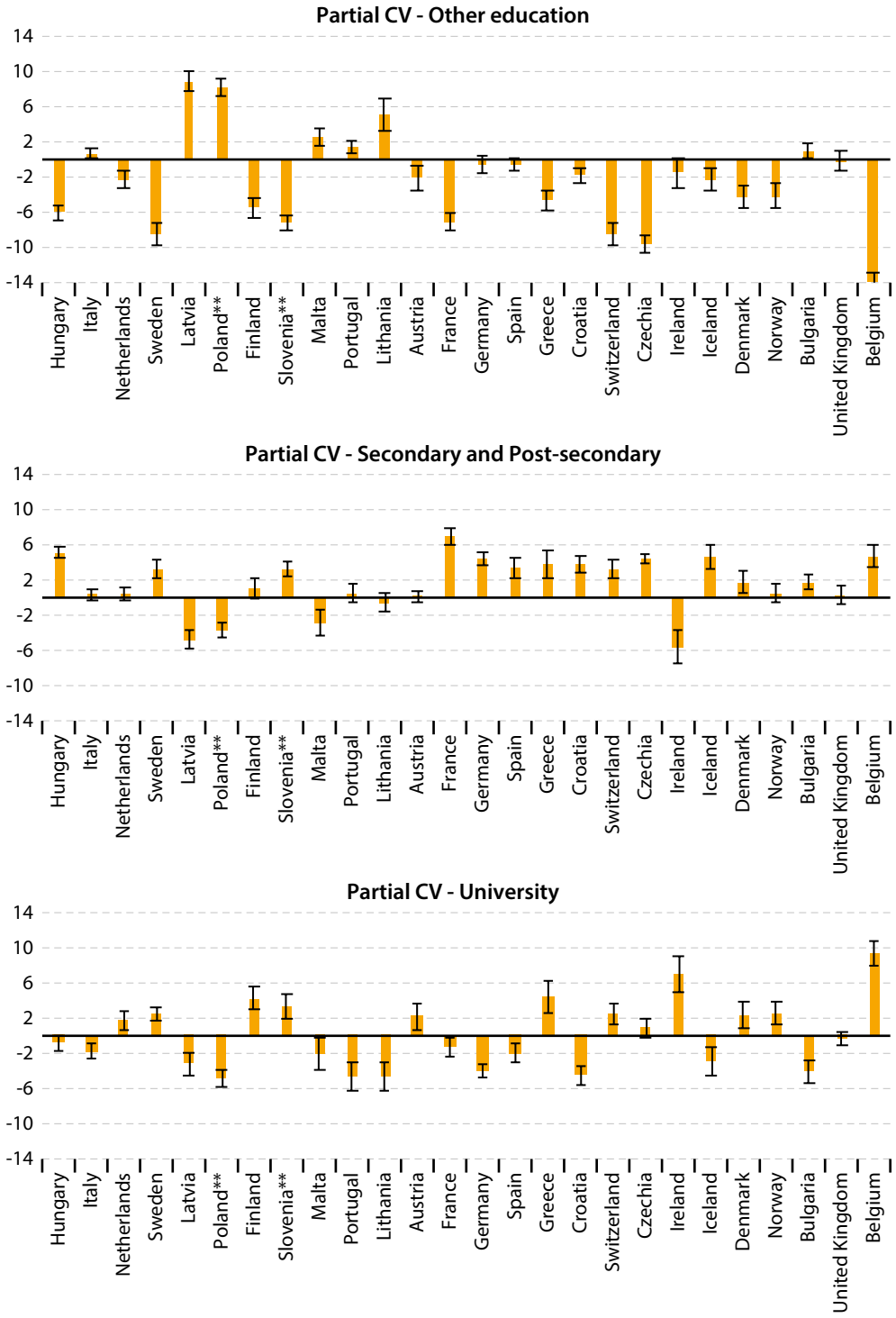
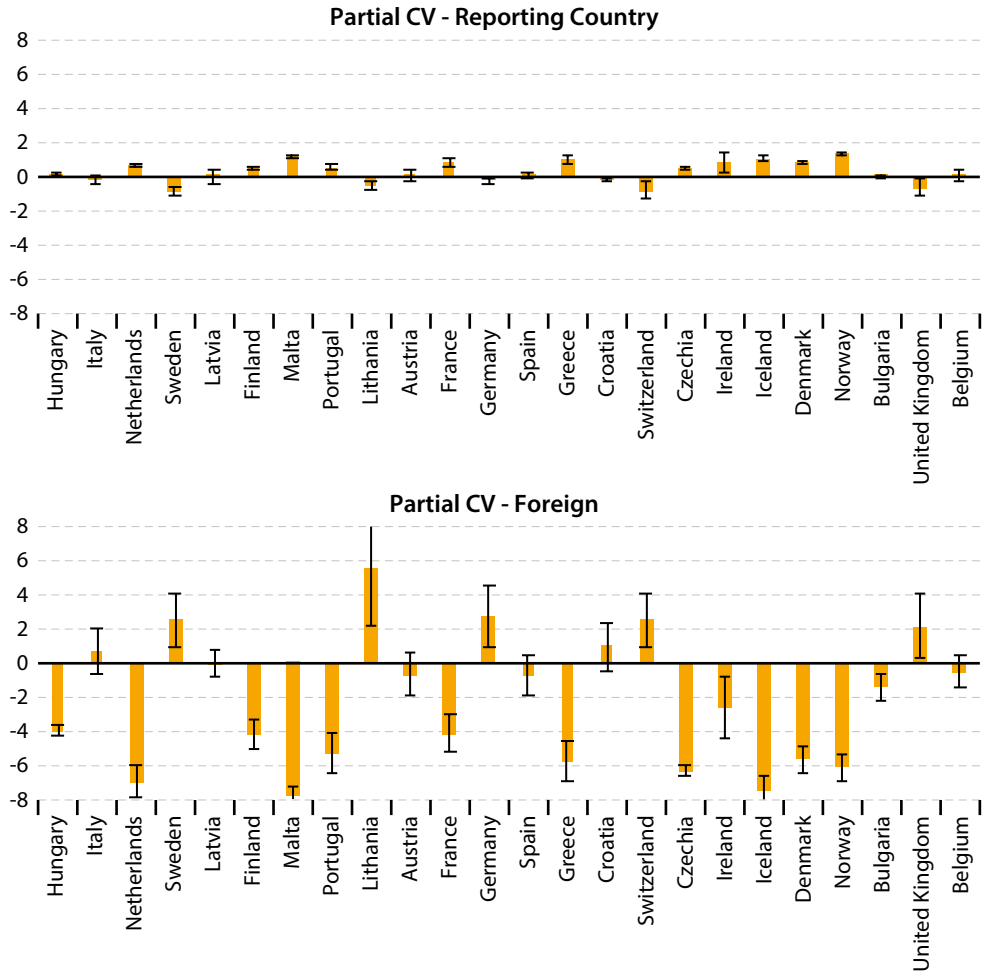


Figure 5.9: Unconditional partial category-level CV for categories of citizenship (multiplied by 100)



6 Impact on representativeness of EU-SILC panel attrition

Barry Schouten and Annemieke Luiten ([55])

6.1. Introduction

Between wave 1 and wave 4 of the European Union Statistics on Income and Living Conditions (EU-SILC) survey, panel members may (temporarily) drop out and not provide any survey data. This chapter investigates the impact of such attrition on the representativeness of the responding panel sample relative to the wave 1 panel sample. Chapter 5 reports on the representativeness of cross-sectional samples relative to the full target population. In the panel attrition assessments, we consider both standard census variables and EU-SILC-specific wave 1 variables. In order to appreciate and understand the findings of this chapter, we also recommend reading Chapter 5.

Representativeness of the responding EU-SILC panel is crucial for comparability and accuracy of key EU-SILC survey statistics. Two benchmarks for representativeness are set. One is the general population aged 16 years and older from which EU-SILC draws samples. Another is the wave 1 EU-SILC response from which subsets of respondents provide longitudinal information about changes in key EU-SILC survey statistics. The population benchmark is obviously the more important of the two; when the wave 1 response is strongly non-representative, then subsequent waves estimate changes in a specific subpopulation.

([55]) Barry Schouten is employed by Statistics Netherlands and is a professor by special appointment at Utrecht University (bstn@cbs.nl). Annemieke Luiten is employed by Statistics Netherlands (jrbs@cbs.nl). This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the views expressed, which are solely those of the authors.

Representativeness of the panel/survey response is a vaguely defined property. Here, as in Chapter 5, we define representativeness as a concept that depends on a specified set of variables. The response is representative when response propensities are constant for the selected variables. In other words, when we cannot find that the variables provide any explanation of response, then it is representative relative to this set of variables. Consequently, representativeness is meaningful only with specification of the variables. Obviously, response only needs to be representative for the variables of interest.

The preconditions for assessment are the availability of sample or population distributions of the selected variables and the absence of context measurement effects in observations of the variables. The latter means that measurement of the variables is not related to being a respondent or non-respondent. In practice, the set of variables for which these conditions hold may be limited, leaving more room for non-response to impact representativeness on key survey variables without signalling this on selected variables. In an evaluation of 14 survey data sets, Schouten et al. (2016) conclude that surveys that are less representative on selected variables also tend to be less representative on non-selected variables, even after statistical adjustment on the selected variables. Thus, on average, weaker representation is a signal of a bigger problem.

In the assessment of EU-SILC panel representativeness relative to wave 1, we can select the EU-SILC wave 1 main survey variables. Representativeness for such variables does not guarantee representativeness for change over time in the same main EU-SILC variables. However, since EU-SILC variables are

more closely related to change in EU-SILC variables than general variables available outside EU-SILC, non-response is less likely to impact representativeness without detection.

As we like to translate panel representativeness into the full population benchmark, we also evaluate representativeness of the EU-SILC panel relative to wave 1 for census variables alongside the evaluation for the main EU-SILC variables.

We perform the evaluation for the following 25 European Statistical System (ESS) countries: Austria, Belgium, Bulgaria, Croatia, Czechia, Denmark, France, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Norway, Poland, Romania, Slovenia, Slovakia, Spain, Sweden, Switzerland and the United Kingdom. The other seven ESS countries were not available in our data set. Panel attrition is evaluated for panels that started in 2012, namely those for which EU-SILC waves 2, 3 and 4 were conducted in 2013, 2014 and 2015, respectively.

In Section 6.2, we elaborate on our methodology for assessing representativeness. In Section 6.3, we provide results for the EU-SILC panel waves in the available ESS countries. We end with conclusions in Section 6.4.

6.2. Assessment of representativeness

We evaluate representativeness using sample-based R-indicators and coefficients of variation (CVs) (Schouten, Cobben and Bethlehem, 2009). SAS and R code and a manual can be found on the web page dedicated to representative indicators for survey quality (www.risq-project.eu).

Both indicators are distance measures of representative response on a specified set of variables. Representative response is defined as constant response propensities over all subpopulations formed by the selected variables. The measures are defined at population level – that is, assuming that response to a survey is independent of being sampled – but are estimated for each survey sample. As a consequence, the estimated R-indicators and CVs have a sample size-dependent precision. In the presentation of the measures, confidence intervals are provided in order to reflect the precision. We abbreviate the two measures to R and CV.

Both R and CV are a function of the standard deviation of response propensities $S(\rho_X)$, where ρ_X denotes the response propensities for the vector of variables X. R is defined as $R(X)=1-2S(\rho_X)$ and is estimated by inserting estimated response propensities and the weighted sample standard deviation: $\hat{R}(X)=1-2\hat{S}(\hat{\rho}_X)$.

The CV is also a function of the average response propensity, $\bar{\rho}_X$, and is defined as $CV(X)=\frac{S(\rho_X)}{\bar{\rho}_X}$, where the estimator employs the weighted response rate (RR): $\widehat{CV}(X)=\frac{\hat{S}(\hat{\rho}_X)}{RR}$.

What is the conceptual difference between the two measures? R measures the variation in response propensities. When non-response is like a toss of a coin, where the coin is not fair but is the same for all sample units, then variation is zero. This implies that non-response is equivalent to random subsampling and induces no bias. R is a measure that is not related to any specific population parameter, and it is assumed that more variation of response propensities is generally bad for any estimator of any population parameter. CV, however, is closely linked to the bias of weighted response means as estimators for population means. The expected non-response bias of a response mean for a variable Y equals the covariance between response probabilities and Y divided by the (weighted) response rate. It can be shown that, in an absolute sense, the expected non-response bias can be bounded by the CV of the true (unknown) response probabilities multiplied by the population standard deviation of Y. The population standard deviation of Y is a fixed parameter that is not dependent on non-response and is therefore ignored. The true response probabilities are unknown and are replaced by response propensities.

R and CV can be visualised through so-called response representativeness plots or RR-plots. In such plots, the R-indicators are plotted against the weighted response rates during data collection or across different waves. Downward diagonal lines starting at RR=0 and R=1 correspond to constant CVs.

Two further remarks need to be made. First, the weighting of the standard deviation and the response rate is carried out with respect to the sampling design (i.e. the inclusion or design weights). This is done because R and CV are population parameters and may be biased due to the sampling design, for example by oversampling or undersampling subpopulations with different response propensities. Second, the response propensities are estimated using variables that are available for the full sample, both respondents and non-respondents – hence the term sample-based R and CV. Estimators for R and CV based on population totals for the selected variables are called population-based R and CV, but they are not employed in this chapter.

Both measures can be developed for individual variables and categories of variables by decomposing the variance of response propensities into between and within variances (see Schouten, Shlomo and Skinner, 2011). The resulting variable-level and category-level measures are called partial R and partial CV. The between decomposition leads to so-called unconditional partial R/CV, as it measures the variation by the variable or category. The within decomposition gives the conditional partial R/CV, because it measures the unique contribution to variance adjusted for the other selected variables.

We will compute both overall R and overall CV, as well as partial R and partial CV. All R code will be made available on the web page dedicated to representative indicators for survey quality (www.risq-project.eu).

6.3. EU-SILC data

At the time of writing, we had access to the longitudinal EU-SILC data for 25 of the 32 ESS countries. In our data, Cyprus, Estonia, Finland, Germany, the Netherlands, Portugal and Turkey are missing. The data we used for the 25 countries were retrieved from the 2015 EU-SILC user database. From these data, we extracted the EU-SILC panel that started in 2012 ([56]).

([56]) The rotation group labels for these vary per country. It is label 1 for Croatia, Denmark, Greece, Ireland, Romania and Switzerland; label 2 for Bulgaria, Norway and Slovenia; label 3 for Hungary, Latvia, Lithuania, Luxembourg, Poland, Slovakia, Sweden and the United Kingdom; and label 4 for Austria, Belgium, Czechia, Iceland, Italy, Malta and Spain. It is groups 3–8 for France.

The longitudinal data for 2014 (which include the panels that started in 2011) were not available to us, so we are unable to link the panel attrition directly to the 2011 census. In linking our analysis to the 2011 cross-sectional evaluation, we assume that ESS populations did not change much between 2011 and 2012 in terms of EU-SILC main survey variables. As the benchmark in this chapter is the wave 1 panel population, our analysis itself is not affected by the starting year.

We consider two sets of auxiliary variables, which we will refer to as the census set and the EU-SILC set. The census set consists of four variables that are also used in the wave 1 assessment of representativeness (Chapter 5):

- age (PX020), which is recoded into seven classes – 16–24, 25–34, 35–44, 45–54, 55–64, 65–74 and 75 and older;
- sex (RB090) – female or male;
- highest educational level attained (PE040), which is recoded into (pre-)primary or lower secondary, upper secondary or tertiary, with missing values recoded to a separate category;
- activity status (RB210) – at work, in retirement, other inactive or unemployed.

The EU-SILC set consists of five variables:

- disposable household income (HY020), which is recoded to quintiles per country based on the wave 1 income distribution, that is, the quintiles vary per country;
- household can make ends meet (HS120) – very easily, easily, fairly easily, with some difficulty, with difficulty or with great difficulty;
- house has a leaking roof (HH040) – yes or no, with missing values recoded to a separate category;
- self-assessed health (PH010) – very good, good, fair, bad or very bad, with missing values set to a separate category;
- can afford unexpected expenses (HS060) – yes or no.

The EU-SILC set is a selection of variables from the relevant topics in EU-SILC. Representativeness on these variables is imperative within longitudinal comparisons.

Representativeness is assessed at the person level. From the longitudinal data, one analysis data set is created per country. The data set is created at the person level by linking and combining the D-files, H-files, P-files and R-files. All wave 1 respondents aged 16 years and older are selected. People with missing values on any of the census variables, except educational level, are deleted. People with missing values on EU-SILC variables disposable income, making ends meet and unexpected expenses are also deleted. In all cases, the numbers of deleted records were small. Missing values on self-assessed health and on a leaking roof are recoded into a separate category; the numbers of missing values were relatively large for these two variables. Three binary indicators per person are created, representing response in waves 2, 3 and 4. The personal base weight (PB020) is included in weight response rates, R-indicators and CVs.

6.4. Results

6.4.1. Representativeness for census and EU-SILC variables

We start by inspecting the CVs per wave and country for the census variables and for the EU-SILC variables. Table 6.1 shows the CVs for all 25 countries for waves 2, 3 and 4, relative to the wave 1 response. Estimated standard errors are between 0.005 and 0.018 depending on the country sample size. With a few exceptions, all CVs show a significant influence of attrition on representativeness. For context, Table 6.2 shows the corresponding response rates, conditional on participation in wave 1. It can be seen that the relationship between response rate and CV is far from simple.

Table 6.1: CVs for the 25 countries for wave 2 (2013), wave 3 (2014) and wave 4 (2015) relative to wave 1

Country	Census set			EU-SILC set		
	Wave 2	Wave 3	Wave 4	Wave 2	Wave 3	Wave 4
Austria	0.05	0.03	0.04	0.07	0.06	0.07
Belgium	0.05	0.05	0.06	0.07	0.08	0.08
Bulgaria	0.02	0.03	0.04	0.04	0.05	0.07
Croatia	0.07	0.13	0.14	0.08	0.15	0.17
Czechia	0.01	0.00	0.00	0.03	0.05	0.03
Denmark	0.02	0.06	0.05	0.02	0.03	0.03
France	0.03	0.03	0.07	0.04	0.04	0.04
Greece	0.03	0.03	0.05	0.02	0.04	0.05
Hungary	0.02	0.02	0.03	0.05	0.08	0.08
Iceland	0.05	0.05	0.05	0.05	0.05	0.05
Ireland	0.08	0.12	0.21	0.06	0.18	0.39
Italy	0.04	0.05	0.08	0.09	0.09	0.15
Latvia	0.00	0.03	0.03	0.04	0.04	0.03
Lithuania	0.03	0.07	0.10	0.06	0.08	0.08
Luxembourg	0.03	0.03	0.03	0.12	0.10	0.11
Malta	0.04	0.07	0.10	0.04	0.07	0.13
Norway	0.04	0.04	0.04	0.04	0.04	0.04
Poland	0.04	0.04	0.04	0.02	0.04	0.03
Romania	0.01	0.02	0.02	0.01	0.02	0.02
Slovakia	0.02	0.02	0.02	0.03	0.03	0.03
Slovenia	0.06	0.06	0.05	0.07	0.08	0.07
Spain	0.04	0.04	0.05	0.03	0.03	0.05
Sweden	0.02	0.14	0.13	0.05	0.10	0.10
Switzerland	0.23	0.23	0.23	0.10	0.10	0.10
United Kingdom	0.13	0.09	0.10	0.09	0.07	0.08

NB: Nearly all CVs are significantly different from zero at the 5% level. The only exceptions are the Census set for waves 2, 3 and 4 in Czechia and for wave 2 in Latvia and Romania and the EU-SILC set for wave 2 in Romania. Five classes for values of the CV are distinguished: ●, $0.00 \leq CV < 0.05$; ●, $0.05 \leq CV < 0.10$; ●, $0.10 \leq CV < 0.15$; ●, $0.15 \leq CV < 0.20$; ●, $CV \geq 0.20$. The higher the CV is, the greater the variation in response propensities and the less representative the sample, in terms of the variables *X*. If, for example, the response rate (Table 6.2) is 0.8, a CV of 0.05 implies that S, the standard deviation of the response propensities, is 0.04. This, in turn, implies that around 95 % of response propensities are between 0.72 and 0.88.

Table 6.2: Response rates for the 25 countries for wave 2 (2013), wave 3 (2014) and wave 4 (2015) relative to wave 1 (%)

Country	Wave 2	Wave 3	Wave 4
Austria	76.8	68.2	65.0
Belgium	76.5	69.6	59.1
Bulgaria	92.0	85.4	80.3
Croatia	76.2	62.8	58.4
Czechia	92.9	88.7	87.0
Denmark	100.0	99.6	99.5
France	81.4	75.3	67.7
Greece	93.1	83.3	75.6
Hungary	82.3	68.4	59.6
Iceland	92.0	91.4	90.0
Ireland	81.1	41.1	26.5
Italy	71.7	68.2	59.8
Latvia	88.2	81.8	76.4
Lithuania	93.2	89.1	83.3
Luxembourg	76.5	68.6	57.6
Malta	86.7	80.4	75.2
Norway	86.8	86.8	86.8
Poland	87.5	82.0	74.3
Romania	99.4	98.2	97.8
Slovakia	93.8	88.8	88.5
Slovenia	74.9	61.9	52.4
Spain	78.0	75.6	70.3
Sweden	88.5	81.4	81.6
Switzerland	81.4	72.8	72.7
United Kingdom	65.2	41.8	32.9

The larger the value of the CV is, the greater the risk of non-response bias. A few observations can be made here. First, the majority of countries have relatively small CV values, indicating a relatively low risk of bias from panel dropout. The CVs for Ireland and Switzerland are the largest and indicate a relatively high risk of panel attrition bias. Other countries with higher CVs are Croatia, Italy, Luxembourg, Malta, Sweden and the United Kingdom. Second, the CV tends to increase from wave 2 to wave 4, but in many countries this increase is modest. Exceptions are Ireland, Croatia and Italy, where the wave 2 CV is much lower than the wave 4 CV. Third, the patterns detected by the census variables and by the EU-SILC variables are often in line. When EU-SILC variables show a change in the CV, then so do census variables. This is an important finding, as sociodemographic variables are often available, whereas substantive variables are not.

As a second step, we investigate what caused large CVs and changes in CVs. For this purpose, we list the wave response rates relative to wave 1 in Table 6.1 and produce RR-plots in which we show the three waves for the census set and the EU-SILC set. From Table 6.1, we conclude that wave response rates vary greatly between countries. Two countries, Denmark and Romania, have close to 100 % response rates according to the data we used. Consequently, there is also almost no room for variation in response propensities.

Figure 6.1 displays the RR-plots for the 25 countries. Blue circles represent the census set and red circles represent the EU-SILC set. Since panel attrition lowers the response rates in all countries, with the exception of Sweden in waves 3 and 4, the values of waves 2–4 read from right to left. The 95 % confi-

dence intervals are included but are generally very narrow due to the large sample sizes.

The RR-plots contain additional information. A large CV can be the result of strong variation in response propensities, a small response rate or both. In Figure 6.1, we can see that, indeed, some larger CVs result from stronger variation and some result from smaller response rates. Examples of countries with relatively high variation (i.e. a smaller R-indicator) are Sweden and Switzerland. Note that Switzerland's response rates for wave 3 and wave 4 are almost the same and corresponding points overlap. Examples of countries with relatively low response rates are Ireland, except for wave 2, and Luxembourg. Croatia and Italy have a mix – attrition rates increase and R-indicators decrease from wave 2 to wave 4. There are also examples of countries, such as Slovenia, where attrition rates grow, but response propensity variation remains very small and counteracts any impact.

Figure 6.1: RR-plots for the 25 selected ESS countries for waves 2–4 (from right to left)

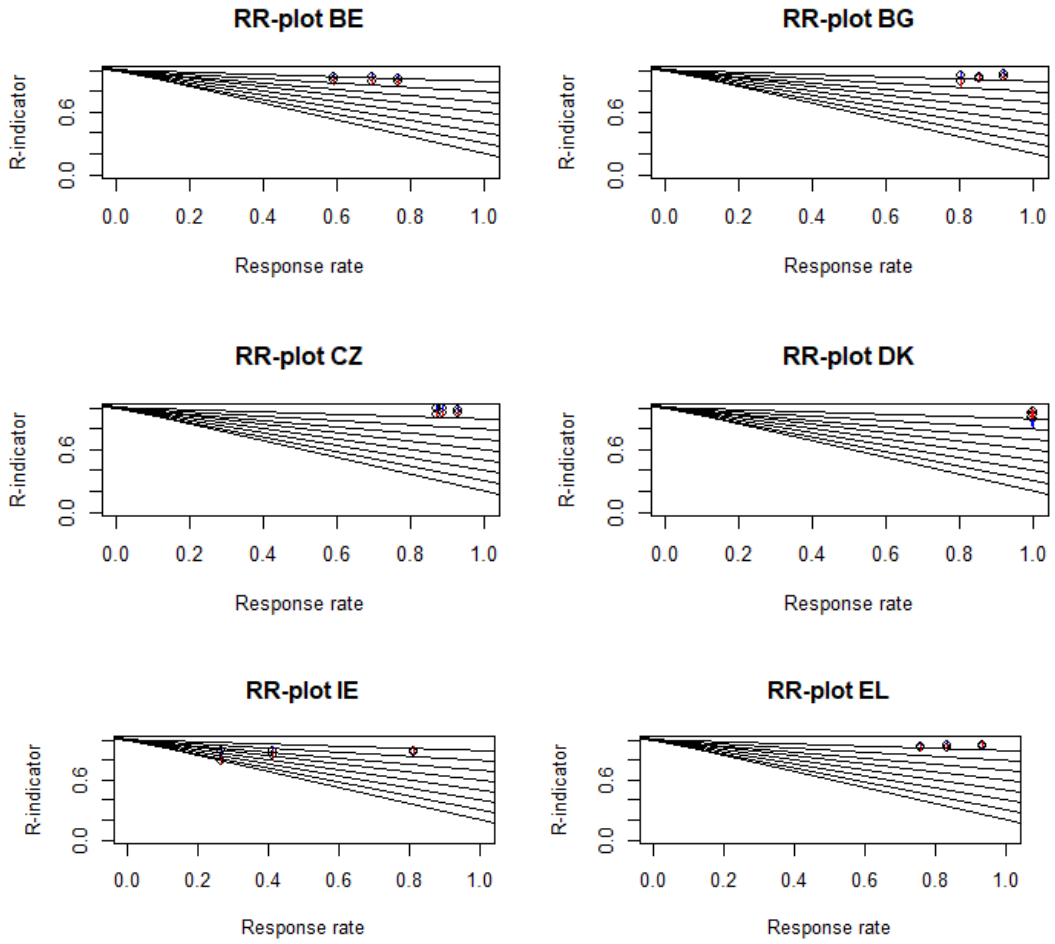


Figure 6.1 continued

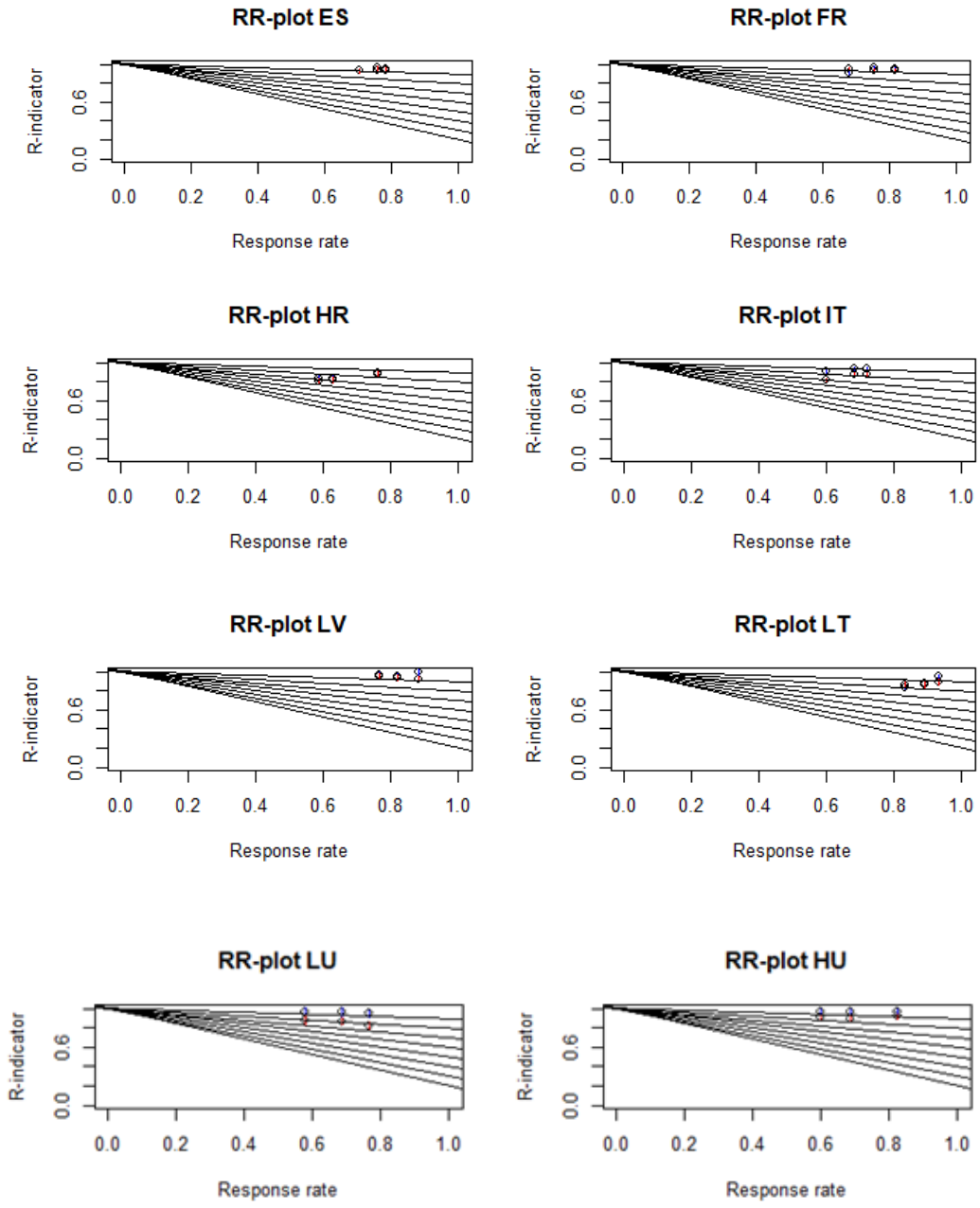


Figure 6.1 continued

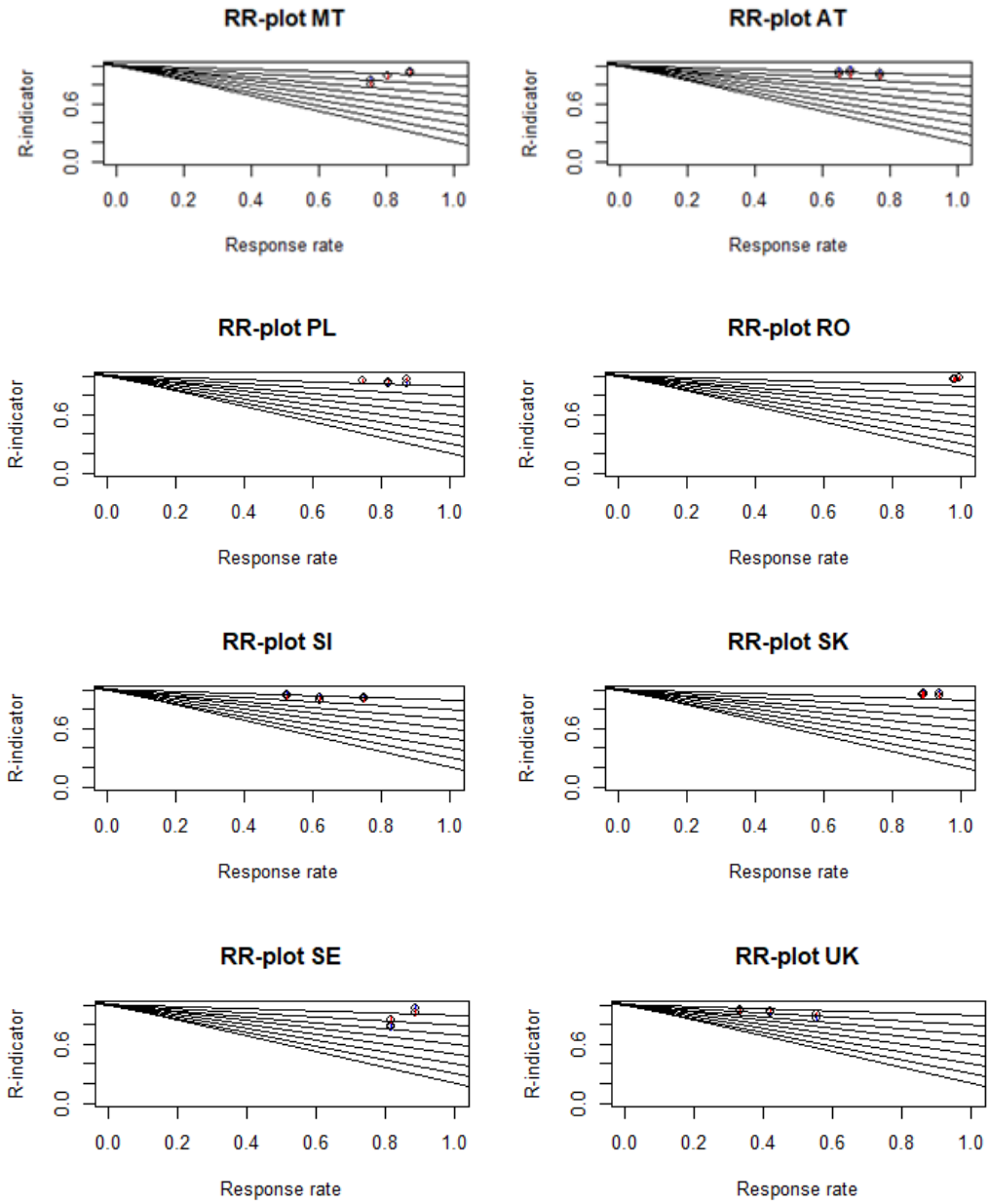
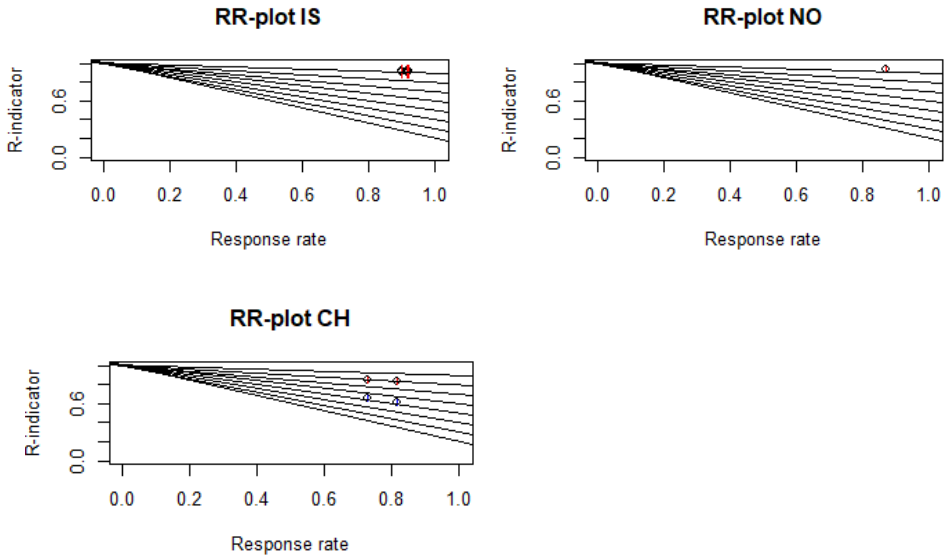


Figure 6.1 continued



NB: Blue circles correspond to the census set and red circles correspond to the EU-SILC set. The 95 % confidence intervals based on normal approximations are depicted.

6.4.2. Contribution of census and EU-SILC variables to response propensity variation

In the previous section, we concluded that some countries show more risk of non-response bias. For these countries, we go a step further and consult the variable-level and category-level partial CVs. We look at the countries that have at least two yellow or orange cells in Table 6.1: Croatia, Ireland, Italy, Luxembourg, Malta, Sweden, Switzerland and the United Kingdom.

We start by inspecting the variable-level partial CV. As mentioned in Section 6.2, there is an unconditional and a conditional partial CV. The unconditional partial CV shows contributions of variables to response propensity variation without any adjustment for collinearity with the other variables. The conditional partial CV estimates the unique contributions of variables. It must be noted, however, that it is important to avoid such collinearity beforehand in the selection of the variables. Both unconditional and conditional values fall in the interval [0, 0.5], where a value of 0.5 corresponds to maximal variance.

Figure 6.2 presents partial CVs for the subset of eight countries for wave 4, that is, at the end of the panel. Both unconditional and conditional values are shown. Conditional values have a similar size to unconditional values in most cases, indicating relatively weak collinearity between the variables in the census and EU-SILC sets. The variable that contributes most often to increased CVs is educational level. Census variable sex and EU-SILC variable leaking roof rarely contribute.

From the variable-level partial CV, we get the impression that none of the census or EU-SILC variables stand out very clearly as the strongest contributors to response propensity variation. There are a number of combinations of variables and countries where contributions are larger. These are educational level (EduClas) for Croatia, Ireland and Switzerland, the ability to make ends meet (MakeEndsMeet) for Ireland, income (IncClas) for Ireland, age (AgeClas) for Sweden and activity status (BasicActivity) for Ireland. For the other four countries no particular variable stands out, and it is the compound impact of all variables that leads to higher CVs.

Here, we take a closer look at variable–country combinations that show a larger impact. We do this by inspecting the category-level partial CV. Like the variable-level indicators, the category-level indicators can be calculated unconditionally and conditionally. Unlike the variable-level indicators, it is possible to add a plus or minus sign to the unconditional indicators in order to indicate over-representation or under-representation. Again, in the absolute sense, the values are smaller than or equal to 0.5, where 0.5 indicates maximal variance.

The seven panels in Figure 6.3 show variable–country combinations with relatively large contributions to response propensity variation in wave 4 of EU-SILC. All variable–country combinations are included where the variable-level CV is larger than 0.10. Negative unconditional partial CVs point to categories that are under-represented. In most cases, the categories that are strongly negative are those that show under-representation in many surveys. Those less educated drop out more often in the Irish and Swiss cases, the younger panel members show weaker representation in the Swedish case, and the lower income classes, the inactive and unemployed and households finding it more difficult to make ends meet are under-represented in the Irish case. The exception is those with a higher level of education in the Croatian case, who show stronger attrition during the EU-SILC panel.

The population subgroups to which the categories correspond are relevant for EU-SILC, and stronger dropout implies a risk of bias in longitudinal comparisons. The results from the category-level partial CV may form clues as to how to improve representation and reduce risk of attrition bias.

Figure 6.2: Variable-level partial CV for Croatia, Ireland, Italy, Luxembourg, Malta, Sweden, Switzerland and the United Kingdom for the census set and the EU-SILC set in wave 4

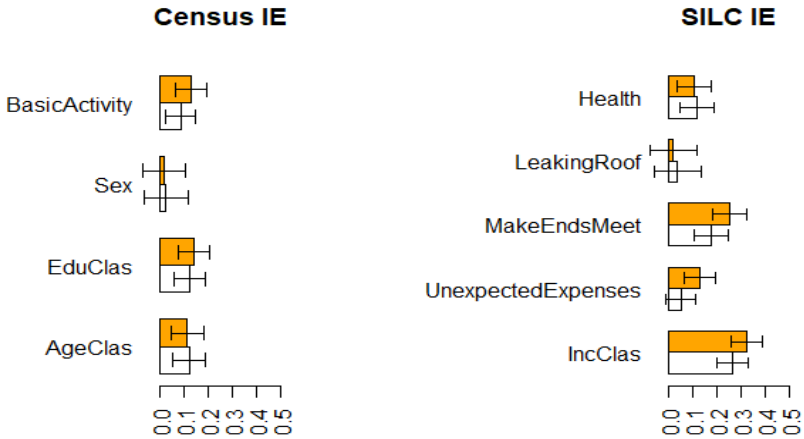


Figure 6.2 continued

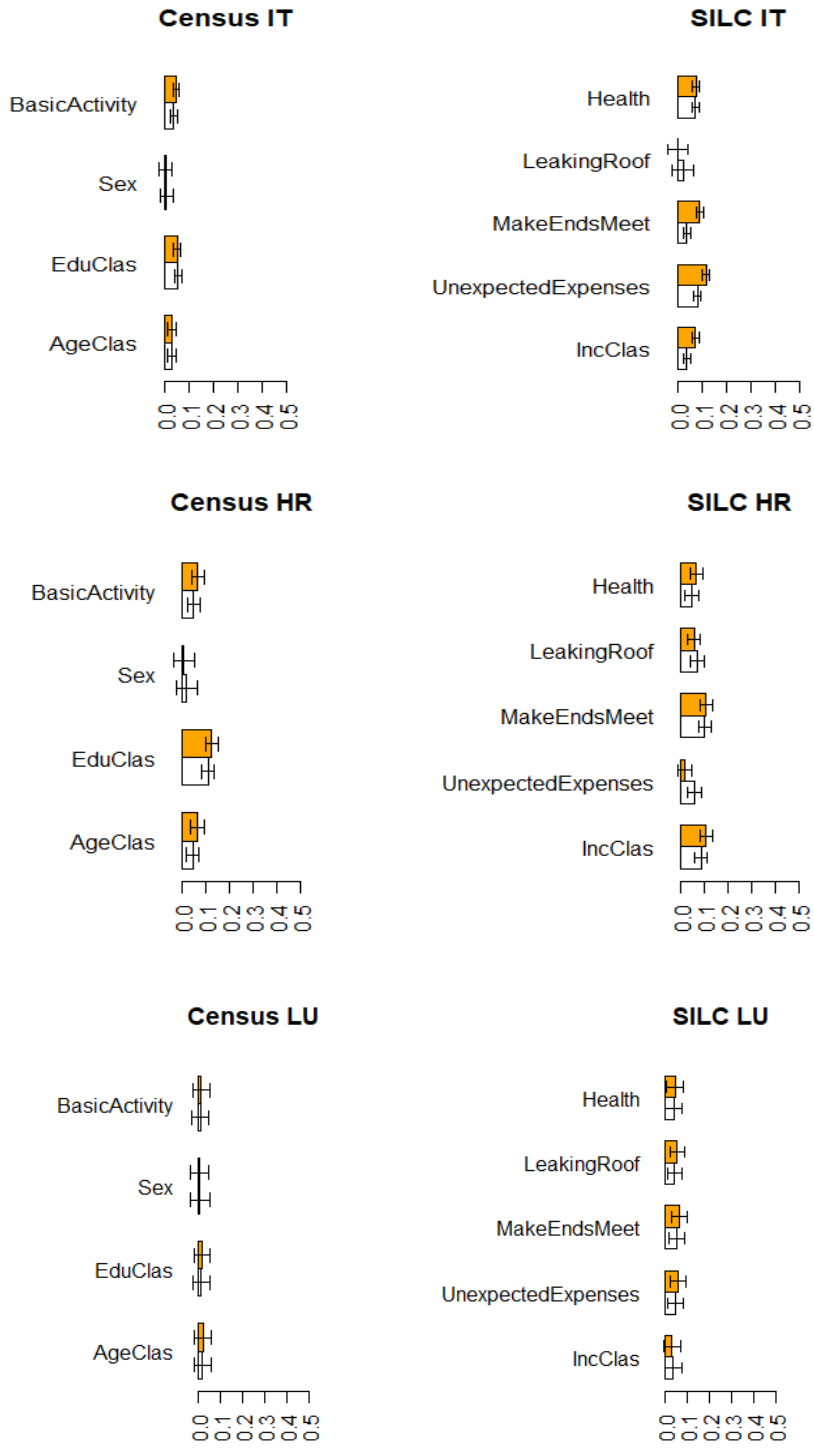


Figure 6.2 continued

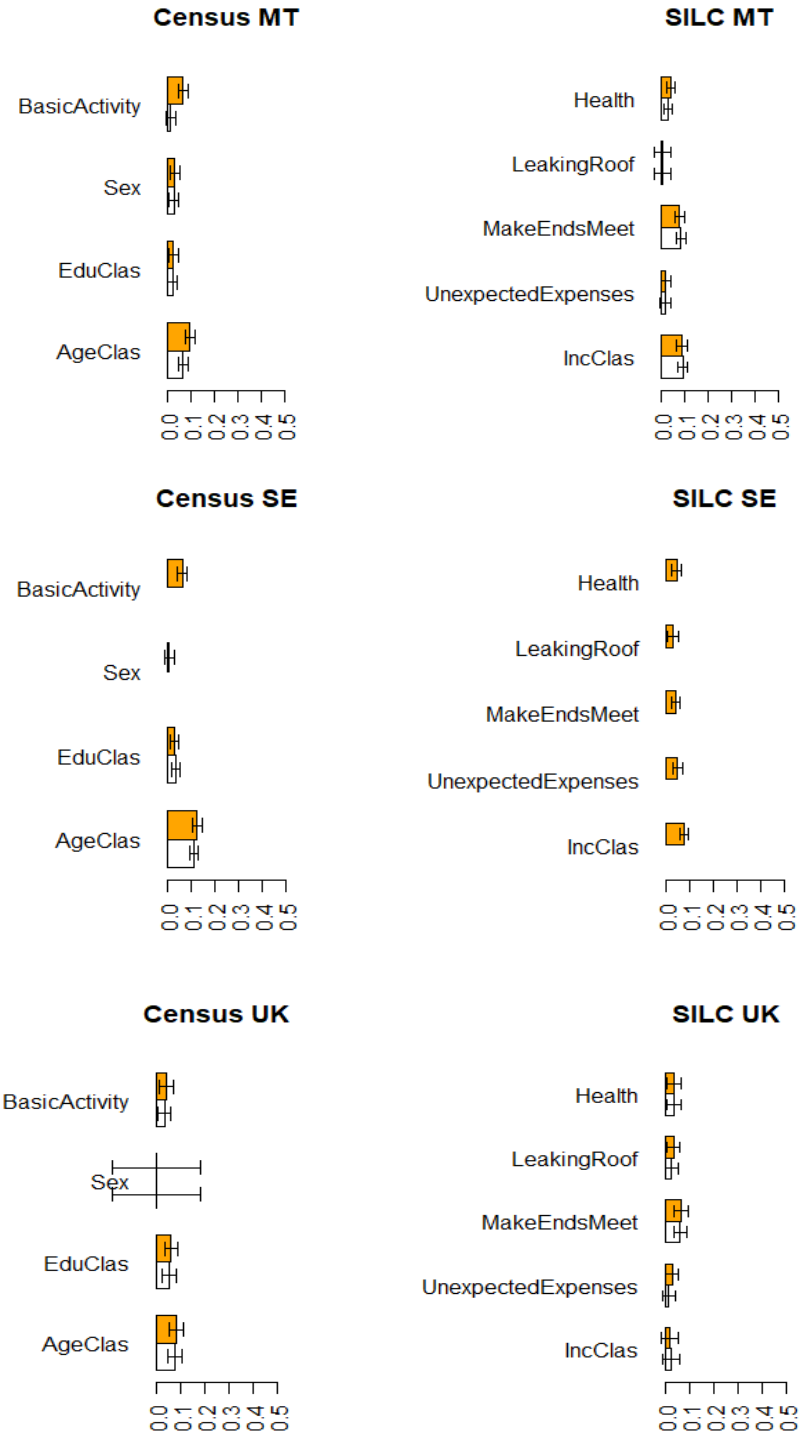
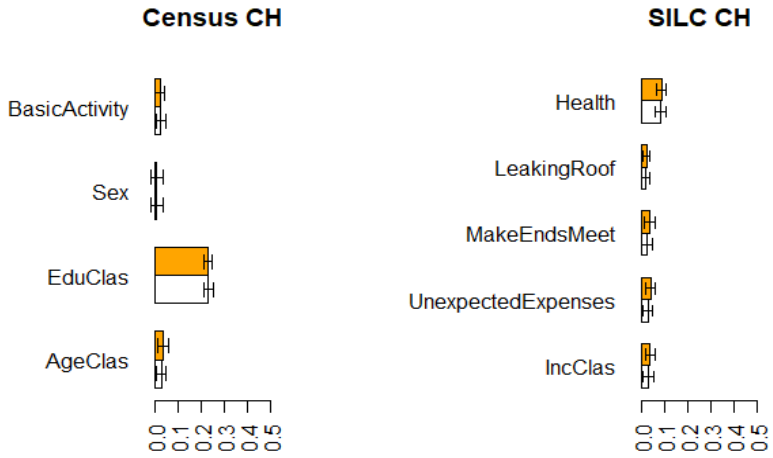


Figure 6.2 continued



NB: Yellow bars represent unconditional values and white bars represent conditional values. The 95 % confidence intervals based on normal approximations are included.

Figure 6.3: Category-level partial CV for variables and countries

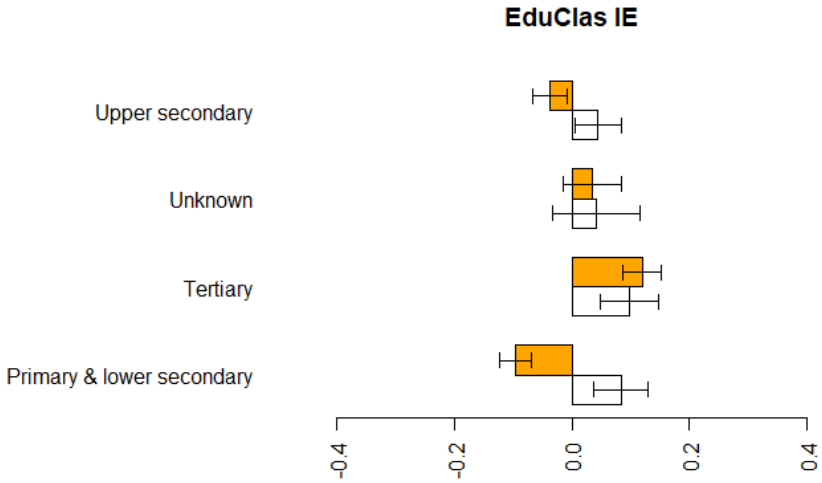


Figure 6.3 continued

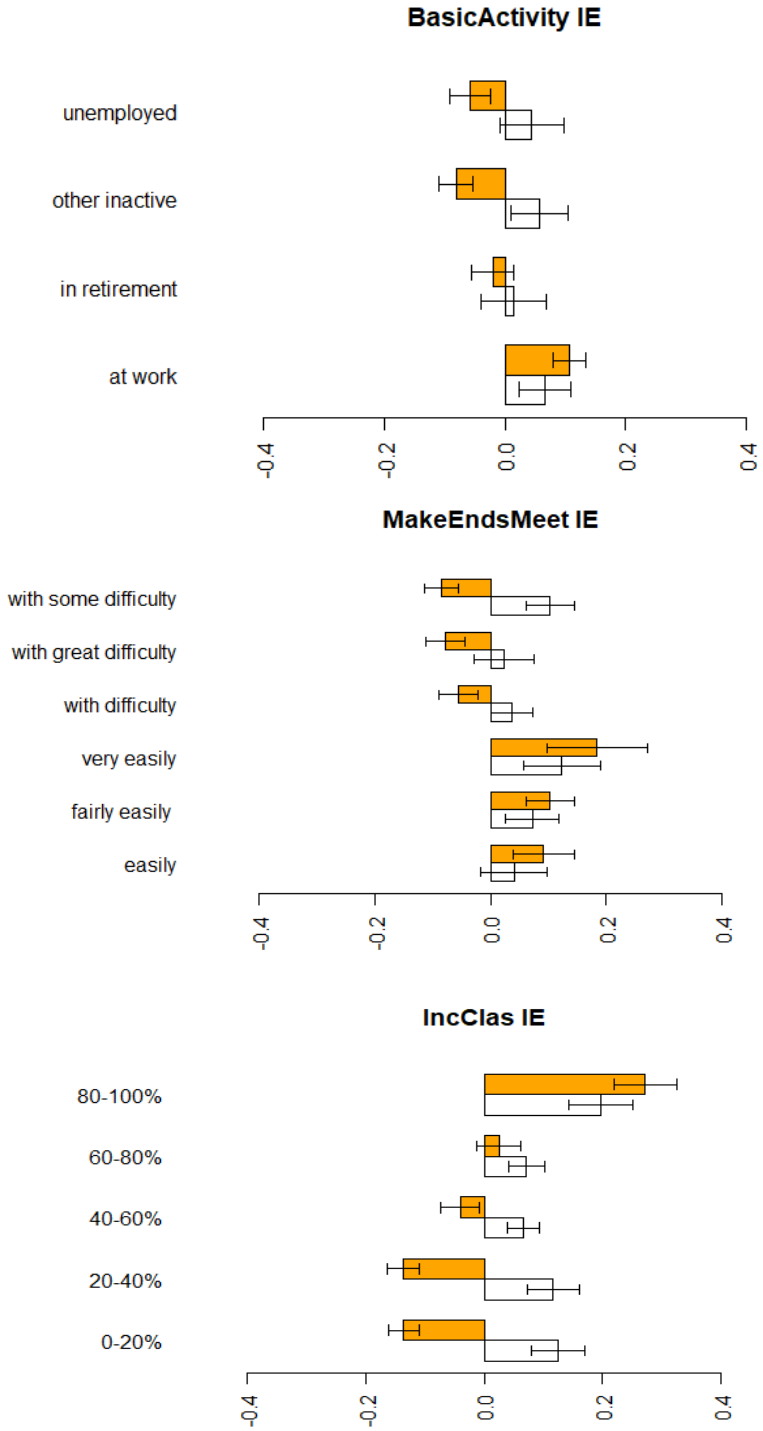
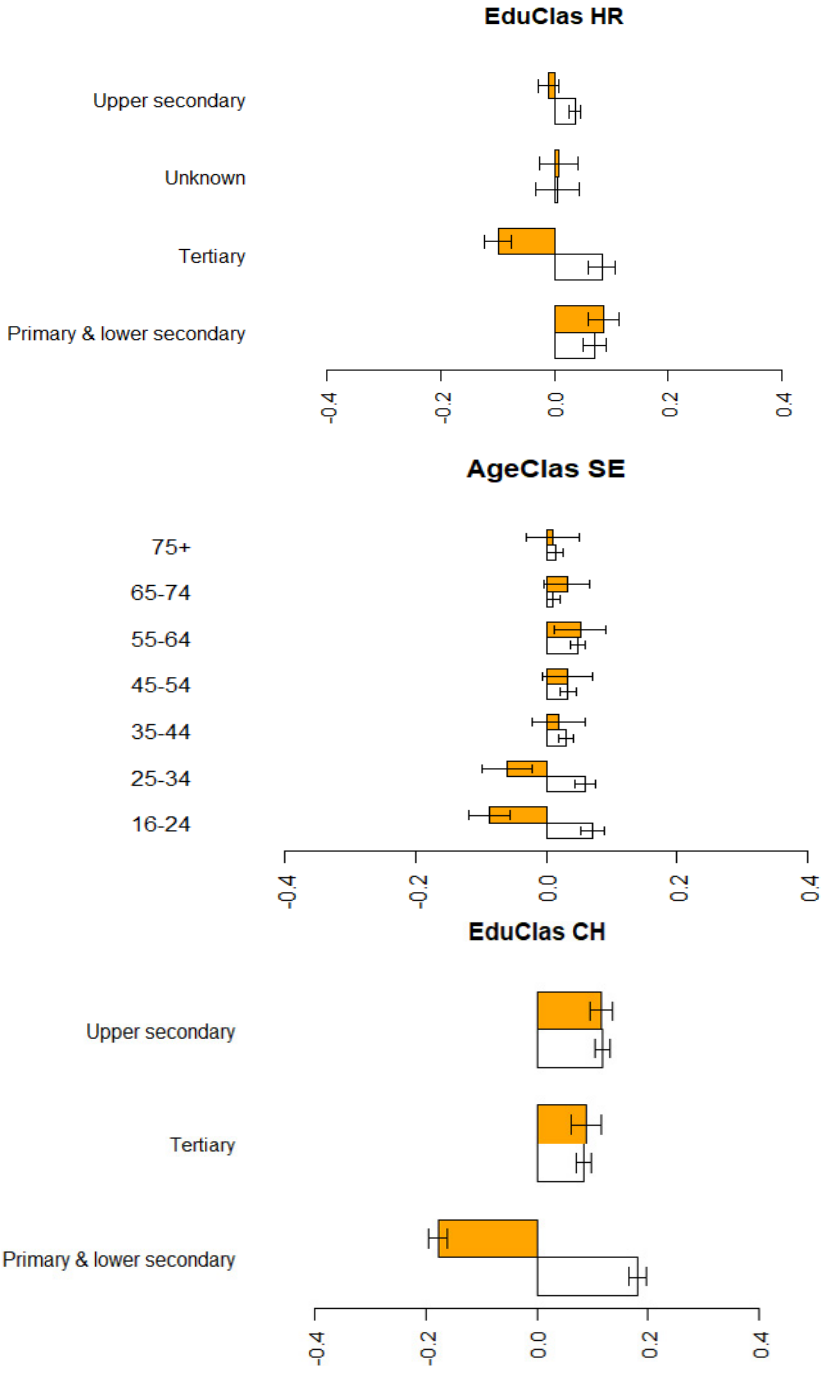


Figure 6.3 continued



NB: Unconditional values are given as orange bars and conditional values are given as white bars. The 95 % confidence intervals resulting from normal approximations are shown.

6.5. Conclusions

In this chapter, we have looked at panel dropout in EU-SILC. We inspected waves 2, 3 and 4 for the panel that started in 2012. It must be noted, therefore, that results about increased risk of bias apply to this period and may have changed in later panel years. Nonetheless, our inspection provides insight into the risk of non-response bias for the 25 available countries.

Our general impression is that risk of bias due to panel attrition is relatively small in many countries and panel waves. Both general census variables and key EU-SILC variables show relatively small CVs for the majority of countries and waves. In some cases, panel dropout is considerable but does not lead to a strong increase in response propensity variation. In some cases, both dropout and variance are small to begin with. We also found a number of countries that show an increased risk of bias in the selected panel years. Given the earlier remark that we inspected only 2012–2015, these findings may no longer be relevant. However, an analysis of the sort we have presented here may usefully be repeated for other years or even become a standard analysis for all longitudinal EU-SILC data sets. The R code that we used for the analysis can be shared for this purpose. A dashboard may be added to aid visualisation and more direct inspection.

A positive finding is that census variables and EU-SILC variables give similar signals about bias. Although this does not necessarily translate into wave 1 representativeness, it is useful knowledge. Obviously, other EU-SILC variables may be included in the analysis. We selected five variables here, but this set may be revised or expanded.

Perhaps one of the most important EU-SILC variables is disposable household income. This variable was categorised into quintiles and included in the analysis. In general, the variable did not show large contributions to CVs in most countries, which again is a positive result. In this deliverable, we chose to categorise income into country-specific quintiles. This choice implies that the income class variable is country specific, adding a country component to the representativeness comparison as well. We leave it open for debate whether income should be characterised differently.

References

Schouten, B., Cobben, F. and Bethlehem, J. (2009), 'Indicators for the representativeness of survey response', *Survey Methodology*, Vol. 35, No 1, pp. 101–113.

Schouten, B., Shlomo, N. and Skinner, C. (2011), 'Indicators for monitoring and improving representativeness of response', *Journal of Official Statistics*, Vol. 27, No 2, pp. 231–253.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016), 'Does balancing survey response reduce nonresponse bias?', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 179, No 3, pp. 727–748.

7 The effect of proxy responses on non-response error

Peter Lynn ([57])

7.1. Introduction: proxy response

Proxy response refers to the situation when data pertaining to one individual (the target individual) are provided by a different individual (the respondent). Allowing proxy responses in a survey can involve a trade-off between non-response error and measurement error. This chapter seeks to identify the extent and nature of proxy responding in EU-SILC and its impact on non-response error. Specifically, variation in the extent of proxy responding over time, between countries and across survey waves is studied, and the characteristics of sample members for whom a proxy response is obtained are compared with the characteristics of those who give a personal interview. Additionally, the impact of proxy responses on survey estimates is illustrated. Considerable variation in the extent of proxy responding is found between countries and over time, including apparently extensive non-compliance with European Union Statistics on Income and Living Conditions (EU-SILC) regulations. Propensity for proxy response is strongly associated with age, gender and activity status, with consequent potential for non-response error reduction.

In some surveys, proxy responding is not permitted and cannot occur, but in many surveys this situation does arise. However, the circumstances in which proxy responding occurs, and hence the frequency and extent of proxy responding, vary greatly between surveys. Some surveys are designed such that certain items are always collected by proxy. A common example is that information about children is often reported in surveys by the parents or guardians of the children. Another example is that surveys sometimes ask respondents to report basic information about other members of their household. Other surveys prefer not to collect information by proxy but allow proxy responding as a 'second-best' or 'last-resort' method of data collection, in the event that the preferred option of personal reporting proves impossible.

EU-SILC falls into this last category. Aside from the situation in which variables are extracted from registers, personal interviewing of each adult member of a EU-SILC household is preferred, but proxy responding is permitted, though the rules differ between types of data/modules. With respect to income (all individuals in a household) and health and labour measures (one person per household), Regulation (EC) No 1177/2003 (European Parliament, 2003) states that the mode of collection should be 'Personal information collected from all household members aged 16 or over (proxy as an exception for persons temporarily away or incapacitated) or extracted from registers'.

Thus, it is clear that proxy responding should be allowed only in exceptional circumstances when there is no possibility of carrying out a personal interview with the respondent. By implication, this does not include situations in which the target individual is simply hard to find at home: in

([57]) Peter Lynn is with the Institute for Social and Economic Research at the University of Essex, Colchester, United Kingdom. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Peter Lynn (plynn@essex.ac.uk).

such situations interviewers must persist with multiple contact attempts. The rules are, however, less stringent with respect to the basic data and education and labour information that are to be collected regarding all household members. Here, the regulation states that the mode of collection should be 'Preferably by personal contact but proxy accepted as a normal procedure or extraction from registers'.

The methodological guidelines (DocSILC 065 ([58])) expand further on the reasoning behind the rules and the situations in which proxy responses are acceptable:

> *Set (iv) variables will normally be collected through direct personal interview in all countries. These are too complex or personal in nature to be collected by proxy; nor are they available from registers or other administrative sources.* (p. 25)
>
> *The proxy rate shall be kept as limited as possible for the income personal variables and for any variables required for at least one household member aged 16 or over.* (p. 32)
>
> *Sample persons aged 14+ who reside temporarily in a collective household or institution but who are still considered as members of a private household are traced and, if aged 16+, are to be interviewed by proxy.* (p. 51)
>
> *It is only under special circumstances (absence, illness, incapacity, …) where the individual is unable to directly provide the requested information through personal interview, that a personal interview with another member of the household (proxy), a telephone interview with the individual or a self-administration of the questionnaire by the respondent are the recommended methods.* (p. 80)
>
> *Proxy interviews are to be especially avoided for both income variables, health and detailed labour information.* (p. 80)
>
> *If the relevant persons are temporarily absent or incapacitated, proxy interviews are allowed as an exception.* (p. 347)

([58]) https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/c91ee3a0-4a48-41f2-917f-a176c1bb3c1a/details

7.2. Advantages and disadvantages of proxy responses

The benefits of allowing proxy responses relate to costs and non-response (error). In the situation in which one household member is found to be available and willing to be interviewed but one or more other household members are not immediately available, it is clearly more convenient and less costly for the interviewer to ask the available household member to provide proxy responses for the absent member(s) than to make further attempts to contact the other household member(s) in person. Proxy responses may also help to boost response rates if some of the target individuals for whom proxy responses are obtained would not have participated in the survey had a personal interview been required. This higher response rate could also translate into reduced non-response error if the sample members for whom proxy responses are obtained are systematically different in important ways from other respondents. There is some evidence from Understanding Society – the UK household longitudinal study – that proxy respondents are indeed systematically different from others: for example, 14.0 % are aged 20–24, compared with 6.6 % of personal respondents (in most of these cases, the proxy response is provided by a parent).

However, the main disadvantage of proxy responses is that they can be inferior to personal responses in terms of quality (Blair, Menon and Bickart (1994), Moore, 1988). It is of particular concern if proxy responding introduces systematic measurement error. It is because of this concern that surveys often do not permit, or often discourage, proxy responding.

Cobb (2018) concludes that quite a lot is known about the situations in which the quality of proxy responses may differ from that of first-person responses, although serious gaps remain in our knowledge. For example, proxy respondents are more likely to provide incomplete data (King, Cook and Hunter Childs, 2012) or answers that are estimates rather than recalled values (Schwarz and Wellens, 1997), and are more likely to provide the

same response that the target respondent would have given if they spend more time with that person (Amato and Ochiltree, 1987; Cohen and Orum, 1972) or discuss the topic with that person (Bickart, Phillips and Blair, 2006). Consistent with the above, Imara (2013) found that agreement in responses between a self-report and a proxy report was more likely when the proxy respondent was a spouse rather than another household member. On employment-related topics, Boehm (1989) found agreement rates that ranged from 92 % regarding whether the target respondent was paid an hourly rate or a salary to 67 % regarding whether they had worked overtime last week. Similarly, Dawe and Knight (1997) found 94 % agreement on whether the target respondent worked full-time or part-time, but only 75 % agreement on occupation (three-digit code level). Grootendorst, Feeny and Furlong (1997) concluded that the quality of proxy responses was unacceptably low for measures of morbidity, while Highton (2005) found that the quality of proxy responses of voter turnout was comparable to that of personal responses.

Thomsen and Villund (2011) attempt to estimate simultaneously the effects of proxy responses on both measurement error and non-response error, using external register data as a benchmark. They conclude that, overall, proxy responses improve estimates of employment rate: the reduction in non-response error outweighs the increase in measurement error.

7.3. Objectives of this chapter

An important question for EU-SILC is whether the advantages of proxy responding outweigh the disadvantages. An evaluation of this question could lead to reconsideration of the EU-SILC rules and guidance on proxy interviewing. This chapter takes some initial steps towards such an evaluation. In particular, the chapter aims to:

- document the extent of proxy responding in EU-SILC;
- identify patterns and trends in proxy responding over time, across waves and between countries;
- provide illustrative estimates of the effects of proxy responding on marginal non-response error.

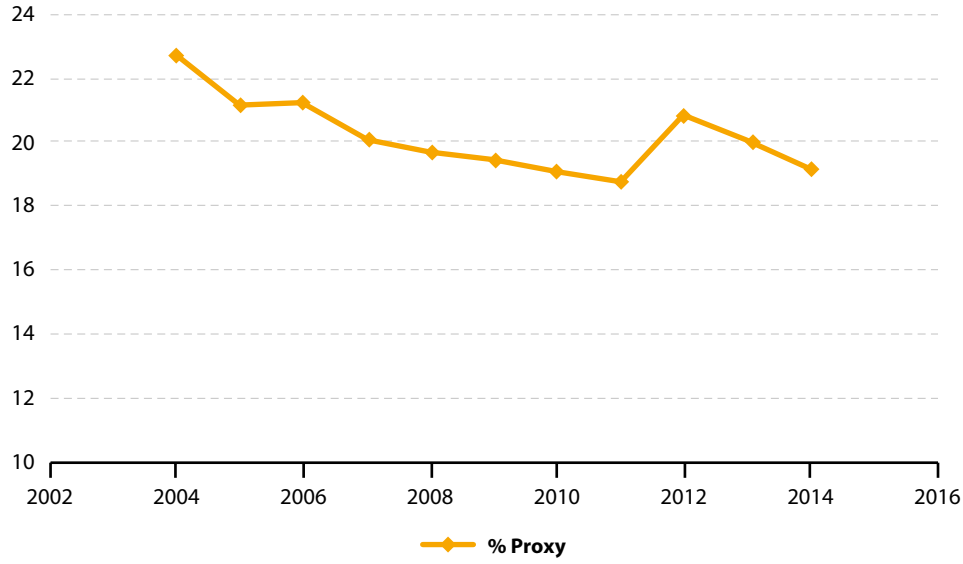
The chapter does not address measurement error.

7.4. Levels of proxy responding over time, between countries and across waves

The overall extent of proxy responding in EU-SILC is high. Across all interviews in the user database (UDB) carried out between 2004 and 2014, 20.1 % were carried out by proxy (959 247 interviews out of 4 781 514). This proportion varied somewhat over the years (Figure 7.1), falling gradually from 22.8 % in 2004 to 18.8 % in 2011, before increasing again to 20.9 % in 2012 and then falling again to 19.1 % in 2014. It should be noted at this point that the indicator of whether response was provided in person or by proxy (RB260) is defined at the level of a data record (sample individual year), and we must therefore assume that all survey data within that record were obtained in the same way. In other words, although the regulations allow for proxy responding for basic and education data but not for income or health data (except in exceptional circumstances), it would appear that the extent of proxy reporting is the same for both types of data.

The extent of proxy reporting varies greatly between countries. Figure 7.2 shows that proxy reporting is extremely common (over 40 % of interviews are carried out by proxy) in Denmark, Finland and Croatia (listed in descending order), but rather uncommon (6 % or less) in Greece, Slovakia, Switzerland and Sweden (listed in descending order), whereas there are apparently no proxy interviews at all in Iceland. Between those extremes are a set of six countries with proxy reports in the range of 25–30 %, 15 countries in the range of 15–25 % and three countries in the range of 10–15 %.

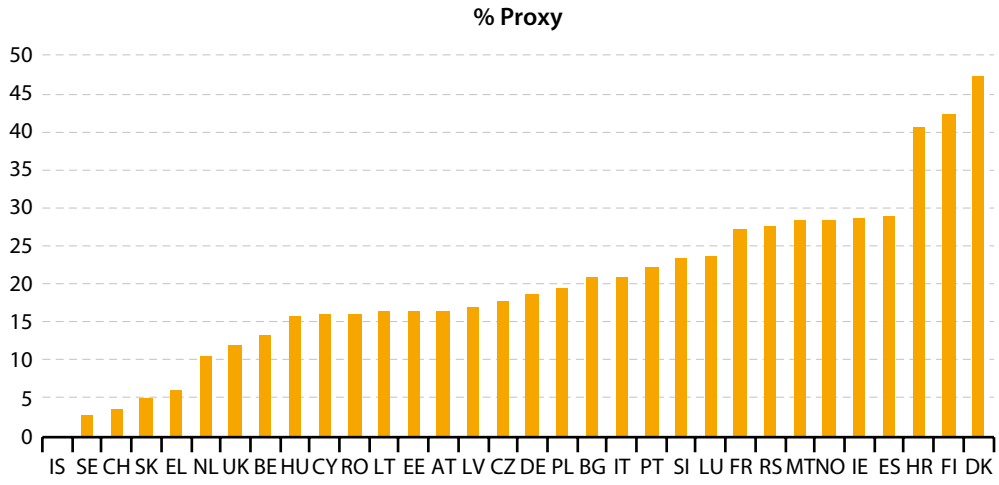
Figure 7.1: Percentage of interviews carried out by proxy, by survey year, 2004–2014



NB: The *y*-axis indicates the percentage of individual interviews that were carried out by proxy in each year.

Source: EU-SILC, pooled cross-sectional UDBs, 2004–2014.

Figure 7.2: Percentage of interviews carried out by proxy, by country, 2004–2014



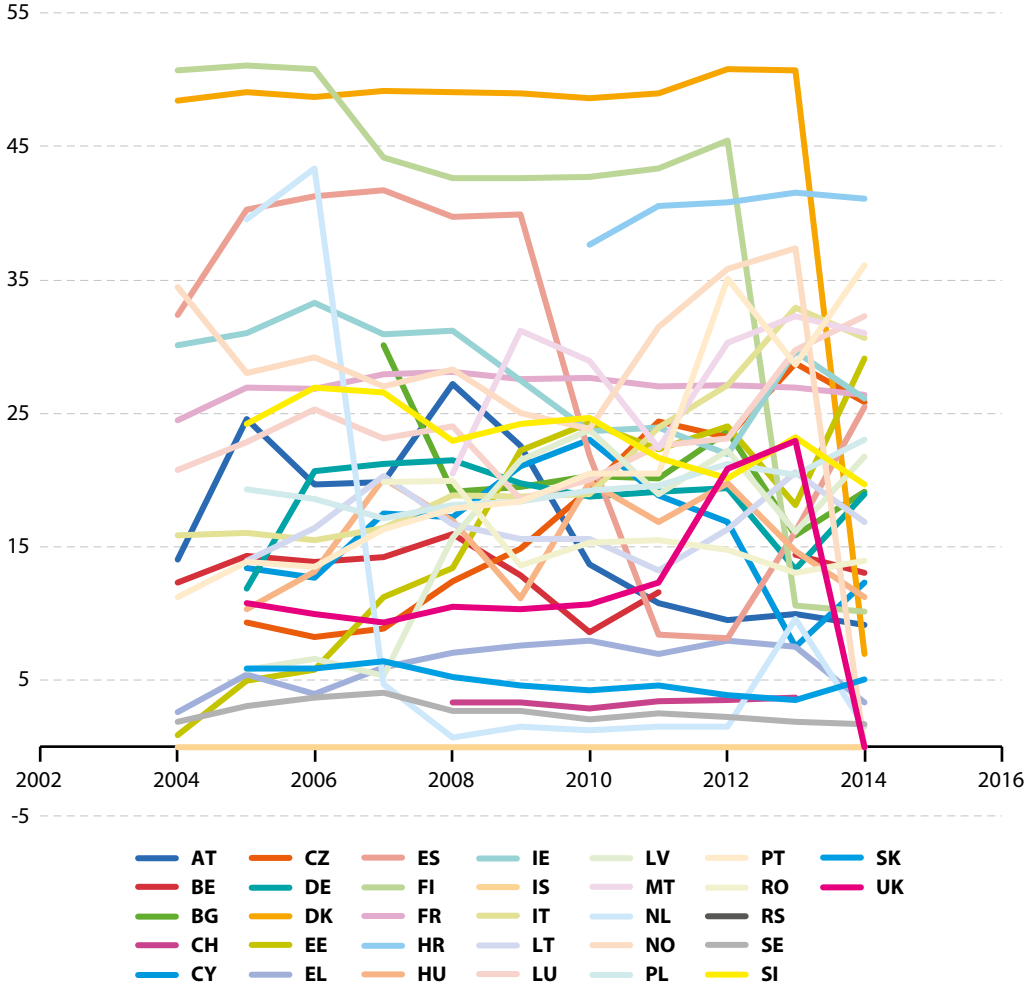
NB: Countries are presented in ascending order of the percentage of proxy responses. The *y*-axis indicates the percentage of individual interviews that were carried out by proxy in each country.

Source: EU-SILC, pooled cross-sectional UDBs, 2004–2014.

As well as differing in overall levels of proxy reports, countries also differ in terms of the trend over time (Figure 7.3). Some countries have experienced a steady increase in the extent of proxy reporting over the years; others have experienced a steady decrease; some have experienced stability; and a few have experienced big jumps at some point. Specifically, there are six countries that exhibit a steady upward trend over time (Figure 7.4): Czechia, Estonia, Italy, Latvia, Portugal and the United Kingdom. (The United Kingdom is an interesting case, as the percentage appears to fall to zero in 2014, but rather than reflecting reality it seems that there was simply a failure to populate the proxy indicator variable.)

Four countries can be seen to have experienced a sharp drop in the levels of proxy reporting, from a previous higher level to a new lower level (Figure 7.5). This happened in the Netherlands in 2007, Spain in 2010 and 2011, Finland in 2013 and Denmark in 2014.

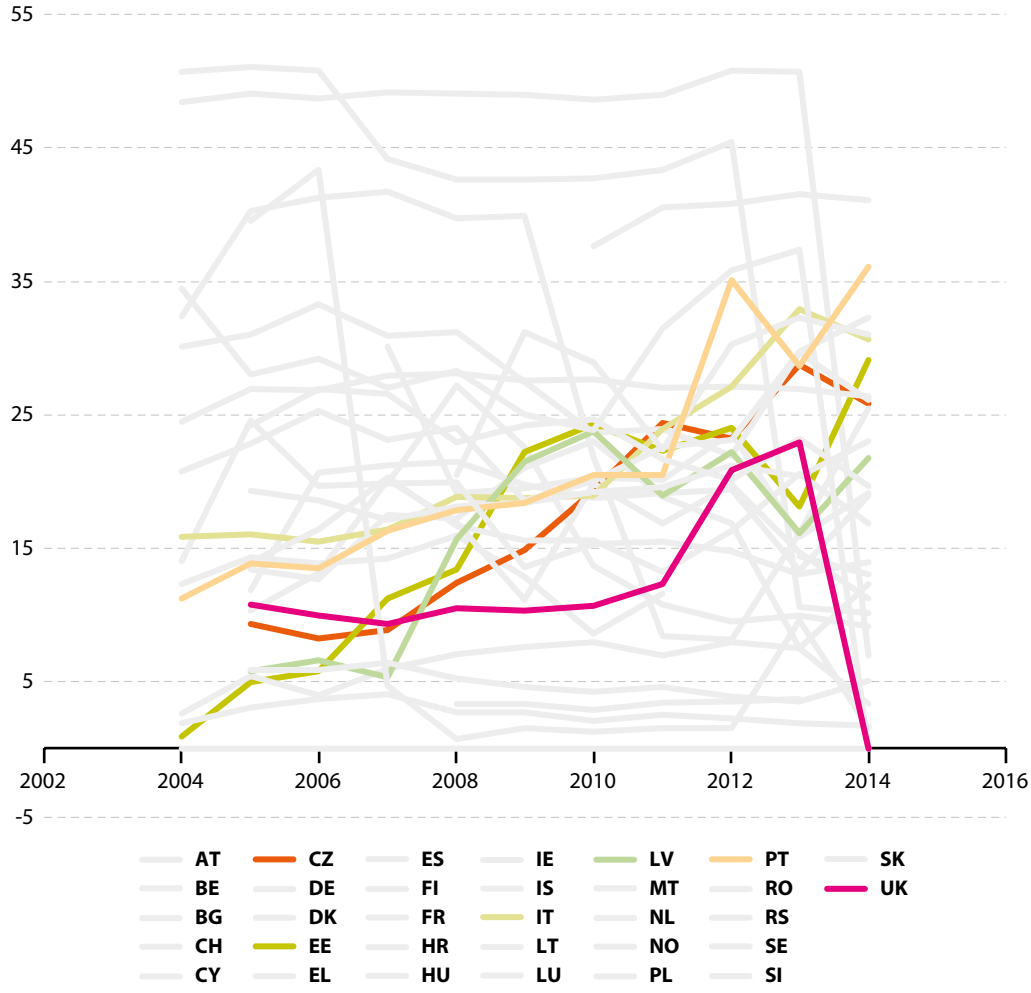
Figure 7.3: Percentage of interviews carried out by proxy, by survey year and country, 2004–2014



NB: The *y*-axis indicates the percentage of individual interviews that were carried out by proxy. Each line in the chart represents a different country, as indicated by the key below the figure.

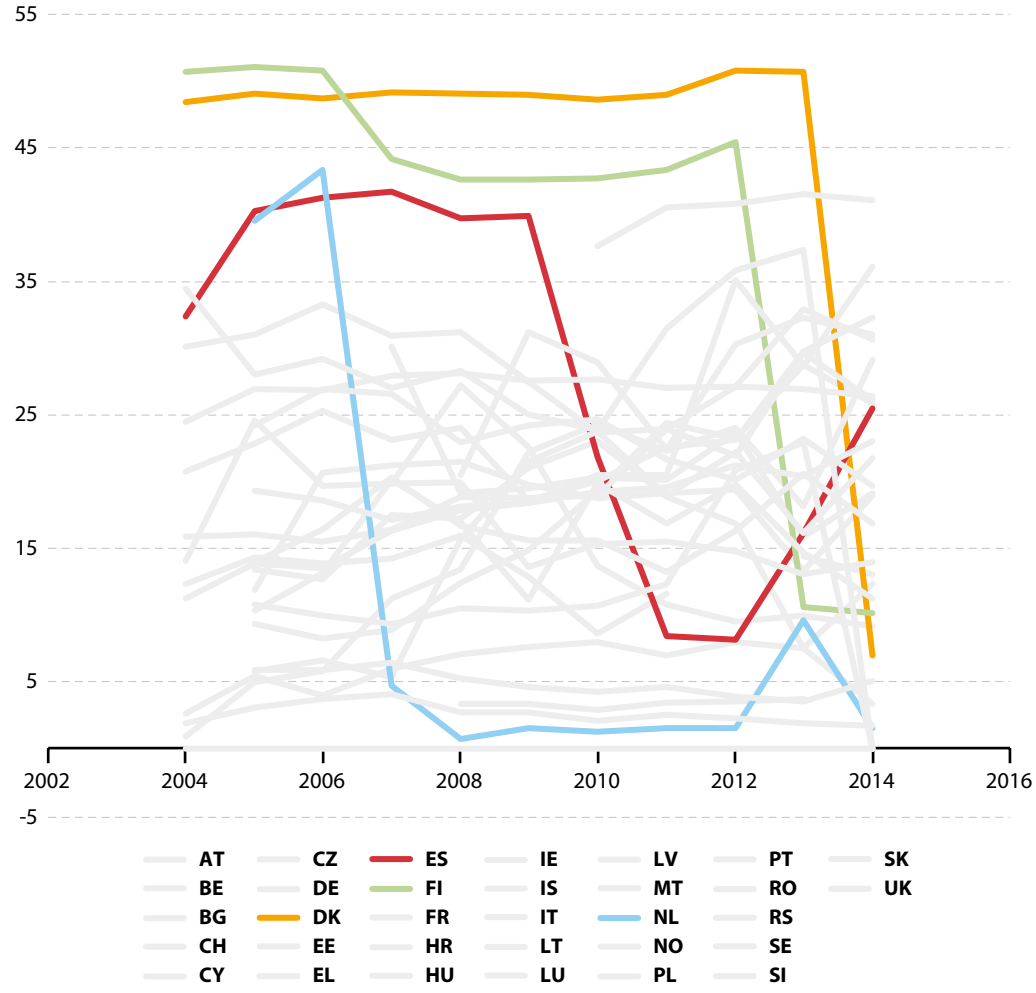
Source: EU-SILC, pooled cross-sectional UDBs, 2004–2014.

Figure 7.4: Countries with an upward trend in the percentage of proxy reports, 2004–2014



NB: See Figure 7.3 for notes and source.

Figure 7.5: Countries with a sharp drop in the percentage of proxy reports, 2004–2014

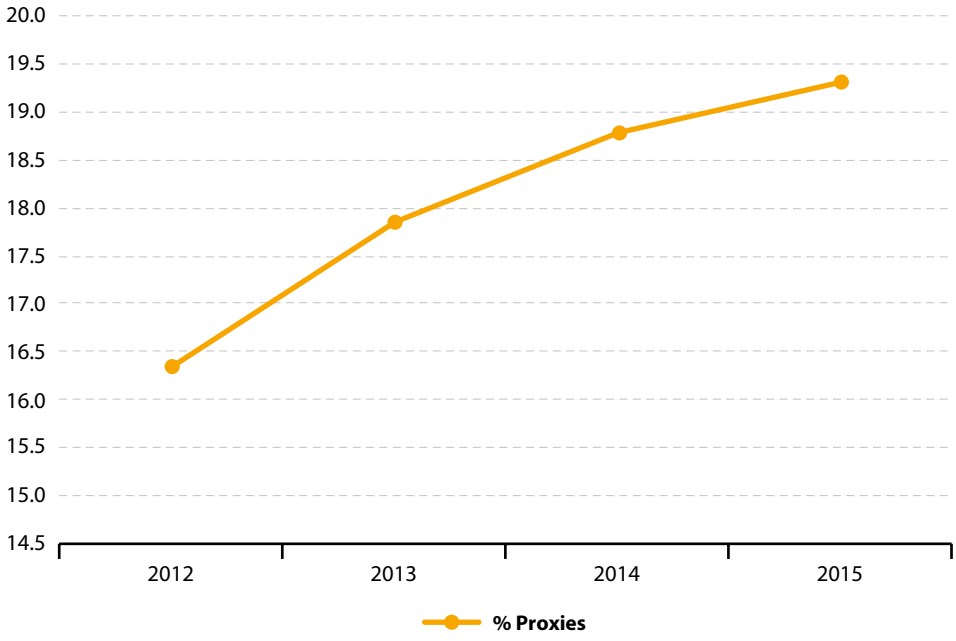


NB: See Figure 7.3 for notes and source.

There is a modest tendency towards increasing levels of proxy response across waves of a panel. The analysis presented in Figure 7.6 is restricted to the balanced panel, by which we mean the subset of sample individuals who participated in all four waves between 2012 and 2015. This is done in order to isolate the effect of survey wave from any possible selectivity due to attrition. For the 2012–2015 balanced panel, the percentage of proxy responses rises from 16.3 % at wave 1 to 19.3 % at wave 4.

However, this overall variation across waves does not reflect the pattern in any individual country (Figure 7.7). Only 8 of the 27 countries – Estonia, Spain, Italy, Latvia, Lithuania, Malta, Austria and Poland – show increasing levels of proxy responses across waves (Figure 7.8), and the pattern is far from consistent even in these countries. Interestingly, six countries – Belgium, Czechia, Croatia, Italy, Lithuania and Slovenia – appear to show a peak in proxy responding at wave 2 (Figure 7.9).

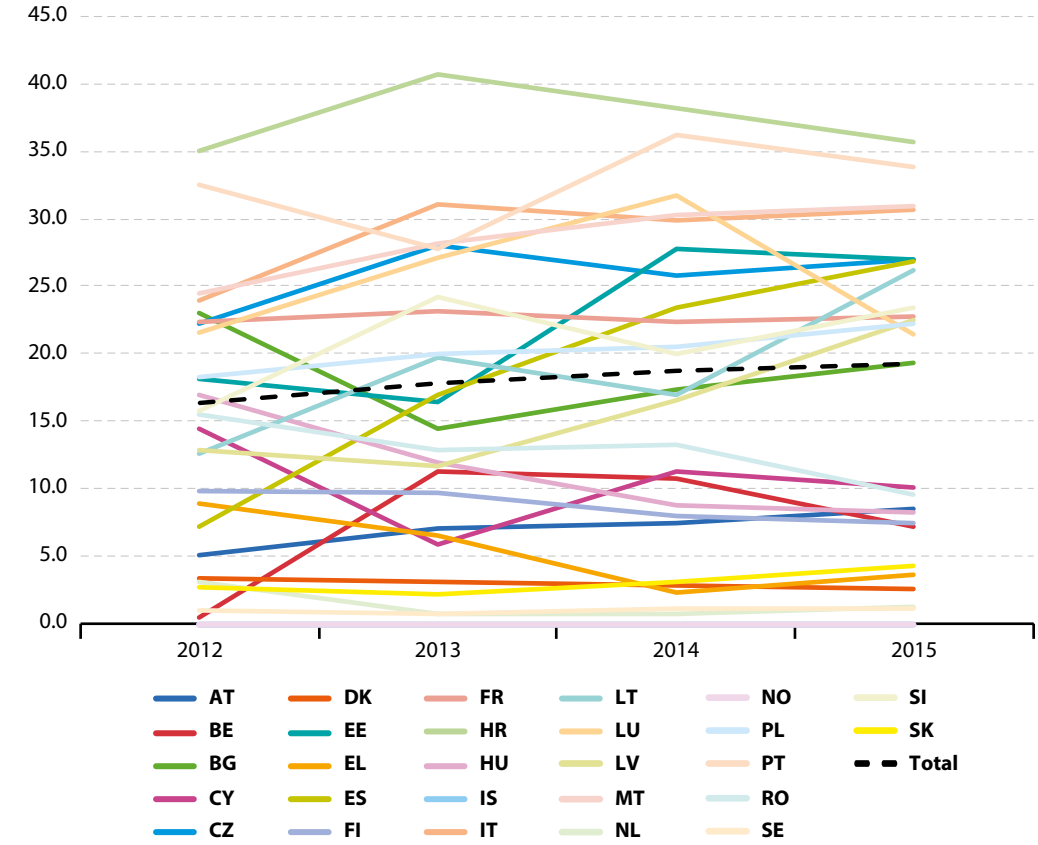
Figure 7.6: Variation in the percentage of proxy reports across waves, 2012–2015 balanced panel



NB: The analysis is restricted to individuals who participated in EU-SILC in each of the survey years 2012, 2013, 2014 and 2015, and for whom 2012 was the first year of inclusion of their rotational group. We refer to this as the balanced panel. *n* = 75 107. The *y*-axis indicates the percentage of individual interviews that were carried out by proxy.

Source: EU-SILC, longitudinal UDB, 2015.

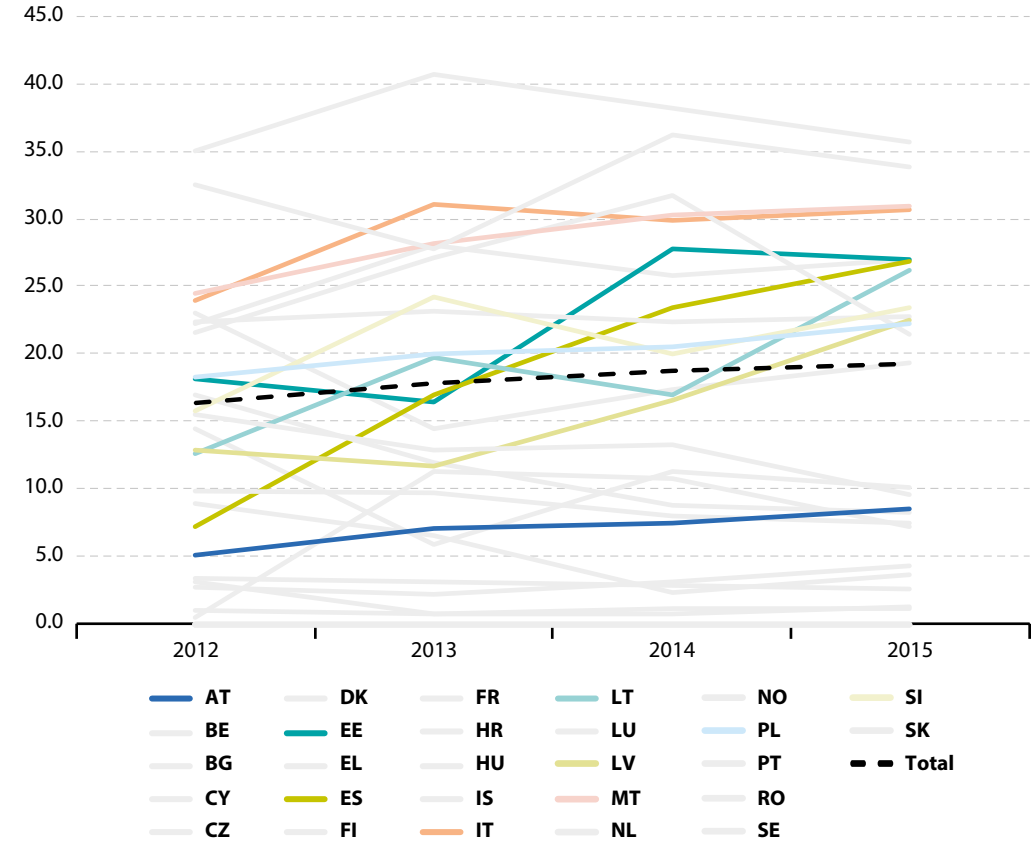
Figure 7.7: Variation across waves and between countries, 2012–2015 balanced panel



NB: See Figure 7.6 for notes.

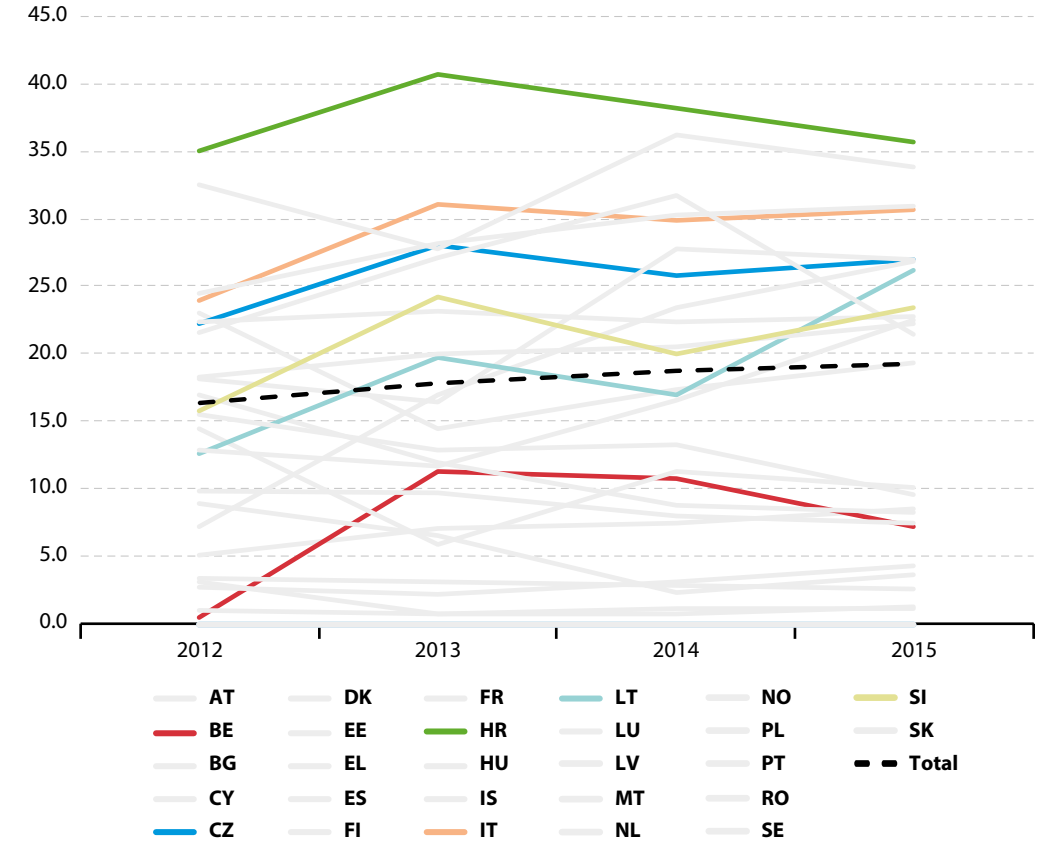
Source: EU-SILC, longitudinal UDB, 2015.

Figure 7.8: Countries with an increase across waves, 2012–2015 balanced panel



NB: See Figure 7.6 for notes.
Source: EU-SILC, longitudinal UDB, 2015.

Figure 7.9: Countries with a peak at wave 2, 2012–2015 balanced panel



NB: See Figure 7.6 for notes.
Source: EU-SILC, longitudinal UDB, 2015.

7.5. Characteristics of sample members with proxy reports

In this section, sample members for whom proxy reports are obtained are compared with sample members who give personal interviews, in terms of a range of important sociodemographic characteristics. This analysis gives a first impression of the possible impact of proxy responses on non-response error, under the – possibly unrealistic – assumption that these data would not have been obtained if proxy interviewing was not permitted. To interpret the findings as an indication of potential selection effects also requires an assumption that there are no systematic differences in measurement error between personal interviews and proxy interviews. The literature summarised in Section 7.2 suggests that this assumption may also not necessarily be warranted, although measurement effects are likely to be minimal for the variables reported here. In the absence of experimental data, however, there is no way of testing the assumption.

It can be seen (Table 7.1) that proxy responses are more likely to be obtained for younger sample members and for males. The age gradient is particularly noticeable. Proxy responses are also far more likely for co-residents (46 %) than for sample individuals (18 %), although co-residents are of course a very small group.

Proxy response is especially common when the target individual is either a student or on military service (Table 7.2), and relatively uncommon for people who are retired or in part-time employment. The differences in the rate of proxy responses between different groups are substantial: the rate ranges from 11.2 % for retired people to 38.8 % for students. Differences are also observable by educational level (Table 7.2), although they are not so pronounced as for activity status. The proxy rate is relatively low for those with post-secondary or tertiary education (13.4 % and 14.6 %, respectively) and highest for those with only pre-primary education (20.4 %) or for whom information on the highest educational level is missing (23.9 %).

Table 7.1: Proxy responses, by age, gender and sample status, 2012–2015 balanced panel, individuals

Characteristic	Personal (%)	Proxy (%)	Base
Age group (derived from PB140)			
16–24	58.3	41.7	32 460
25–34	78.7	21.3	37 264
35–44	82.0	18.0	50 760
45–54	84.4	15.6	56 540
55–64	87.3	12.7	58 164
65–74	89.4	10.6	40 256
75+	87.2	12.8	24 984
Gender (RB090)			
Male	76.7	23.3	140 048
Female	86.5	13.5	160 344
Sample status (RB100)			
Sample person	82.0	18.0	299 632
Co-resident	54.3	45.7	796
Total	81.9	18.1	300 428

NB: Proxy status derived from RB260. Figures are row percentages. For example, 58.3 % of interviews with people aged 16–24 were carried out in person and 41.7 % were carried out by proxy.

Source: EU-SILC UDB, 2012–2015 balanced panel, all countries.

Table 7.2: Proxy responses, by activity status and highest educational qualification, 2012–2015 balanced panel, individuals

Characteristic	Personal (%)	Proxy (%)	Base
Main activity (derived from PB140)			
Full-time employment	80.3	19.7	108 767
Part-time employment	87.7	12.3	15 051
Self-employment	81.2	18.8	21 680
Unemployed	78.2	21.8	20 182
Student	61.2	38.8	21 465
Retired	88.8	11.2	80 750
Disabled	82.5	17.5	9 420
Military service	65.9	34.1	420
Care/home	84.8	15.2	22 641
International Standard Classification of Education (derived from PE040)			
Pre-primary	79.6	20.4	4 971
Primary	83.1	16.9	35 139
Lower secondary	80.3	19.7	52 873
Upper secondary	80.3	19.7	127 943
Post-secondary non-tertiary	86.6	13.4	9 556
Tertiary	85.4	14.6	67 431
Missing	76.1	23.9	2 515
Total	81.9	18.1	300 428

NB: Proxy status derived from RB260. Figures are row percentages. For example, 80.3 % of interviews with people in full-time employment were carried out in person and 19.7 % were carried out by proxy.

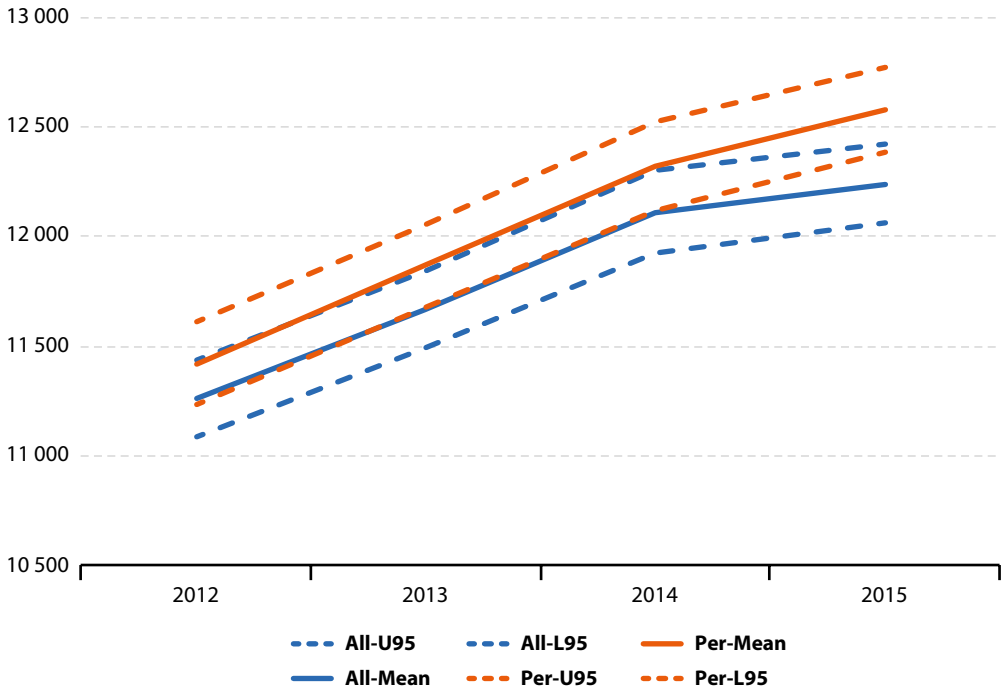
Source: EU-SILC UDB, 2012–2015 balanced panel, all countries.

7.6. Effects of proxy reports on estimates

Here, we assess the effects on selected estimates of excluding proxy responses. This analysis is intended to be indicative of the likely nature of any marginal non-response bias that would be introduced if proxy responding was no longer permitted, although the assumptions mentioned earlier should be borne in mind.

Example 1. Using the 2012–2015 balanced panel, we analyse the effects of proxy responses on the mean amount of old-age benefit received (among those receiving a non-zero amount) and the growth in old-age benefit across the four waves. We see (Figure 7.10) that adding in proxy interviews reduces the mean income from old-age benefit. However, it does so fairly uniformly across the waves, so the observed growth is similar with or without the proxy responses. There are, nevertheless, some small differences between countries. For example, proxy responses reduce the mean income from old-age benefit in Spain, but increase it in both France and Italy (Figure 7.11), although none of these effects are statistically significant.

Figure 7.10: Effect across waves of proxy responses on mean income from old-age benefit, 2012–2015 balanced panel, individuals

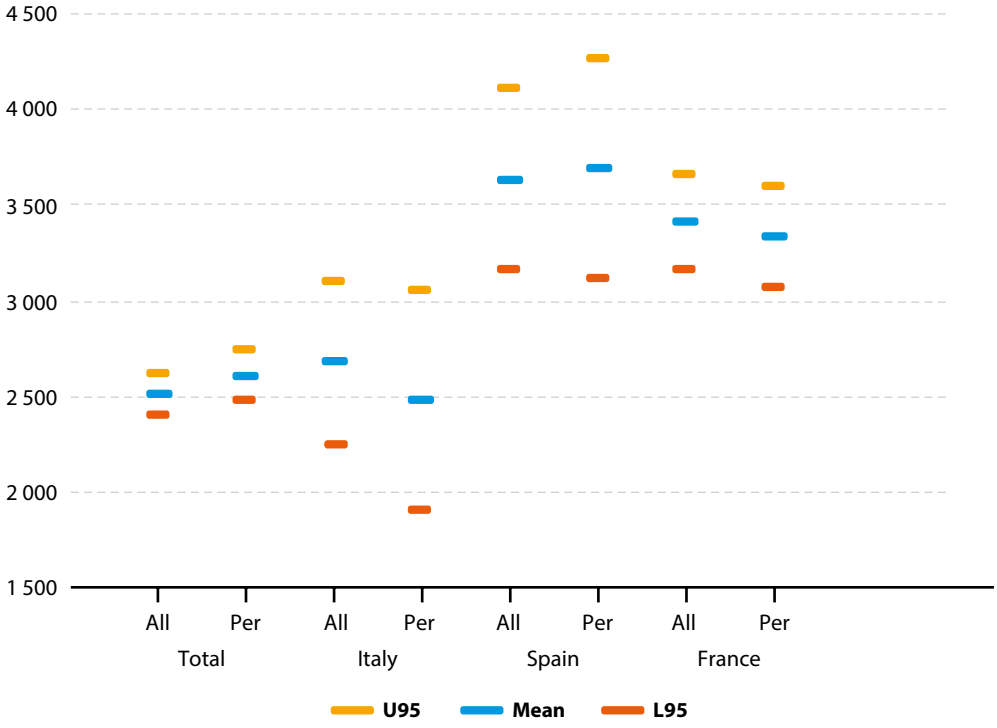


NB: Income from old-age benefit given by EU-SILC variable PY100g, in euro. Individuals with zero reported income from this source are excluded. The *y*-axis plots mean income from old-age benefit. The orange lines represent sample members who gave a personal interview, whereas the blue lines represent all sample members, including those for whom a proxy interview was obtained.

Source: EU-SILC UDB, 2012–2015 balanced panel, all countries.

Example 2. Based on the same balanced panel, we examine the effects of proxy responses on the mean amount of cash or near-cash income from employment (PY010n). Across all countries where this measure is obtained from survey data, we observe that including proxy responses raises the mean slightly (from EUR 6 117 to EUR 6 245) and reduces the standard error (from EUR 27.0 to EUR 24.4). This suggests that the acceptance of proxy responses reduces non-response bias (the mean including proxy responses is well outside the 95 % confidence interval for the mean when proxy responses are excluded) and increases the precision of estimation (due to the larger sample size available). However, we again observe that the effects are not uniform across countries. Figure 7.12 shows estimated 95 % confidence intervals for the mean amount of cash or near-cash income from employment for countries with a mean (all respondents) in the range EUR 2 000–2 600. There are five such countries, and it can be seen that the inclusion of proxy responses raises the mean noticeably for Croatia and Poland, raises it slightly for Hungary, and hardly affects it at all for Greece and Lithuania. The narrowing of the confidence interval is perceptible for all countries except Greece and is largest for Croatia. The differences stem largely from two sources, namely the extent to which countries are reliant on proxy responses and the extent to which proxy respondents are distinctive in terms of earnings. We have already seen (Figure 7.2) that Croatia has one of the highest levels of proxy response, whereas Greece has one of the lowest. This is reflected in the effect of proxy responses on the width of the confidence interval. Proxy respondents are particularly distinctive in

Figure 7.11: Country differences in effect of proxy responses on growth in mean income from old-age benefit, 2012–2015 balanced panel, individuals



NB: Income from old-age benefit given by EU-SILC variable PY100g, in euro. Individuals with zero reported income from this source are excluded. The *y*-axis plots the mean value of the growth in income from old-age benefit between 2012 and 2015. 'Per' is the mean among sample members who gave a personal interview, whereas 'All' is the mean among all sample members, including those for whom a proxy interview was obtained. L95 and U95 present the lower bounds and upper bounds, respectively, of the 95 % confidence interval around the estimate.

Source: EU-SILC UDB, 2012–2015 balanced panel, all countries.

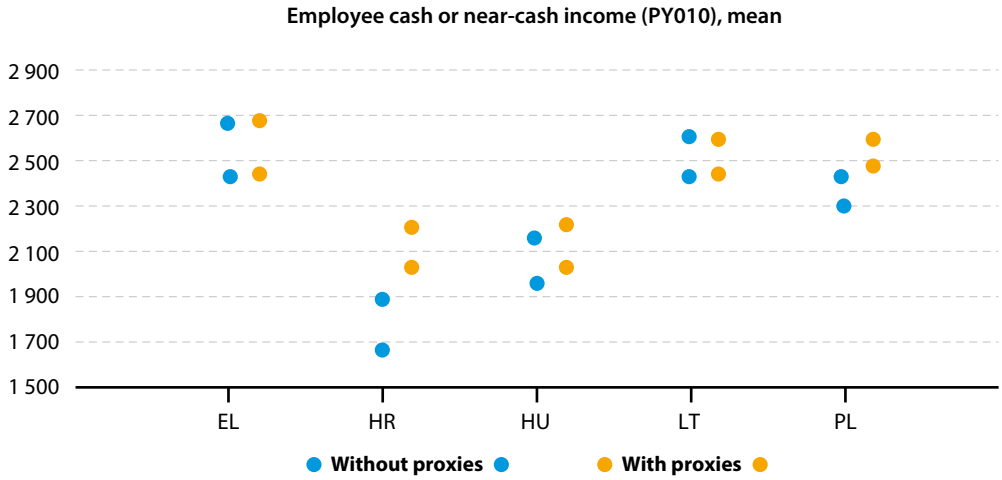
Croatia and Poland, with earnings that are 51 % and 26 % higher, respectively, than those of personal respondents, but not at all distinctive in Lithuania, where earnings are 0.3 % lower for proxy respondents. These systematic differences drive the shift in location of the confidence intervals in Figure 7.12.

7.7. Conclusions

The levels of proxy responding are high in EU-SILC. This appears to be the case even for modules in which proxy responding is supposed to be accepted only in exceptional circumstances, in countries that collect this information from interviews as opposed to registers. These high levels of proxy responding raise some concerns regarding compliance with the EU-SILC methodological guidelines, but, more importantly, they indicate that careful consideration of the use of proxy reporting in EU-SILC and the implications for data analysts may be warranted. This chapter has provided some initial steps in this direction.

The observed levels of proxy responding vary greatly between countries and – in many cases – over time within countries. There are several countries in which proxy reporting seems to have become gradually more prevalent over time, whereas there are some in which a sudden drop in the prevalence is observed. The latter cases are mainly explained

Figure 7.12: Country differences in effect of proxy responses on confidence intervals for mean employee cash or near-cash income, 2012–2015 balanced panel, individuals



NB: Employee cash or near-cash income given by EU-SILC variable PY010n, in euro. For illustrative purposes, the display is restricted to countries with a mean in the range EUR 2 000–2 600. The *y*-axis plots the mean value of employee cash or near-cash income. Orange marks represent the bounds of the 95 % confidence interval for the mean among sample members who gave a personal interview; blue marks represent the equivalent interval for estimation based on all sample members, including those for whom a proxy interview was obtained.

Source: EU-SILC UDB, 2012–2015 balanced panel.

by major changes to data collection protocols in those countries, for example the wide-scale introduction of online data collection.

Within a rotational group (panel), levels of proxy responding increase across the waves in some countries, whereas in other countries the level seems to peak at wave 2. This finding would suggest that longitudinal measures – such as measures of persistent poverty – could be affected if the accuracy of reports differs between personal respondents and proxy respondents. The centrality of such longitudinal measures to EU-SILC would suggest that a study of the measurement quality of proxy responses should be undertaken.

Certain characteristics of sample members are associated with an increased likelihood of a proxy response. These include being a student or on military service, aged under 25, male and a co-resident, and having pre-primary or missing education. It is clear that proxy interviews are not distributed at random. Proxy respondents are systematically different from personal respondents. There is therefore potential for proxy responses to reduce non-response error. However, if proxy responding introduces measurement effects, these could systematically bias comparisons of subgroups with a different propensity to respond by proxy (such as age groups or countries). Further work is recommended to better assess the likely impact of proxy responses on non-response bias.

In addition, ways should be sought to improve compliance with the regulations and guidance regarding the use of proxy interviews. The guidance that proxy reports should be accepted only in exceptional circumstances for income and health and labour measures is for good reason. The scientific evidence suggests that the accuracy and comparability of reports on these topics are likely to be compromised if proxy reporting is accepted. Reducing the extent of proxy reporting on these topics to an absolute minimum should therefore improve the quality and comparability of EU-SILC data.

References

Amato, P. R. and Ochiltree, G. (1987), 'Interviewing children about their families: a note on data quality', *Journal of Marriage and the Family,* Vol. 49, No 3, pp. 669–675.

Bickart, B., Phillips, J. M. and Blair, J. (2006), 'The effects of discussion and question wording on self and proxy reports of behavioral frequencies', *Marketing Letters*, Vol. 17, No 3, pp. 167–180.

Blair, J., Menon, G. and Bickart, B. (1994), 'Measurement effects in self vs. proxy response to survey questions: an information-processing perspective', in Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A. and Sudman, S. (eds), *Measurement Error in Surveys*, Wiley, New York, pp. 145–166.

Boehm, L. M. (1989), 'Reliability of proxy response in the current population survey', in *Proceedings of the American Statistical Association, Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 486–489.

Cobb, C. (2018), 'Answering for someone else: proxy reports in survey research', in Vannette, D. L. and Krosnick, J. A. (eds), *The Palgrave Handbook of Survey Research*, Palgrave Macmillan, London, pp. 87–93.

Cohen, R. S. and Orum, A. M. (1972), 'Parent-child consensus on socioeconomic data obtained from sample surveys', *Public Opinion Quarterly*, Vol. 36, No 1, pp. 95–98.

Dawe, F. and Knight, I. (1997), 'A study of proxy response in the Labour Force Survey', *Labour Force Survey User Guide*, Vol. 1, No 11, Office for National Statistics, London, pp. 50–59.

European Parliament (2003), Regulation (EC) No 1177/2003 of the European Parliament and of the Council of 16 June 2003 concerning community statistics on income and living conditions (EU-SILC), OJ L 165, 3.7.2003, p. 1.

Grootendorst, P. V., Feeny, D. H. and Furlong, W. (1997), 'Does it matter whom and how you ask? Inter- and intra-rater agreement in the Ontario Health Survey', *Journal of Clinical Epidemiology,* Vol. 50, No 2, pp. 127–135.

Highton, B. (2005), 'Self-reported versus proxy voter turnout in the current population survey', *Public Opinion Quarterly,* Vol. 69, No 1, pp. 113–123.

Imara, M. A. (2013), 'Quality of proxy reports on Palestinian Labor Force Survey – case study: Tulkarm governorate', thesis, Birzeit University, Palestine.

King, T., Cook, S. and Hunter Childs, J. (2012), 'Interviewing proxy versus self-reporting respondents to obtain information regarding living conditions', in *Proceedings of the American Statistical Association, Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 5667–5677 (http://www.asasrms.org/Proceedings/y2012/Files/400243_500698.pdf).

Moore, J. C. (1988), 'Self/proxy response status and survey response quality – a review of the literature', *Journal of Official Statistics*, Vol. 4, No 2, pp. 155–172.

Schwarz, N. and Wellens, T. (1997), 'Cognitive dynamics of proxy responding: the diverging perspectives of actors and observers', *Journal of Official Statistics*, Vol. 13, No 2, pp. 159–179.

Thomsen, I. and Villund, O. (2011), 'Using register data to evaluate the effects of proxy interviews in the Norwegian Labour Force Survey', *Journal of Official Statistics*, Vol. 27, No 1, pp. 87–98.

8 Current best practice in minimising non-response error in panel surveys

Peter Lynn ([59])

8.1. Introduction

This chapter aims to summarise what is known currently about best practice in minimising non-response error in panel surveys and to make some related recommendations regarding EU-SILC. The chapter therefore emphasises aspects that are particularly pertinent to rotating panels, to face-to-face and mixed-mode data collection, to between-wave intervals of around 1 year and to cross-national surveys – in other words, to the context of EU-SILC. The chapter draws on the results of projects carried out within the Third Network for the Analysis of EU-SILC as well as a review of the broader research literature.

8.1.1. Non-response error

Non-response error is the difference between the numerical value of a survey estimate and the value that would have been obtained if the survey had achieved a 100 % response rate. This difference is therefore caused by the fact that not all sample units participate in the survey. However, it should be clear that non-response error may be different for different estimates from the same survey, so it is hard to make general statements about the level of non-response error associated with a particular survey. If we consider, for simplicity, simple design-based estimation (so we assume that no weighting, calibration or other statistical adjustment is applied), then the non-response error associated with any given estimate $\hat{Y}$ (a sample statistic y that provides an estimate of a population parameter Y) is the product of two components. The first component is the response rate, namely the proportion of sample elements providing data to contribute to the estimate. The second component is the difference between the responding units and the non-responding units in the sample statistic y. We can see this in the following expression for the non-response error:

$$\left(y_r - y_t\right) = \left(\frac{n_t - n_r}{n_t}\right)\left(y_r - y_{(t-r)}\right)$$

where y_r denotes the (observed) statistic y based on the responding sample of size n_r; y_t denotes the (unobserved) statistic y based on the total sample of size n_t; and $y_{(t-r)}$ denotes the (unobserved) statistic y based on the $\left(n_t - n_r\right)$ non-responding units.

The non-response error can therefore be seen to depend both on the (estimate-specific) response rate and on the similarity of responding and non-responding units in terms of the statistic. It is possible for a low response rate to result in little or no non-response error for a particular estimate. This will happen if the responding and non-responding units are similar. However, it is also possible for a high response rate to result in considerable non-response error. This will happen if the non-respondents are rather distinctive in character, that is, if they are systematically different in terms of the variable(s) that are used to produce statistic y. The risk of non-response error probably increases with increasing response rates, but empirical research

([59]) Peter Lynn is with the Institute for Social and Economic Research at the University of Essex, Colchester, United Kingdom. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Peter Lynn (plynn@essex.ac.uk).

has failed to find a clear relationship between response rates and non-response error (Groves and Peytcheva, 2008).

To reduce non-response error, it would seem that one could aim to reduce either the non-response rate, $\left(\frac{n_t - n_r}{n_t}\right)$, or the difference between respondents and non-responents, $\left(y_r - y_{(t-r)}\right)$. However, it must be noted that these two quantities are not independent. If efforts are made to improve response rate, for example from 65 % to 70 %, it is entirely possible that the characteristics of the non-respondents will change. The 30 % of sample members who remain non-respondents after the response rate has been increased could be more distinctive than the 35 % who were previously non-respondents. Thus, although the first component of non-response error has reduced, the second component may have increased, leaving uncertainty as to whether non-response error overall will have reduced or increased. Recognition of this uncertainty has led survey researchers to focus on improving the representativeness (articles often refer to composition or balance) of the responding sample rather than merely improving response rates. In this way, provided that response rates do not decline (much), we can be sure that we are having a positive impact on reducing non-response error. It is with this focus in mind that Chapters 5 and 6 examine the nature of sample composition in EU-SILC. Specifically, Chapter 5 looks at the effect of initial (wave 1) non-response on sample representativeness, whereas Chapter 6 looks at the effect of subsequent attrition between wave 1 and waves 2, 3 and 4.

8.1.2. Causes of non-response

The causes of non-response can be categorised conceptually into three main types (Lepkowski and Couper, 2002): non-location, non-contact and refusal to cooperate.

Non-location is the failure to locate a sample member. In EU-SILC, this can be due to the sample member's location details (i.e. residential address, phone number, etc.) being incorrect in the sampling frame (applies to wave 1) or due to these details having changed since the previous wave at which they were successfully located (applies to wave 2 and subsequent waves). The prominence of non-location at waves other than the first is therefore a function of the extent of mobility in the study population and the between-wave interval (Couper and Ofstedal, 2009).

Face-to-face interviewing remains the dominant mode for EU-SILC data collection (see Chapter 24). In this situation, making contact with a sample member requires the sample member to answer the door when the interviewer visits. With computer-assisted telephone interviewing, contact requires the sample member to answer the telephone when the interviewer calls. In both cases, the chance of successful contact is a function of the interaction between when the interviewer attempts to make contact and when the respondent is at home. With mail or web questionnaires, contact requires the sample member to receive and pay attention to the email or letter inviting them to take part in the survey.

The decision whether to cooperate with a survey will depend partly on survey-specific factors and partly on external and situational factors. A unique feature of any panel survey is that the decision to participate in waves subsequent to the first will be strongly influenced by the experience of previous participation. Thus, respondents' perceptions of the time taken to participate, the cognitive burden, the sensitivity of the questions, and so on, are likely to affect the chance of continued participation. Panel surveys that are particularly burdensome, that are uninteresting or that induce embarrassment or anxiety are therefore at increased risk of suffering from attrition due to non-cooperation. In addition, some sample members, even if willing in principle, are unable to take part in a survey. This can be for a variety of reasons, including poor health or an inability to communicate with the interviewer or to read and respond to the questionnaire. Interviewer-administered modes are generally more inclusive of people with low literacy skills or a lower cognitive ability than self-completion modes. Language barriers are more likely to be a cause of non-response if a survey is offered in only one language.

8.1.3. Approaches to minimising non-response error

Non-response error can be tackled both in the data collection phase of a survey and in the processing/analysis phase. The two approaches should not be thought of as alternatives; they are complementary and are typically used in combination. In the data collection phase, the focus is often on maximising response rates, based on the assumption that higher response rates are more likely to equate to higher quality data, namely lower non-response bias. As we have seen in Section 8.1.1, this assumption is not warranted. Recognising this, surveys sometimes instead aim to improve sample composition. This can be done by focusing resources or attention on population subgroups that are known or expected to be under-represented in the responding sample (see Section 8.4). Such targeting can be carried out only when subgroup membership is known in advance of data collection, based on information either from the sampling frame or, in the case of a panel survey, from a previous wave (Lynn, 2019).

In the processing/analysis phase, statistical adjustment methods such as forms of weighting or calibration can be used to improve the (weighted) sample composition and thereby reduce non-response error (see Chapter 13). Tackling non-response, whether in the data collection phase or in the processing/analysis phase, will reduce non-response error only to the extent that the adjusted variables (the subgroups targeted for response rate enhancement, or the weighting classes) are correlated with survey estimates. The choice of these variables is therefore particularly important.

8.2. Office procedures to minimise non-response

8.2.1. Sample management database

Any panel survey needs to be able to track its sample members for the duration of the panel and to re-contact them at each wave. To facilitate this, some form of sample management database is designed, maintained and used to manage fieldwork at each wave, to carry out tracking between waves and to send emails or letters to panel members. Initially, the database will contain, for each sample unit, whatever contact information is available from the sampling frame or through linkage to other registers. Thereafter, it should be augmented with any relevant additional information collected at each survey wave. This information may be of two types: updated or additional contact information and outcomes from the survey process, such as the timing and outcome of each interviewer visit/call. Both types of information can help to increase the chances of participation at subsequent waves. Information may also need to be added to the database between waves, such as when the survey office is informed that a panel member has died, wishes to withdraw from the survey or has simply moved home.

Systems must be in place for rapidly accessing and using relevant information from the sample management database when the need arises. This need arises primarily during the data collection phase, for example when an interviewer reports that the household is no longer at the same address or that the supplied phone number no longer works. Alternative phone numbers or addresses (such as the phone number or address of the sample member's parents in the case of a young adult who was living with his or her parents at the time of wave 1) may then be supplied to the interviewer. Alternatively, tracking may be undertaken by office staff, who may write emails or letters to the sample household.

8.2.2. Mailings

It is good practice to write to sample members in advance of each wave of data collection, to forewarn them that they may expect a call or visit from an interviewer, and after each wave, to thank them for their participation (Laurie, Smith and Scott, 1999). When survey waves are at annual intervals, as with EU-SILC, it is also common practice to send a mailing midway between waves. In addition to reminding the recipient about the survey, this is also an opportunity to collect additional information

that may aid the data collection process at the next wave, such as any change (realised or imminent) of postal address, phone number or email address.

The advance letter should introduce the survey and explain what participation will involve, as well as providing means for sample members to raise queries (email address, phone number). It is also used to convey any necessary statutory information, for example concerning data storage, confidentiality and the voluntary nature of the survey. Well-designed advance letters can motivate sample members, allay concerns and help minimise non-response (Lynn, Turner and Smith, 1998; de Leeuw et al., 2007). The phone number provided to respondents should be answered by people who are knowledgeable about the survey and trained in addressing concerns. If possible, the aim should be to persuade concerned sample members to allow an interviewer to call and explain the survey in more detail, emphasising that they will still have the opportunity to decline to take part at that stage if they wish.

Between-wave mailings are designed primarily to maintain the salience of the survey for respondents. Survey findings, or reporting of the survey in the media, are often included. In addition, these mailings are used to obtain updates to contact details, particularly for households that have recently moved or already have a move planned in the near future. The most effective ways to obtain new contact information from panel members include offering multiple modes for respondents to report changed details (reply-paid postcard, email, telephone or web form) and requesting change-of-address details from people who have moved rather than asking all sample members to confirm their address (McGonagle, Couper and Schoeni, 2011; Fumagalli, Laurie and Lynn, 2013; McGonagle, Schoeni and Couper, 2013; Cleary and Balmer, 2015). Between-wave mailings can boost response rates at the subsequent wave if well designed (Fumagalli, Laurie and Lynn, 2013).

All respondent mailings can be sent by post and/or by email if email addresses are available for panel members. Email addresses can be collected at the end of the first interview (see Section 8.3.2) and used for subsequent contact. This enables more, and more frequent, contact to be made.

8.2.3. Survey design and interviewer training

Interviewers play an important role in locating, contacting and persuading sample members to participate in survey interviews (see Chapter 24). The processes of selecting (recruiting), training and motivating survey interviewers are therefore influential (Morton-Williams, 1993; Groves and Couper, 1998). The importance of having skilled and motivated interviewers should not be underestimated.

However, once a sample member has participated in their first (wave 1) interview, the experience of having done so will have a major influence on their willingness to participate again at subsequent waves. The perceived enjoyment, difficulty and burden of taking part are all likely to affect the decision regarding whether to take part again (Kalton et al., 1990; Hill and Willis, 2001; Olsen, 2005). Leaving a good impression of a reasonably enjoyable experience is therefore vital for panel surveys. When the survey is administered by an interviewer, the interviewer plays a key role in creating this impression, but in any mode the design of the survey materials and questionnaire is also influential. Questionnaires that jump between topics without explanation, that do not provide answer categories that reflect the respondent's circumstances or that use complex language or complex sentence construction will not leave a good impression. The concept of questionnaire useability (Willis, 2015) encompasses many of the relevant dimensions of the experience of survey respondents and is something that should be tested before any (substantially changed) questionnaire is fielded (Hansen and Couper, 2004; Presser et al., 2004; Tourangeau, Conrad and Couper, 2013). The cross-national context of EU-SILC also brings additional considerations for questionnaire design, including the need for some degree of consistency between countries in the objectives and methods for questionnaire testing (Fitzgerald and Zavala-Rojas, 2019; Smith, 2019).

8.3. Field procedures to minimise non-response

The mode in which sample members are approached and in which they are asked to partici-

pate will influence both resource requirements and the likely extent and nature of non-response. Face-to-face data collection offers the greatest chance to locate sample members (Couper and Ofstedal, 2009) and also tends to produce the highest co-operation rates (de Leeuw, 2005). However, it is also the most expensive data collection mode. Self-completion data collection (e.g. web or mail) is much less expensive but offers fewer opportunities for locating a sample member who has moved or for persuading a reluctant sample member to continue participating in a panel such as EU-SILC. Increasingly, panel surveys are seeking a balance between the cost and attrition implications of different modes by using mixed-mode designs, with personal home visits taking place only when other cheaper modes have been unsuccessful in securing participation (Lynn, 2013; Jäckle, Lynn and Burton, 2015; see also Chapter 24). If personal home visits are not possible, for budgetary or other reasons, a mix of web and telephone modes can achieve respectable response rates (Burton, Lynn and Benzeval, 2020) but may be less likely to achieve measurement equivalence.

8.3.1. Obtaining the interview

In an interviewer-administered survey, non-contact rates can be reduced by increasing the number of contact attempts and by spreading those attempts over different days of the week and times of day. With a self-completion survey, the same principles apply (more reminder emails or letters, and diversity in the timings of these mailings tends to improve response), although the effects on the non-contact rate are assumed rather than measured.

Securing the cooperation of sample members is a key task of survey interviewers. Training, experience and confidence are key (Jäckle et al., 2013). The provision of respondent incentives, monetary or otherwise, can assist the interviewers in this task (Laurie and Lynn, 2009). Incentives tend to play a larger role in web and mail surveys, when there is no interviewer contact (Singer and Ye, 2013). For panel survey interviews, there is no evidence that modest differences in the length of the previous interview influence the willingness of sample members to participate again (Lynn, 2014a), nor that a change of interviewer will necessarily have a negative impact (Lynn, Kaminska and Goldstein, 2014). As discussed in Section 8.2.3, it is likely that the cognitive and emotional experience of the previous interview has greater influence.

In some countries and contexts, an inability to communicate adequately in the language of the interviewer/questionnaire can be a non-negligible reason for non-response. To combat this, translated questionnaires and multilingual interviewers may be made available, but this is costly, and the cost-effectiveness of a multilingual approach to survey data collection will depend on the national prevalence of minority languages and the sample size.

8.3.2. During and after each interview

In panel surveys, each respondent at each wave is typically asked to provide a range of contact details. These may include home telephone number, mobile telephone number, work telephone number, email address, and address and phone number of at least one 'stable contact' (parent, sibling, friend, etc.) who is likely to know where the sample member is if they move. National household panel surveys generally find that a majority of respondents are willing to provide at least some of this information, although each particular type of information may be provided only by a minority (Laurie, Smith and Scott, 1999). Some contact details are given more readily than others, on average, but respondents vary in the extent to which they are willing to give different types of information. For example, far fewer respondents typically give a work telephone number than a home telephone number, but some give only a work telephone number, so if the question about work telephone numbers were omitted, there would be no telephone number available for such respondents.

At each wave, respondents can be asked if they expect to move in the next 12 months. If they respond 'yes', they should be asked if they know when or where they will move. If they already know where they will move, the new address should be recorded. If they know when but not where, the month of the move should be recorded, as this can be used to trigger a between-wave contact.

As soon as possible after a wave is completed, all relevant information that could help with future tracing should be copied to the sample management database (see Section 8.2.1).

At each wave, respondents should be asked to inform the national statistical institute (NSI) if they change address before the next annual interview has taken place and should be given easy means to do so. Typically, respondents are given a reply-paid postcard to report address changes, a freephone telephone number, a link to a web form and an email address to which they can write. A small incentive may be paid for reporting an address change.

Interviewers often hand over some kind of 'thank you' gift at the end of an interview before leaving the house. It is important that sample members realise how grateful the interviewer (on behalf of the survey team) is for their participation. A separate 'thank you' mailing may also be sent.

8.3.3. Tracking and tracing

It is inevitable that some sample members will be found, in the course of survey fieldwork, to have changed address. Procedures must be in place for attempting to locate such sample members at their new address. Often, the field interviewer will be the initial agent for such procedures. He or she may ask current residents at the address whether they know the new address of the mover. When the mover has split from a household and some of the members of the household are still at the address, it is usually simple to trace the mover. When the whole household has moved, the new residents at the address may know the new address of the out-moving household but may be unwilling to reveal this address to an interviewer. In this situation, they can be asked to mail a letter to the sample household asking them to get in touch with the research team. For this to work, the interviewer must be equipped with letters, postage-paid envelopes and reply cards for this purpose, so that they can prepare the mailing immediately and hand the sealed envelope to the new resident, ready for them to write the address and mail it.

If there is no one at the address who knows the new address of a mover, the interviewer may consider asking immediate neighbours. Again, they should be prepared to implement the mailing procedure described above.

If these in-field procedures do not yield the new address, then other methods will be needed. The next step may be to make use of other contact details known for the sample member (from the sample management database), starting with personal contact details such as a mobile phone number or an email address and then, if that is not successful, using the contact details of other (ex-)household members or other people for whom the respondent gave details previously.

Phone calls and emails to trace a mover can be carried out either by the field interviewer or by central office staff. Having the interviewer do this means that tracing activities can begin immediately, and, if successful, the outcome can be acted on immediately if the respondent is still in the same geographical area and therefore accessible to the interviewer (in many countries a large proportion of movers move within a few kilometres). For example, if the sample member is successfully contacted by phone, it is efficient to immediately make an appointment to conduct the interview, which only the interviewer can do. A disadvantage of having the interviewer conduct tracing activities is that the interviewer needs access to personal contact details. This may raise ethical and security issues, depending on the technology in use and possibly on the nature of the interviewer's employment contract and training. If tracing activities are carried out by office staff, these people will be more experienced in tracing (as any one interviewer is likely to have a very small number of cases that require tracing, if any, at any particular wave), and they may also have access to additional resources such as online tracing services or databases. However, this model requires rapid and structured communication between interviewers and the specialist tracing staff.

It is sometimes possible to collect additional information from other sources that may help with future tracing. One example is when the sampling frame is a population or other register that includes information such as telephone numbers or details of other (non-sampled) household or family members.

8.4. Targeted procedures

8.4.1. Basic ideas of targeting

Many survey design features have been shown to have effects that are heterogeneous across subgroups of sample members. These include the form and value of incentives (VanGeest, Johnson and Welch, 2007; Singer and Ye, 2013), the length of the invitation letter (Kaplowitz et al., 2012) and interviewer calling patterns (Campanelli, Sturgis and Purdon, 1997; Bennett and Steel, 2000).

Survey researchers have begun to exploit this heterogeneity of effects by moving away from standardised survey designs in which every sample member is treated identically to designs in which some features are targeted at different subgroups of the sample (Lynn, 2014b). When done well, this can provide considerable efficiency gains (the same survey outcomes for a lower cost or better outcomes for the same cost).

The idea of targeting is simple: if a particular design feature with an associated cost (e.g. extra calls on Sundays) is effective only for a particular sample subgroup, then it should be applied only to that subgroup; if different versions of a feature (e.g. the wording of a reminder letter) are optimal for different subgroups, then a different version should be applied to each subgroup. Targeting can be applied to more than one design feature on the same survey, possibly using different target groups (e.g. Luiten and Schouten, 2013). Furthermore, it should be noted that targeting can be used to improve either response rates or non-response error, or both (see Section 8.1.1). To minimise the level of non-response, each sample subgroup can be assigned the design features that should maximise participation rates. To minimise attrition bias, costly but effective design features can be restricted to subgroups that would otherwise suffer from a higher level of non-response, thereby improving sample representativeness.

Targeting design features requires knowledge of membership of relevant subgroups and the relative effectiveness of features across subgroups. Panel surveys are in a strong position to meet these requirements, given the wealth of relevant information collected about sample members at previous waves, including substantive measures, paradata and participation behaviour. Surveys that fail to target design features are likely to find it harder to achieve representative samples and good response rates within budget constraints.

8.4.2. Sample subgroups for targeting

Targeted subgroups should either be subgroups for which distinct non-response strategies can be designed or include groups with particularly low response rates for which a relatively costly strategy is likely to be effective. Data from previous annual rounds of EU-SILC can be used to identify groups with low cooperation rates, low contact rates or a high propensity to move home, or that require many contact attempts. The groups can be defined by survey measures from earlier waves or, if possible, sampling frame variables.

Targeted messaging may include, for example, mentioning to parents of young children that the survey is used to improve the well-being of children, while mentioning to elderly people that the survey is used to improve the well-being of elderly people. Other groups for whom survey materials can be targeted can be defined by demographics, economic circumstances, geographical location or other features that may be context specific, such as apparent eligibility for a particular government scheme or benefit.

8.4.3. Field procedures for targeting

A wide variety of survey field procedures have the potential to be targeted at subgroups. These include the messaging within respondent communications, respondent incentives, extra contact attempts, use of alternative data collection modes, extra between-wave contacts, interviewer incentives, field priority, call scheduling and interviewer allocation. Good examples of targeted procedures that can be implemented in panel surveys include:

- sending extra between-wave mailings to sample units predicted to be likely to move home (McGonagle, Couper and Schoeni, 2011; Lynn, 2012);

- ensuring targeted messaging in advance letters based on subgroup membership (Fumagalli, Laurie and Lynn, 2013; Cleary and Balmer, 2015; Lynn, 2016);
- tailoring the timing of contact attempts based on paradata from previous waves (Lagorio, 2016);
- prioritising contact attempts with sample units predicted to be hard to contact (Calderwood et al., 2012) or hard to locate (Walejko and Wagner, 2018);
- allocating the least cooperative sample units to the most experienced/successful interviewers (Luiten and Schouten, 2013);
- offering bonus payments to interviewers for carrying out interviews with the lowest propensity groups (Peytchev et al., 2010; Calderwood, Carpenter and Cleary, 2013);
- offering alternative data collection modes to subgroups predicted to be less likely to participate in the main mode (Luiten and Schouten, 2013; Rosen et al., 2014; Lynn, 2017).

8.5. Conclusions and recommendations

There are many aspects of a survey's design and implementation that can contribute to non-response error. There are correspondingly many survey procedures that can be adapted to help tackle non-response and reduce non-response error. We have seen in Chapters 5 and 6 that a considerable degree of non-response error exists in the EU-SILC data. The wave 1 participating sample is not fully representative in most countries, particularly with respect to age, economic activity status and level of education (Chapter 5). Attrition subsequent to wave 1 tends to further reduce representativeness, and increasingly so across the waves, though the magnitude of sample imbalance is, in most cases, modest (Chapter 6). However, these tendencies are not uniform between countries. For example, in some countries attrition is related to income and in others it is not. These findings suggest that there is certainly scope for considering adaptations to survey procedures that may tend to reduce non-response error in EU-SILC. Optimally, at least some of these adaptations may be country specific, reflecting the country differences in attrition and sample balance.

Some of the general best practice identified in this chapter is probably already implemented in EU-SILC in many countries, whereas other best practices may be less common. Several aspects of field implementation are known to vary considerably between countries (Chapter 24). The focus on non-response avoidance skills during the recruitment and training of interviewers, for example, is believed to vary considerably between NSIs. However, this is an institutional-level issue, not something that can be tackled solely within the NSI's EU-SILC team. The institutional context is a common constraint on many of the best practices identified in this chapter. The design and use of a sample management database, however, is survey specific, as are many of the procedures concerning sample mailings, interviewer communications, and tracking and tracing sample members. These may be more easily amenable to survey-specific adaptation. Allowing proxy responses in extreme cases in order to avoid the non-response that would otherwise occur (Chapter 7) is a practice that should continue, but there should be tighter controls on the extent to which countries comply with the guidance, in order to reduce unnecessary between-country variation in data quality.

There would appear to be a strong case for reviewing the non-response best practices identified in this chapter, with a view to including the most important ones in the EU-SILC guidance documents. In addition, NSIs should be encouraged to review their own non-response outcomes and consider implementing targeted non-response procedures. Such targeted procedures provide an opportunity for NSIs to improve fieldwork efficiency, reduce field costs and/or reduce non-response error.

References

Bennett, D. J. and Steel, D. (2000), 'An evaluation of a large-scale CATI household survey using random digit dialing', *Australian and New Zealand Journal of Statistics*, Vol. 42, No 3, pp. 255–270.

Burton, J., Lynn, P. and Benzeval, M. (2020), 'How Understanding Society: the UK Household Longitudinal Study adapted to the COVID-19 pandemic', *Survey Research Methods*, Vol. 14, No 2, pp. 235–239.

Calderwood, L., Carpenter, H. and Cleary, A. (2013), 'Experiments to improve response rates among likely refusers on longitudinal surveys: does assigning better interviewers and paying interviewer incentives work?', paper presented at the International Workshop on Household Survey Nonresponse, September, London.

Calderwood, L., Cleary, A., Flore, G. and Wiggins, R. D. (2012), 'Using response propensity models to inform fieldwork practice on the fifth wave of the Millennium Cohort Study', paper presented at the International Panel Survey Methods Workshop, 4 July, Melbourne.

Campanelli, P., Sturgis, P. and Purdon, S. (1997), *Can you hear me knocking? An investigation into the impact of interviewers on survey response rates*, Social and Community Planning Research, London.

Cleary, A. and Balmer, N. (2015), 'Fit for purpose? The impact of between-wave engagement strategies on response to a longitudinal survey', *International Journal of Market Research*, Vol. 57, No 4, pp. 533–554.

Couper, M. P. and Ofstedal, M. B. (2009), 'Keeping in contact with mobile sample members', in Lynn, P. (ed.), *Methodology of Longitudinal Surveys*, Wiley, Chichester, pp. 183–203.

de Leeuw, E. D. (2005), 'To mix or not to mix data collection modes in surveys', *Journal of Official Statistics,* Vol. 21, No 2, pp. 233–255.

de Leeuw, E., Callegaro, M., Hox, J., Korendijk, E. and Lensvelt-Mulders, G. (2007), 'The influence of advance letters on response in telephone surveys: a meta-analysis', *Public Opinion Quarterly,* Vol. 71, No 3, pp. 413–443.

Fitzgerald, R. and Zavala-Rojas, D. (2019), 'A model for cross-national questionnaire design and pretesting', in Beatty, P., Collins, C., Kaye, L., Padilla, J. L., Willis, G. and Wilmot, A. (eds), *Advances in Questionnaire Design, Development, Evaluation and Testing*, Wiley, Hoboken, NJ, pp. 493–520.

Fumagalli, L., Laurie, H. and Lynn, P. (2013), 'Experiments with methods to reduce attrition in longitudinal surveys', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 176, No 2, pp. 499–519.

Groves, R. M. and Couper, M. P. (1998), *Nonresponse in Household Interview Surveys*, Wiley, New York.

Groves, R. M. and Peytcheva, E. (2008), 'The impact of nonresponse rates on nonresponse bias: a meta-analysis', *Public Opinion Quarterly,* Vol. 72, No 2, pp. 167–189.

Hansen, S. E. and Couper, M. P. (2004), 'Usability testing to evaluate computer-assisted instruments', in Presser, S., Rothgeb, J., Couper, M. P., Lessler, J., Martin, E. A., Martin, J. and Singer, E. (eds), *Methods for Testing and Evaluating Survey Questionnaires*, Wiley, New York, pp. 337–360.

Hill, D. H. and Willis, R. J. (2001), 'Reducing panel attrition: a search for effective policy instruments', *Journal of Human Resources,* Vol. 36, No 3, pp. 416–438.

Jäckle, A., Lynn, P. and Burton, J. (2015), 'Going online with a face-to-face household panel: effects of a mixed mode design on item and unit nonresponse', *Survey Research Methods*, Vol. 9, No 1, pp. 57–70.

Jäckle, A., Lynn, P., Sinibaldi, J. and Tipping, S. (2013), 'The effect of interviewer experience, attitudes, personality and skills on respondent cooperation with face-to-face surveys', *Survey Research Methods*, Vol. 7, No 1, pp. 1–15.

Kalton, G., Lepkowski, J., Montanari, G. E. and Maligalig, D. (1990), 'Characteristics of second wave nonrespondents in a panel survey', in *Proceedings of the American Statistical Association, Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 462–467.

Kaplowitz, M. D., Lupi, F., Couper, M. P. and Thorp, L. (2012), 'The effect of invitation design on web survey response rates', *Social Science Computer Review*, Vol. 30, No 3, pp. 339–349.

Lagorio, C. (2016), 'Call and response: modelling longitudinal contact and cooperation using wave 1 call records data', *Understanding Society Working Papers*, No 2016-01, University of Essex, Colchester.

Laurie, H. and Lynn, P. (2009), 'The use of respondent incentives on longitudinal surveys', in Lynn, P. (ed.), *Methodology of Longitudinal Surveys*, Wiley, Chichester, pp. 205–233.

Laurie, H., Smith, R. and Scott, L. (1999), 'Strategies for reducing nonresponse in a longitudinal panel study', *Journal of Official Statistics*, Vol. 15, No 2, pp. 269–282.

Lepkowski, J. M. and Couper, M. P. (2002), 'Nonresponse in the second wave of longitudinal household surveys', in Groves, R. M., Dillman, D. A., Eltinge, J. L. and Little, R. J. A. (eds), *Survey Nonresponse*, Wiley, New York, pp. 259-272.

Luiten, A. and Schouten, B. (2013), 'Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 176, No 1, pp. 169–189.

Lynn, P. (2012), 'Failing to locate panel sample members: minimising the risk', paper presented at the International Workshop on Household Survey Nonresponse, September, Ottawa.

Lynn, P. (2013), 'Alternative sequential mixed mode designs: effects on attrition rates, attrition bias and costs', *Journal of Survey Statistics and Methodology*, Vol. 1, No 2, pp. 183–205.

Lynn, P. (2014a), 'Longer interviews may not affect subsequent survey participation propensity', *Public Opinion Quarterly,* Vol. 78, No 2, pp. 500–509.

Lynn, P. (2014b), 'Targeted response inducement strategies on longitudinal surveys', in Engel, U., Jann, B., Lynn, P., Scherpenzeel, A. and Sturgis, P. (eds), *Improving Survey Methods: Lessons from recent research*, Psychology Press, Abingdon, pp. 322–338.

Lynn, P. (2016), 'Targeted appeals for participation in letters to panel survey members', *Public Opinion Quarterly,* Vol. 80, No 3, pp. 771–782.

Lynn, P. (2017), 'Pushing household panel survey participants from CAPI to web', paper presented at the 28th International Workshop on Household Survey Nonresponse, August, Utrecht.

Lynn, P. (2019), 'Tackling panel attrition', in Vannette, D. L. and Krosnick, J. A. (eds), *The Palgrave Handbook of Survey Research*, Palgrave Macmillan, London, pp. 143–153.

Lynn, P., Kaminska, O. and Goldstein, H. (2014), 'Panel attrition: how important is interviewer continuity?', *Journal of Official Statistics,* Vol. 30, No 3, pp. 443–457.

Lynn, P., Turner, R. and Smith, P. (1998), 'Assessing the effects of an advance letter for a personal interview survey', *Journal of the Market Research Society*, Vol. 40, No 3, pp. 265–272.

McGonagle, K. A., Couper, M. P. and Schoeni, R. F. (2011), 'Keeping track of panel members: an experimental test of a between-wave contact strategy', *Journal of Official Statistics*, Vol. 27, No 2, pp. 319–338.

McGonagle, K. A., Schoeni, R. F. and Couper, M. P. (2013), 'The effects of a between-wave incentive experiment on contact update and production outcomes in a panel study', *Journal of Official Statistics*, Vol. 29, No 2, pp. 261–276.

Morton-Williams, J. (1993), *Interviewer Approaches*, Dartmouth Publishing, Aldershot.

Olsen, R. J. (2005), 'The problem of respondent attrition: survey methodology is key', *Monthly Labour Review*, Vol. 128, pp. 63–70.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010), 'Reduction of nonresponse bias in surveys through case prioritization', *Survey Research Methods*, Vol. 4, No 1, pp. 21–29.

Presser, S., Couper, M. P., Lessler, J., Martin, E. A., Martin, J., Rothgeb, J. and Singer, E. (2004), 'Methods for testing and evaluating survey questionnaires', *Public Opinion Quarterly,* Vol. 68, No 1, pp. 109–130.

Rosen, J. A., Murphy, J., Peytchev, A., Holder, T., Dever, J. A., Herget, D. R. and Pratt, D. J. (2014), 'Prioritizing low-propensity sample members in a survey: implications for nonresponse bias', *Survey Practice*, Vol. 7, No 1.

Singer, E. and Ye, C. (2013), 'The use and effects of incentives in surveys', *Annals of the American Academy of Political and Social Science*, Vol. 645, pp. 112–141.

Smith, T. W. (2019), 'Optimizing questionnaire design in cross-national and cross-cultural surveys', in Beatty, P., Collins, C., Kaye, L., Padilla, J. L., Willis, G. and Wilmot, A. (eds), *Advances in Questionnaire De-*

sign, Development, Evaluation and Testing, Wiley, Hoboken, NJ, pp. 471–492 (https://onlinelibrary.wiley.com/doi/10.1002/9781119263685.ch19).

Tourangeau, R., Conrad, F. G. and Couper, M. P. (2013), *The Science of Web Surveys*, Oxford University Press, New York.

VanGeest, J. B., Johnson, T. P. and Welch, V. L. (2007), 'Methodologies for improving response rates in surveys of physicians: a systematic review', *Evaluation and the Health Professions*, Vol. 30, No 4, pp. 303–321.

Walejko, G. and Wagner, J. (2018), 'A study of interviewer compliance in 2013 and 2014 census test adaptive designs', *Journal of Official Statistics*, Vol. 34, No 3, pp. 649–670.

Willis, G. (2015), 'Pretesting of health survey questionnaires: cognitive interviewing, usability testing, and behavior coding', in Johnson, T. (ed.), *Handbook of Health Survey Methods*, Wiley, Hoboken, NJ, pp. 217–242.

Statistical adjustment: weighting and imputation

9 Review of EU-SILC weighting methods

Mārtiņš Liberts ([60])

9.1. Introduction

The aim of this chapter is to assess approaches used for weighting in EU-SILC, drawing on the *Methodological Guidelines and Description of EU-SILC Target Variables* (DocSILC065) published by Eurostat (2018a) and illustrating the merits of different approaches through examples from selected countries. The information on current weighting practices in countries was collected through the Net-SILC3 online consultation (third round), in which national statistical institutes (NSIs) were asked to take part (see Section 9.4 for more details). The aim of the third round of consultation was to collect information about weighting and imputation. The information collected about weighting is used in this chapter. National quality reports were used as the secondary source of information for this study.

9.2. DocSILC065

DocSILC065 presents methodological guidelines to assist EU Member States in the preparation of the EU-SILC operation. Moreover, it describes in detail all target variables to be collected for EU-SILC and also all variables included in the ad hoc modules. The document is prepared for each survey round. All versions of the document are available at the Communication and Information Resource Centre for Administrations, Businesses and Citizens forum ([61]).

The document is organised into three chapters. The first chapter – 'EU-SILC methodological guidelines' – describes the following topics of the survey.

- **Introduction.** This section describes aim, time reference, reference population, legal basis, timely and flexible data delivery, the integrated design, selection of the sample, survey units, sample size, sample implementation, survey duration and time, weighting, tracing rules, imputation, transmission of data and data availability.
- **General definitions.** This section provides definitions and describes income data.
- **General description.** This section describes domains and areas, reference periods, units, modes of collection, flags, income flags (income flags: total household income variables; income flags: gross income variables; income flags: net income variables), and identification numbers and record of persons.

The second chapter – 'Description of EU-SILC target variables' – describes all EU-SILC core variables. Table 9.1 summarises the EU-SILC variables used in this study.

([60]) Mārtiņš Liberts is with the Central Statistical Bureau of Latvia. The author would like to thank Gareth James, Peter Lynn, Lars Lyberg and participants of the Net-SILC3 International Best Practice Workshop 'Unit non-response and weighting' for their very useful suggestions and comments. All errors are the author's responsibility. The author would also like to thank all respondents who took part in the Net-SILC3 online consultation (third round). This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Mārtiņš Liberts (martins.liberts@csb.gov.lv).

([61]) https://circabc.europa.eu/

Table 9.1: Summary of EU-SILC weight variables used in the study

Variable	Description	Formula	Notes
DB080	Household design weight	= 1/PROB	PROB – sampling probability
DB080(N)	Household weight after final non-response adjustment	= DB080/R	R – response propensity
DB090	Household cross-sectional weight	= DB80(N) × g	g – calibration factor (g-weight)
DB095	Household longitudinal weight		Is not described in the section 'Weighting'
RB050	Personal cross-sectional weight	= DB090	The result of the 'integrative' calibration
RB060	Personal base weight		Denoted as $\omega_t^{(RB)}$
RB062	Longitudinal weight (2-year duration)		
RB063	Longitudinal weight (3-year duration)		
RB064	Longitudinal weight (4-year duration)		
PB040	Personal cross-sectional weight	= RB050	
PB050	Personal base weight	= RB060	Denoted as $\omega_t^{(RB)}$
PB060	Personal cross-sectional weight for selected respondent	= PB070 × DB090/DB080 = PB070/R × g	Only when a sample of people is used
PB070	Personal design weight for selected respondent		Only when a sample of people is used
PB070(N)	Personal weight for selected respondent after non-response adjustment	= PB070 × DB080(N)/DB080 = PB070/R	Only when a sample of people is used
PB080	Personal base weight for selected respondent		Only when a sample of people is used; denoted as $\omega_t^{(SB)}$

Source: Eurostat (2018a).

The third chapter – 'Description of ad-hoc modules' – describes variables of EU-SILC modules included in the relevant year (a new edition of DocSILC065 is issued for each year of data collection). Classifications and additional information are available in annexes to DocSILC065.

The section 'Weighting' in Chapter 1 outlines a unified structure for the whole weighting procedure of the standard integrated EU-SILC design, covering the initial sample and its cross-sectional and longitudinal development. The section is structured into five subsections that provide information on the following topics:

- 'Introduction';
- 'Weighting for the first year of each subsample':
  - 'Design weights (Household weights DB080 and "Selected respondent" weights PB070)',
  - 'Adjustments for non-response',
  - 'Adjustment to external sources (calibration): SILC target variables DB090 and PB060',
  - 'Personal weights (SILC target variables RB050 and PB040)';
- 'Computation of base weights (SILC target variables RB060, PB050 and PB080)';
- 'Cross-sectional weights, year 2 onwards;'
- 'Longitudinal weights (SILC variables RB062, RB063 and RB064)'.

9.3. National quality reports

National quality reports are available for the 27 EU Member States, the United Kingdom and three European Free Trade Association (EFTA) countries. Most of the national quality reports are published by Eurostat ([62]). There are two exceptions. The quality report for EU-SILC in Portugal is not publicly available. The European Commission's main authentication service (ECAS) login with corresponding permissions is required to access the quality report for Portugal ([63]). The quality report for EU-SILC in Ireland is published by the Central Statistics Office of Ireland (2020). The reference survey for the most recent quality report differs by country ([64]).

- For most of the countries (Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Germany, Greece, Finland, France, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Norway, Poland, Romania, Slovakia, Slovenia, Sweden, Switzerland and the United Kingdom), the most recent quality report available concerns the 2016 EU-SILC.
- The most recent quality reports for Belgium, Estonia and Spain concern the 2015 EU-SILC.
- The most recent quality report for Iceland concerns the 2014 EU-SILC.
- The most recent quality report for Portugal concerns the 2013 EU-SILC.

All quality reports are in English with one exception. The quality report for France is in French.

Information from the national quality reports is used in this study as a supplementary source to the information obtained from the Net-SILC3 online consultation.

([62]) https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/7af111b3-b700-4321-9902-695082dcb7e1

([63]) https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/d4337091-442b-45d3-bc77-d6e6295bcb06

([64]) The availability of the quality reports was checked on 21 June 2018.

9.4. Net-SILC3 online consultation

NSIs were asked to take part in the third round of the Net-SILC3 online consultation organised by the Net-SILC3 project. Data collection was carried out from 12 July 2017 to 15 September 2017. The aim of the third round of consultation was to collect information about weighting and imputation. Thirty-four countries were invited to take part:

- the 27 EU Member States (Belgium, Bulgaria, Czechia, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Lativa, Lithuania, Luxembourg, Hungary, Malta, the Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland and Sweden) and the United Kingdom,
- four EFTA countries (Iceland, Liechtenstein, Norway and Switzerland),
- two other countries (North Macedonia and Serbia).

We received answers from 22 countries (a total response rate of 65 %). The response rate amongst the first group, the EU Member States and the United Kingdom, was 75 % (responses were received from Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, the Netherlands, Romania, Slovenia, Slovakia, Spain and Sweden). Among EFTA countries, the response rate was 25 % (only Switzerland responded). Neither of the other countries responded, giving a response rate in this group of 0 %. Belgium responded only to the questions related to imputation. Data from 21 countries (excluding Belgium) are analysed in this section.

Most of the countries gave answers related to the 2016 EU-SILC. Three countries gave answers related to the 2015 EU-SILC, and one country gave answers related to the 2017 EU-SILC. Most of the countries use a household, dwelling or address sample. In four countries (the Netherlands, Slovenia, Finland and Sweden) a person sample is used. The overview of sampling designs used by countries is available in annex 3 to the *EU Comparative Quality Report 2016* (Eurostat, 2018b).

9.4.1. Non-response adjustment for the first wave

Response homogeneous groups are the most common method used for estimating response propensities (they are used in 52 % of all cases). Logistic or other regression models are used by four countries (19 %).

The most common variables used for constructing response homogeneous groups are sampling strata (DB050) and different variables related to geographical properties of sampled households (for example Nomenclature of Territorial Units for Statistics (NUTS) 3 region, degree of urbanisation, size of municipality). This is expected, as strata and variables related to geographical location of a dwelling or a household are common variables available for all sampled cases. Household composition was mentioned in one case. Gender and age groups are used for constructing response homogeneous groups in the case of individual samples.

A wider scope of variables is used in modelling. The variables can be grouped as:

- geographical location variables (for example NUTS 2 region, degree of urbanisation, size of municipality, districts),
- household-type variables (social status, nationality),
- household composition (size of household, number of men, number of women, number of children, identifier if all people aged 16 and over are employees, identifier if all people aged 16 and over are old-age benefit receivers),
- age-related variables (age of the youngest person in the household, age of the oldest person in the household),
- income-related variables (total household income, decile of household income).

Obviously, most of those variables are constructed by linking sample units with data from administrative registers.

France uses a combination of both methods. A logistic regression model is built to model response propensities. The model is used to determine the most significant variables explaining response propensity. The sample is then cross-classified by those variables. Some small groups are grouped together to ensure robust results. Finally, the sample is divided into eight subgroups using the most significant variables explaining the response propensity. A homogeneous response mechanism is assumed within those eight subgroups. This method limits the variance of the non-response corrected weights.

In total, 5 of the 22 responding countries do not directly estimate response propensities. Instead, non-response correction is carried out indirectly by calibration of weights.

All those methods are valid according to DocSILC065. DocSILC065 recommends using two methods (Eurostat, 2018a, p. 34): a classical approach using response homogeneous groups and an alternative method whereby response propensities are estimated using a regression-based approach. Calibration of weights is solved as an optimisation problem, but usually the same result can be achieved as if a generalised regression (GREG) estimator had been used for the estimation (Deville and Särndal, 1992, p. 378; Särndal and Lundström, 2006, p. 63). The most popular calibration estimators can be expressed as GREG estimators. For example, the calibration estimator using the chi-squared distance is equal to the linear GREG estimator. The claim holds true for all cases when population or sample information is used in calibration. We can conclude that calibration of weights fits under a regression-based approach.

The situation regarding the units with unknown eligibility status is diverse. In the most popular approach (used in 48 % of cases), units with unknown eligibility status are counted as eligible. The second most popular approach (38 %) is to count those units as non-eligible. In two cases (10 %), this is not relevant, as eligibility status can be determined using the information from the population frame. Eligibility status is imputed only in one case (Estonia).

DocSILC065 requires imputation of the eligibility status for the units with unknown eligibility status (Eurostat, 2018a, p. 34). Treating all those units as eligible or non-eligible can be described as deterministic imputation. The message of DocSILC065 is that 'Every unit has to be assigned uniquely to one category or the other'. This is achieved for all countries. However, the justification for this requirement

is not given. It is not necessary for the calculation of the over-coverage rate or unit non-response rate. Units with unknown eligibility can be treated as a separate set according to the *ESS guidelines for the implementation of the ESS quality and performance indicators (QPI)* (Eurostat, 2014, pp. 7 and 10). It is possible that the requirement is necessary for the organisation of the follow-up of sample persons who are 'Persons who are no longer members of a private household, or who have moved outside the national territory covered in the survey are dropped from the survey' (Eurostat, 2018a, p. 51).

DocSILC065 allows the use of controlled substitution of non-responding units in exceptional cases (Eurostat, 2018a, p. 31). However, in most countries (86 %) substitution is not allowed. Substitution is allowed in three cases only, and original non-respondents who are successfully substituted are counted as respondents in the estimation of response probabilities. This approach is in line with DocSILC065 (Eurostat, 2018a, p. 35).

9.4.2. Base weights and cross-section weights

Base weights are response-adjusted design weights. In other words, base weights are the product of the design weights and response adjustment factor. However, some countries additionally apply calibration of the base weights.

Base weights are calculated for each panel separately. DocSILC065 states the following (Eurostat, 2018a, p. 36):

> *The base weights are the backbone of the computation of both cross-sectional weights and longitudinal weights. They are computed and updated for a single panel and, as such, they will rarely be used for estimating population parameters. The cross-sectional and longitudinal weights are obtained by combining the base weights in an appropriate way, which will be described later.*

Different methods are used to estimate response propensities for the second-, third- and fourth-wave respondents. The most common option (43 %) is to use response homogeneous groups to estimate response propensities for the second-, third- and fourth-wave respondents. Other options are as follows.

- A logistic or other regression model can be used (29 %).
- Response propensities are estimated indirectly by calibration (two cases).
- Response propensities are not estimated (three cases). However, calibration of base or cross-section weights is applied. We can assume response propensities are estimated indirectly by calibration.
- The logistic regression model and response homogeneous groups can be combined (France).

DocSILC065 recommends calculating the base weights at any subsequent wave conditionally on the previous wave (this is the so-called incremental method). DocSILC065 does not explicitly give a recommendation on how response propensities for the second-, third- and fourth-wave respondents should be estimated. Only one example is given, involving a logit model for estimation of response propensities (Eurostat, 2018a, p. 37). Thus, here countries are free to choose the method for response propensity estimation. The recommendation is to improve DocSILC065 regarding this issue by adding several options for non-response correction for the second, third and fourth waves.

Base weights correspond to each panel separately. Four panels are combined to achieve a cross-sectional data set. Base weights should be adjusted to derive cross-sectional weights. Many countries (48 %) use a scale factor of ¼ for the adjustment. Other countries calculate a scale factor proportional to panel sample size (33 %) or proportional to the number of respondents per panel (19 %).

DocSILC065 explicitly recommends using a scale factor of ¼ for the adjustment of the base weights to derive cross-sectional weights (Eurostat, 2018a, p. 40). It would be good to reconsider this recommendation in DocSILC065. The scale factor of ¼ is optimal in the case of an equally allocated set of respondents in each panel, which is not possible in practice because of panel attrition.

Most countries (62 %) do not apply weight trimming. Three countries (14 %) apply trimming of weights only after non-response correction, and

three countries (14 %) apply trimming of weights only after calibration of weights. Two countries (10 %) apply trimming twice – after non-response correction and again after weight calibration.

DocSILC065 does not explicitly recommend weight trimming (Eurostat, 2018a, p. 40). It is up to each country to decide if trimming of weights should be applied.

9.4.3. Longitudinal weights

According to DocSILC065 (Eurostat, 2018a, p. 15):

> *The reference population of EU-SILC is all private households and their current members residing in the territory of the Member States (MS) at the time of data collection. Persons living in collective households and in institutions are generally excluded from the target population.*

This is the definition of the cross-sectional target (reference) population.

The following statement is also made in DocSILC065 (Eurostat, 2018a, p. 23):

> *For all components of EU-SILC (whether survey or register based), the cross-sectional and longitudinal (initial sample) data shall be based on a nationally representative probability sample of the population residing in private households within the country, irrespective of language, nationality or legal residence status.*

This definition is equivalent to the first definition regarding the cross-sectional population.

The definition of the target population is necessary for successful survey planning and organisation, including stages such as sampling, data collection, weighting and quality evaluation. The target population in survey statistics is a limited and fixed set of units. The definition of the target population is necessary to specify the properties of units belonging to the target population covered by the survey.

The target cross-sectional population is quite well defined by DocSILC065. However, this is not the case for the target longitudinal population. The only reference to the definition of the longitudinal population is given in the subsection 'Weighting of re-entries' (Eurostat, 2018a, p. 43): 'Let us assume for simplicity that all quantities above refer to the "longitudinal" population, i.e. to all units at wave 1 that remain in-scope at waves 2 and 3'. From this sentence, the reader can infer that the longitudinal population is defined as an intersection of corresponding cross-sectional populations. In other words:

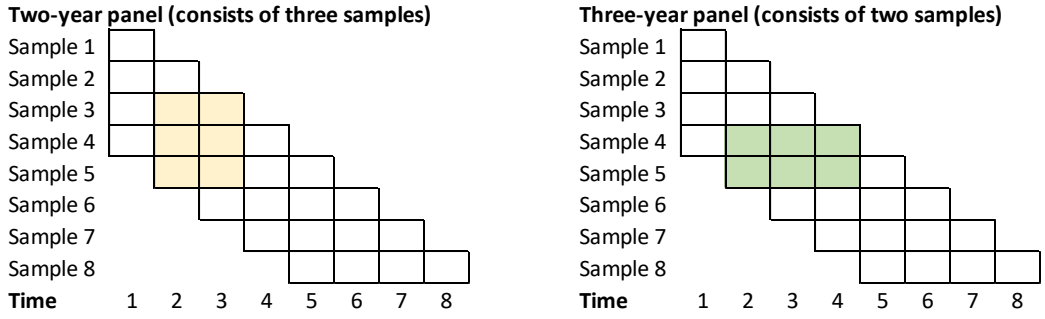
- the target population of the 2-year panel consists of all individuals who were eligible for both years covered by the panel;
- the target population of the 3-year panel consists of all individuals who were eligible for all 3 years covered by the panel;
- the target population of the 4-year panel consists of all individuals who were eligible for all 4 years covered by the panel.

It is important to provide a clear definition of the populations we aim to observe. DocSILC065 should provide an explicit definition of the target EU-SILC longitudinal population.

Different scale factors are used for base weights to derive 2-year panel weights (RB062) and 3-year panel weights (RB063) (Figure 9.1). The most common approach (62 %) is to use the scale factor of 1/3 for a 2-year panel (a 2-year panel consists of data from three samples) and the scale factor of ½ for a 3-year panel (a 3-year panel consists of data from two samples). Other approaches are to calculate scale factors that are proportional to wave sample size (19 % cases) or proportional to number of respondents per wave (14 % cases). Austria uses an approach whereby the sum of RB062 and RB063 is rescaled to the number of people with a continuing presence in the population during the past 2 years and 3 years, respectively (calculated from the population register).

DocSILC065 explicitly recommends using a scale factor of 1/3 to derive weights for a 2-year panel and a scale factor of ½ for a 3-year panel (Eurostat, 2018a, p. 42). It would be good to reconsider this recommendation. Analogous to the argument in Section 9.4.2 regarding cross-sectional weights, the scale factors of 1/3 and ½ are optimal in the case of an equal allocated sample size for each panel, but this is not possible in practice because of wave attrition.

Figure 9.1: 2-year and 3-year panels



NB: The 2-year panel covers 2 years and contains sampled units from three samples. The 3-year panel covers 3 years and contains sampled units from two samples.

Source: Eurostat (2018a) and author's visualisation.

The term 're-entries' in the case of EU-SILC (with four waves) refers to respondents who:

- responded in wave 1, did not respond in wave 2 and responded (re-entered) in wave 3 (response status in wave 4 is not relevant here);
- responded in wave 1 and wave 2, did not respond in wave 3 and responded (re-entered) in wave 4.

Note that in accordance with EU-SILC follow-up rules:

- non-respondents in wave 1 are not followed up in the subsequent waves;
- respondents in wave 1 who do not respond in waves 2 or 3 are not followed up in wave 4.

Conditional response propensity in DocSILC065 is denoted as $P\left(n_j|n_i\right)$, where n_i denotes response in wave i, n_j denotes response in wave j, and $j > i$. Obviously, i can take a value from 1 to 3, and j can take a value from 2 to 4. Most countries (52 %) estimate conditional response propensity $P\left(n_3|n_1\right)$ and apply it for the weighting of re-entries denoted as set C in DocSILC065 (Eurostat, 2018a, p. 43) ([65]). Other approaches are as follows.

- In some countries, there are no re-entries in the data, as no attempt is made to follow up if non-response occurs in any wave.
- In Bulgaria, the base weight of the person from the last year of participation is adjusted using average response propensities from non-participating years for this person.
- In Latvia, re-entries receive the base weight of the re-entered household (no adjustment is made).
- In Austria, re-entries are treated as co-residents when re-entering the population.

DocSILC065 provides only one approximate solution for weighting re-entries (Eurostat, 2018a, p. 44), namely by estimating the conditional response propensities $P\left(n_3|n_1\right)$ and $P\left(n_4|n_2\right)$. Other options could be considered.

9.4.4. Calibration of weights

All countries use calibration of weights. However, strategies for weight calibration differ considerably by country. There are five different scenarios observable from the responses.

- Calibration is applied for base weights (RB060, PB050, PB080), cross-sectional weights (DB090, RB050, PB040, PB060) and longitudinal weights (RB062, RB063, RB064) by six countries. There are some additional comments.

([65]) By mistake, a question in the Net-SILC3 online consultation (third round) about only re-entries corresponding to the pattern 'response at wave 1, non-response at wave 2, response at wave 3' was asked. There is no information collected about re-entries corresponding to the pattern 'response at waves 1 and 2, non-response at wave 3, response at wave 4'.

  - **Italy.** In addition to the calibration for the whole sample (combining all four rotational groups), we apply a calibration for the entering rotational subgroup in the first wave and for the other rotational groups (the author assumes the comment refers to the calibration of the cross-sectional weights).
  - **Finland.** Cross-sectional weights are calibrated separately for each wave.
- Calibration is applied for cross-sectional weights (DB090, RB050, PB040, PB060) and longitudinal weights (RB062, RB063, RB064) by six countries. There are some additional comments.
  - **Austria.** After weight-sharing (of RB060 on the household level) the household weights are calibrated separately for waves 2–4 before combining all waves.
  - **Slovakia.** Base weights are calibrated only in the first wave and then adjusted for attrition.
- Calibration is applied for base weights (RB060, PB050, PB080) and cross-sectional weights (DB090, RB050, PB040, PB060) by four countries. There are some additional comments.
  - **Bulgaria.** Base weight RB060 is calibrated only in the first wave.
  - **Spain.** Household longitudinal weight (DB095) is calibrated.
- Calibration is applied only for **cross-sectional weights** (DB090, RB050, PB040, PB060) by four countries. There are some additional comments.
  - Regarding, cross-sectional weights (DB090), calibration of weights is for each panel separately. Weights are combined proportional to the number of responding households.
- Calibration is applied only **for base weights** (RB060, PB050, PB080) by one country.

We can observe the following.

- Italy and Austria perform extra calibration for waves 2–4 before combining all four waves together.
- Bulgaria and Slovakia calibrate the base weights only for the first wave.
- Spain calibrates the household longitudinal weight (DB095). The calibration of DB095 is not described in DocSILC065.

DocSILC065 recommendations are as follows.

- The household cross-sectional weight (DB090) is computed by calibrating the household weight after final non-response adjustment ($DB080^{(N)}$).
- Personal cross-sectional weight for the selected respondent (PB060) is indirectly calibrated, because $PB060 = PB070 \cdot \frac{DB090}{DB080}$, where DB090 is calibrated.
- Personal cross-sectional weights (RB050 and RB040) are automatically calibrated, as they are set equal to DB090, which is calibrated.

The calibration of longitudinal weights is not mentioned in DocSILC065.

Almost all countries (90 %) use 'integrative' calibration, which is recommended by DocSILC065. The following is an excerpt from DocSILC065 (Eurostat, 2018a, p. 36):

> *The idea [of the 'integrative' calibration] is to use both household and individual external information in a single-shot calibration at household level. The individual variables are aggregated at household level by calculating household totals such as the number of males or females in the household, the number of persons aged of 16 and over. The calibration is then conducted at household level using household variables and the individual variables in their aggregated form. This technique ensures 'consistency' between household and individual estimates by making the household and the individual weights equal.*

The result of the integrative weighting is that the cross-sectional weight at the person level (RB050) will be equal for all individuals from the same household, and it will be equal to the cross-sectional weight of the corresponding household (DB090). It makes the estimation easier, as there is full consistency between household-level and person-level estimates. The recommendation to use integrative calibration is reasonable from this point of view.

One of the reasons for not using integrative calibration is the higher variance for calibration factors (g-weights) and hence higher variance for calibrated weights. Obviously, integrative calibration is an extra constraint for the optimisation problem. Integrative calibration requires the resulting calibrated weights to be equal for all individuals from

a household, which is not the case for a standard calibration at the person level. Whether to use integrative calibration is a trade-off between simplicity of the weighting system and variance of calibrated weights. The variance of calibrated weights with integrative calibration can be reduced by dropping other calibration constraints. For example, some less important population totals could be dropped from the list used in the calibration.

Four different calibration methods are used by countries:

- logit calibration (bounded),
- truncated linear calibration (bounded),
- linear (or regression) calibration,
- raking (or exponential) calibration.

The most common calibration methods are so-called bounded calibration methods, whereby the user can set upper and lower bounds for the calibration factor (known as g-weight).

DocSILC065 does not specify which calibration method to use. The recommendation is to use a bounded calibration method. Thus, both truncated linear calibration and logit calibration are acceptable methods according to DocSILC065.

The most common software for calibration is the SAS macro CALMAR (Institut National de la Statistique et des Etudes Economiques, 2016). It is used in 67 % of all countries. Other solutions mentioned are as follows.

- 'Bascula' used to be a module of the Blaise software (Statistics Netherlands, 2020) up to version 4. R-Bascula is currently being developed in R by Statistics Netherlands. R-Bascula will integrate with Blaise 5 through an R interface.
- SAS macro CLAN was developed by Statistics Sweden (Andersson and Nordberg, 1994).
- SAS macro Estimation of Totals and Order Statistics (ETOS) is an enhancement of CLAN.
- SAS macro Generalised Estimation System (GES) is mentioned (Estevao, Hidiroglou and Särndal, 1995).
- Custom SAS script is another solution.
- R package 'sampling' is mentioned (Tillé and Matei, 2016).
- R Shiny application 'Calif' is another option (Statistical Office of the Slovak Republic, 2020).

DocSILC065 gives only one example regarding the software for calibration, namely the SAS macro CALMAR. This should be revised, because many other tools are available for weight calibration. Giving only one example can lead to the wrong conclusion that SAS macro CALMAR is the only recommended software for weight calibration.

SAS software is required to run CALMAR. The licence to use SAS software is quite expensive. There are freeware options available as alternatives, for example different R packages that perform the same weighting procedures as CALMAR.

9.4.5. Plans for EU-SILC weighting

The online consultation included a question about the plans for EU-SILC weighting in the near future. Some responses were as follows.

- **Austria** planned to expand the number of age categories and maybe add occupational status from registers instead of number of employees/retirees by the 2018 EU-SILC.
- **Bulgaria.** The 2016 EU-SILC in Bulgaria contains six rotational groups. The weighting procedure is analogous to the procedure described in DocSILC065. The country planned to add another calibration variable and try to reduce the standard error of the at-risk-of-poverty or social exclusion (AROPE) indicator.
- **Croatia** planned to test response homogeneity groups.
- **Germany** planned to have a new sample in 2020. The sample size would then increase. That is why it planned to develop a new method for weighting.
- **Greece** was in the process of finding an expert to reassess and make improvements regarding weighting/calibration methods (together with sampling design and variance estimation).
- **Finland.** A grants project concerning this issue was ongoing.
- **France** planned to redesign EU-SILC, which would take effect in the 2020 EU-SILC. Weighting, calibration and imputation would have to be thoroughly revised.

- **Ireland** was reviewing EU-SILC processes and, depending on the outcome of this review, any aspect of current EU-SILC methods could be amended.
- **Italy.** Starting from the 2016 survey, the data collection technique (computer-assisted telephone interviewing / computer-assisted personal interviewing) was planned to be introduced in the non-response correction phase using distinct models to estimate response propensities. Concerning longitudinal weights, the estimation procedure was under revision. In particular, the country was studying how to improve RB062, RB063 and RB064 to better represent $t-1$, $t-2$ and $t-3$ populations, respectively, surviving at time t, using only balanced subsamples. Furthermore, some improvements in RB060 were under study, to allow for fertility and mortality estimations.
- **Romania** planned to apply the trimming of the weights after non-response treatment.
- **Slovakia** planned to put the Calif tool into a new html graphical user interface environment by using the Shiny package (in a few months).
- **Sweden** planned to also include calibrated weights in the longitudinal data sets.
- **Switzerland** planned to use register variables for longitudinal weights (which it had already been doing for transversal weights from 2014 onwards).
- **United Kingdom.** The source of the first wave would change for the 2017 EU-SILC, which would require changes to the weighting. This would affect wave 2 base weights from 2018 onwards.

9.5. Conclusions

In general – but not universally – we can conclude that EU-SILC weighting is performed according to the DocSILC065 recommendations ([66]). There were only some minor incompliances found during the study. In many cases, the reason for those incompliances is because DocSILC065 should be revised and updated. Incompliances and recommendations are summarised in Table 9.2.

([66]) Conclusions are drawn only about the countries that took part in the third round of the Net-SILC3 online consultation.

Table 9.2: The list of incompliances and recommendations

Incompliance	Recommendation to countries	Recommendation to Eurostat
DocSILC065 requires the eligibility status to be determined for all sampled units (p. 34). In other words, every sampled unit must be assigned uniquely to the categories eligible or non-eligible.	None.	The reason for this requirement should be explained.
In several cases (48 %), the weighting of re-entries is not done according to DocSILC065. In some cases, this is because re-entries are not followed or recorded (this topic is outside the scope of this chapter).	Those countries should check if the weighting of re-entries is done in accordance with DocSILC065.	Should advise countries on the weighting of re-entries.
Some countries do not use the integrative calibration approach.	Integrative calibration provides a simpler weighting system with consistent weights at the person and household levels. The recommendation to use integrative calibration is reasonable from this point of view. Obviously, the integrative calibration can increase the variance of calibrated weights if compared with the standard calibration at the person level. If the variance of calibrated weights is a problem, the recommendation is to revise the list of population totals used in the calibration.	None.
Bounded calibration method is recommended by DocSILC065. However, one third of the countries do not use bounded calibration.	Those countries should test if bounded calibration could improve the precision of the main EU-SILC results.	Should check whether bounded calibration should be a recommendation in all cases.
Different methods are used for estimating response propensities for the second-, third- and fourth-wave respondents. However, the logit model is the only method mentioned by DocSILC065.	None.	Should revise and update DocSILC065 regarding this topic.
Different scale factors are used for combing data from different waves. However, DocSILC065 recommends using constant scale factors (¼ for cross-section data, ⅓ for a 2-year panel, ½ for a 3-year panel)	None.	Should revise and update DocSILC065 regarding this topic.
DocSILC065 does not provide an explicit definition of the longitudinal populations EU-SILC aims to observe.	None.	Should update DocSILC065 by adding an explicit definition for longitudinal populations in the scope of EU-SILC.
DocSILC065 does not give any recommendation for calibration of longitudinal weight. However, many countries calibrate longitudinal weights.	None.	Should update DocSILC065 by adding recommendations on how the calibration of longitudinal weights should be done.
The only calibration software mentioned in DocSILC065 is the SAS macro CALMAR.	None.	Should update DocSILC065 by adding other software options that could be used for the weight calibration.

References

Andersson, C. and Nordberg, L. (1994), 'A method for variance estimation of non-linear functions of totals in surveys, theory and software implementation', *Journal of Official Statistics*, Vol. 10, No 4, pp. 395–405.

Central Statistics Office (Ireland) (2020), *Standard report on methods and quality for the 2018 Survey on Income and Living conditions (EU-SILC)* (https://www.cso.ie/en/methods/qualityreports/surveyonincomeandlivingconditions/).

Deville, J. and Särndal, C. (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association*, Vol. 87, No 418, pp. 376–382, doi:10.2307/2290268.

Estevao, V., Hidiroglou, M. A. and Särndal, C.-E. (1995), 'Methodological principles for a generalized estimation system at Statistics Canada', *Journal of Official Statistics*, Vol. 11, No 2, pp. 181–204.

Eurostat (2014), *ESS guidelines for the implementation of the ESS quality and performance indicators (QPI)*, version 1.4 (https://ec.europa.eu/eurostat/web/quality/quality-reporting).

Eurostat (2018a), *Methodological Guidelines and Description of EU-SILC Target Variables – DocSILC065 – 2018 operation (version October 2018)* (https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/e9a5d1ad-f5c7-4b80-bdc9-1ce34ec828eb/details/1.4).

Eurostat (2018b), *EU Comparative Quality Report 2016* (https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/6f7191df-c72d-4537-aee4-81972188497d/details).

Institut National de la Statistique et des Etudes Economiques (2016), 'La macro SAS CALMAR', 6 July (https://www.insee.fr/fr/information/2021902).

Särndal, C. and Lundström, S. (2006), *Estimation in Surveys with Nonresponse*, Wiley, Chichester.

Statistical Office of the Slovak Republic (2020), 'Calibration of weights of statistical surveys – Calif' (https://slovak.statistics.sk/wps/portal/ext/products/software.tools/).

Statistics Netherlands (2020), 'Blaise' (https://www.blaise.com/).

Tillé, Y. and Matei, A. (2016), 'Sampling: survey sampling', R package version 2.8 (https://CRAN.R-project.org/package=sampling).

10 Use of registers in calibration

Mārtiņš Liberts ([67])

10.1. Introduction

The aim of this chapter is to examine whether the addition of registered income data (Jantti, Törmälehto and Marlier, 2013) improves the performance of the weight calibration (Särndal and Lundström, 2005). It expands on Chapter 9, drawing on empirical evaluations of alternative calibration approaches from five different register countries. Two quality and efficiency measures are considered for the evaluation of the calibration process:

- precision gains for the estimates of the main EU-SILC target indicators,
- variability of calibration factors (known as g-weights) and final (calibrated) weights.

The study uses data from Latvia, the Netherlands, Slovenia, Finland and Sweden. The analyses of data from the Netherlands, Finland and Sweden were carried out by the corresponding national statistical institute (NSI); the analyses of data from Latvia and Slovenia were carried out by the author.

10.2. General methodology

The following study was conducted to achieve the research goals. The weighting of recent European Union Statistics on Income and Living Conditions (EU-SILC) data was performed in two settings.

- **Setting 0 (denoted as 'W0').** Calibration of the EU-SILC cross-sectional weight was performed using registered income calibration variables **excluded** from the calibration matrix.
- **Setting 1 (denoted as 'W1').** Calibration of the EU-SILC cross-sectional weight was performed using registered income calibration variables **included** in the calibration matrix.

For the Netherlands, Finland and Sweden, the corresponding NSI chose the calibration variables for W0 and calculated the weight. Weight W1 is the official weight used in the production of statistics. Statistics Slovenia provided anonymised EU-SILC microdata to the author for the duration of the project. The weight calibration to derive weights W0 and W1 for Slovenia was carried out by the author. The weights and results calculated by the author do not match the official weights used by Statistics Slovenia in the production of statistics. Hence, the resultant estimates do not match the official EU-SILC figures published by either Statistics Slovenia or Eurostat.

The two types of weights introduced here are compared for the Netherlands, Slovenia, Finland and Sweden. Data from 2016 are used for the Netherlands and Sweden, whereas 2014–2016 data are

([67]) Mārtiņš Liberts is with the Central Statistical Bureau of Latvia. The author is grateful for the help and support received from Gareth James, Peter Lynn, Tara Junes, Harm Jan Boonstra, Barry Schouten, Jens Malmros, Thomas Helgeson, Rihard Inglic, Rudi Seljak and Eric Marlier. All errors are the author's responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Mārtiņš Liberts (martins.liberts@csp.gov.lv).

used for Slovenia, and 2013–2016 data are used for Finland (variance estimation for Finland has been performed only for the 2016 data). Thus, there are nine country-years for which the two weights are compared in Section 10.4.

The use of registered income data for calibration was being evaluated in Latvia and was not yet part of the EU-SILC production process at the time of the study. Ten different types of calibration weights were compared for Latvia. The case of Latvia is addressed in detail in Section 10.3.

Only cross-sectional weights are analysed in this study. However, similar evaluations can also be carried out for longitudinal weights (if they are calibrated).

The following calibration variables are used in EU-SILC production by the countries taking part in the study. This list of weighting variables therefore corresponds to W1 (the calibration variables used for W0 are a subset of the W1 variables):

- Latvia:
  - age groups (0–5, 6–11, 12–15, 16–17, 18–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75 and over) by sex classes,
  - Nomenclature of Territorial Units for Statistics (NUTS) 3 regions,
  - use of income variables for calibration only since the 2020 EU-SILC (Section 10.3 describes a study of different income variables to be used for weight calibration in the case of EU-SILC in Latvia);
- the Netherlands:
  - personal level:
    - age by sex classes (15 classes for each sex),
    - NUTS 2 regions (12 classes) by age (two classes),
    - ethnicity (three classes),
    - household size (five classes),
    - NUTS 1 regions (four classes) by low-income category (three classes),
    - degree of urbanisation (five classes) by at-risk-of-poverty rate (two classes; at-risk-of-poverty threshold is computed by applying EU-SILC methodology on the registered income variable, which is available for the whole population),
    - NUTS 2 regions (12 classes) by at-risk-of-poverty rate (two classes; at-risk-of-poverty threshold is computed by applying EU-SILC methodology on the registered income variable, which is available for the whole population),
    - NUTS 2 regions (12 classes) by activity status (five classes);
  - household level:
    - NUTS 2 regions (12 classes) by household income (deciles),
    - NUTS 1 regions (four classes) by tenure status (three classes),
    - tenure status (three classes);
- Slovenia:
  - age groups (0–15, 16–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–99),
  - sex,
  - wages,
  - pensions,
  - income from receiving rents,
  - unemployment benefits,
  - scholarships,
  - family-/children-related allowances;
- Finland:
  - NUTS 3 regions, capital region and Helsinki,
  - size of household-dwelling unit (1, 2, 3, 4, 5, 6, 7, 8 and over),
  - degree of urbanisation (designed by Statistics Finland),
  - age by sex – ages 0–4, 5–9, 10–14 … 80–84, 85 and over for each sex (36 classes),
  - low-income people,
  - income 1 – cash or near-cash employee income,
  - income 2 – income 1 > 0 (binary variable indicating positive cash or near-cash employee income),
  - income 3 – pensions,
  - income 4 – unemployment benefits 1,
  - income 5 – unemployment benefits 2,

- income 6 – income 4 > 0 (binary variable indicating positive unemployment benefits 1),
- income 7 – income from self-employment,
- income 8 – capital income 1,
- income 9 – income from agriculture,
- income 10 – income from property and forestry,
- income 11 – other capital income,
- income 12 – income from forestry 2,
- income 13 – capital gains,
- income 14 – pensions > 0 (binary variable indicating positive pension income),
- income 15 – mortgage interests;

- Sweden:
  - age by sex – ages 0–5, 6–10, 11–15, 16–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, 85 and over for each sex (22 classes),
  - civil status – unmarried, married / registered partner, divorced / separated partner, widow / widower / survivor partner (four classes),
  - education level – not registered, basic, secondary, tertiary (four classes),
  - Swedish born / foreign born (two classes),
  - NUTS 2 regions (eight classes),
  - income (deciles) – deciles for the income year (10 classes),
  - income (amount),
  - financial aid – receives financial aid, does not receive financial aid (two classes),
  - housing allowance – receives housing allowance, does not receive housing allowance (two classes),
  - sick pay – receives sick pay, does not receive sick pay (two classes).

Point estimation and standard error estimation for the 23 main EU-SILC indicators (Eurostat, 2018, Annex 4) were performed using both weights (W0 and W1). All those indicators refer to the population of individuals. Table 10.1 shows the main EU-SILC indicators. The estimation of standard errors and related precision measures for Latvia and Slovenia was carried out using the R package *vardpoor* (Breidaks, Liberts and Ivanova, 2020).

Table 10.1: Main EU-SILC indicators

Indicator code	Indicator name	Domain
AROPE	At risk of poverty or social exclusion	Total
AROPE	At risk of poverty or social exclusion	Age 0–17
AROPE	At risk of poverty or social exclusion	Age 18–64
AROPE	At risk of poverty or social exclusion	Age 65 and over
AROPE	At risk of poverty or social exclusion	Females
AROPE	At risk of poverty or social exclusion	Males
ARPT60	At-risk-of-poverty rate	Total
ARPT60	At-risk-of-poverty rate	Age 0–18
ARPT60	At-risk-of-poverty rate	Age 18–64
ARPT60	At-risk-of-poverty rate	Age 65 and over
ARPT60	At-risk-of-poverty rate	Females
ARPT60	At-risk-of-poverty rate	Males
LWI	Low work intensity	Total
LWI	Low work intensity	Age 0–18
LWI	Low work intensity	Age 18–60
LWI	Low work intensity	Females
LWI	Low work intensity	Males
Sev_Dep	Severe material deprivation rate	Total
Sev_Dep	Severe material deprivation rate	Age 0–18
Sev_Dep	Severe material deprivation rate	Age 18–64
Sev_Dep	Severe material deprivation rate	Age 65 and over
Sev_Dep	Severe material deprivation rate	Females
Sev_Dep	Severe material deprivation rate	Males

Source: Annex 4 in Eurostat (2018).

10.3. The case of Latvia

For Latvia, register income data are available. Therefore, it was possible to experiment with several approaches regarding how to construct calibration variables.

10.3.1. Data availability

Three data sets are used for calibration purposes.

Sample. A dwelling sample is drawn in the case of Latvia. The frame of private dwellings is created using the data from the statistical dwelling register (SDR). The SDR is maintained by the Central Statistical Bureau of Latvia (CSB). The main data sources used for SDR maintenance are the population register and the address register. Private dwellings with at least one declared permanent resident are included in the sampling frame for EU-SILC. It is possible to compile a list of declared permanent residents for all dwellings from the dwelling population frame. The personal identification codes are available for all declared permanent residents. Some other auxiliary information is also available from the SDR, for example geographical location of a dwelling (territory and geographical coordinates), sex, date of birth, citizenship, nationality, marital status, children, parents, and so on. The sample also contains sampling probabilities and hence the design weights.

Survey data. The survey data are used to classify all sampled cases into three groups, namely respondents (R), non-respondents (N) and over-coverage cases (O). The following classification rules are used:

- R – (DB120 = = 11 and DB130 = = 11),
- N – (DB120 = = 11 and DB130 < > 11) or DB120 = = 21 or DB120 = = 22,
- O – (DB120 = = 23).

Administrative register data. This study uses data from the State Revenue Service (SRS). There is an agreement between the CSB and the SRS about regular data delivery of register microdata for statistical purposes. A wide range of data is received from the SRS. However, only one data set, namely annual income data for individuals, is used here. The data cover all registered income that natural persons receive during a year; in other words, the personal identification code, type of income and income value are available.

All three data sources contain personal identification codes, which allow a direct link to all data sources.

10.3.2. Calibration strategy

Two data objects are necessary for a calibration function:

- a matrix of calibration variables whereby auxiliary information is aggregated to the level of households, as 'integrative' calibration is used in EU-SILC;
- a vector of calibration totals.

The dimensions of both data objects should be consistent. The length of the vector should be equal to the number of columns in the calibration matrix. It is extremely important that both data objects are created from the same data source. If this is not the case, survey estimates may be biased. However, in practice some exceptions are allowed. One example is the calibration to population counts by sex, age groups and regions. Usually, in this case, the calibration matrix is derived from survey data and totals are derived from the official population statistics. The justification for such an approach is the following.

- Official population statistics are usually treated as reliable statistics of high quality (and with high precision). We assume that, even though data for the calibration matrix and the totals are taken from different sources, there is a good level of comparability between the data sources.
- A secondary purpose of calibration is to achieve comparability and consistency between different social surveys. We want to observe the same population size in EU-SILC, the Labour Force Survey and other social surveys. Even if calibration to population statistics can introduce some bias, the precision loss here is compensated by an increase in comparability and consistency across surveys.

The calibration strategy defines how a calibration matrix and a vector of totals are created. The following steps were taken in this study.

- Register income data were aggregated to the level of individuals. This creates a total registered income for each person in the SRS data.
- The calibration totals cannot be directly computed from the SRS data. This is because the SRS data contain many records that are

out of scope for the EU-SILC target population, for example income for individuals living in institutional dwellings and income for non-residents. Filtering of the SRS data should be applied.

- One possible way of filtering is to use the population frame of dwellings as a filter for the SRS data. A list of individuals declared in the frame dwellings is compiled. Then, the SRS data are linked only to those individuals (the rest of the SRS data are discarded). Income can now be aggregated to the dwelling level. We have computed the total registered income for each dwelling in a population frame. The calibration totals can now be computed from the population frame of dwellings.
- The registered income compiled in the previous step can be linked to each sampled dwelling.
- N cases are excluded from the calibration, because we have not received survey data for those cases. We will try to compensate for these missing data by applying weights.
- O cases are treated as pseudo-respondents in the calibration. We may not exclude O cases from calibration. We can distinguish O cases only in the sample; we cannot distinguish O cases in the population frame, and hence we cannot remove the O part from the calibration totals.
- The calibration matrix is created from the R and O cases. The design weights of R and O cases are calibrated to totals computed from the population frame.
- We treat R and O cases as separate population (frame) domains, where R cases correspond to the EU-SILC target population and O cases are out of scope.
- O cases can be discarded when point estimates are derived, as only R cases are necessary to derive estimates of EU-SILC indicators. However, R and O cases should be used when the variance estimates are derived, as O cases were used in the calibration step. R cases should be treated as a domain of the frame population consisting of eligible population units in variance estimation.

The data from wave 1 of the 2017 EU-SILC were used for the study. The following weighted (with design weights) distribution of the sample was observed: R 51 %, N 40 % and O 9 %.

The simple estimate under an assumption that the over-coverage rate is the same among respondents as non-respondents is that 15 % of the population frame is over-coverage ([68]).

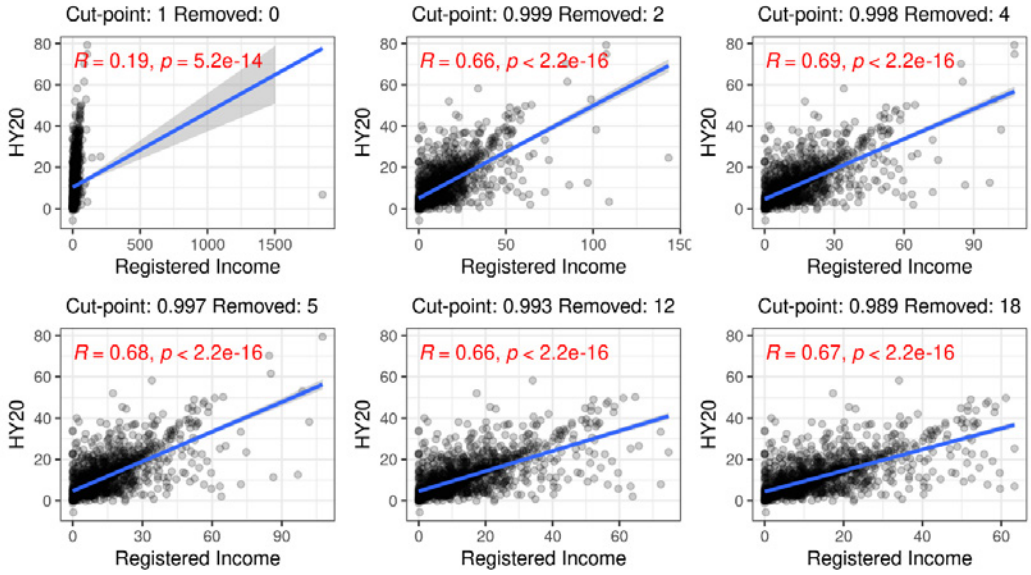
10.3.3. Calibration variables

The registered personal income data can be aggregated to a dwelling as total dwelling income or as equivalised total dwelling income using the so-called Organisation for Economic Co-operation and Development-modified equivalence scale (1.0, 0.5, 0.3) (Haagenars et al., 1994). In the case of respondents, it is possible to compare the registered dwelling income with household income observed in EU-SILC ([69]). The total registered income is compared with the total disposable household income (HY020). The equalised registered income is compared with the equivalised disposable income after social transfers (EQ_INC20). We can observe that the correlation is affected by outliers. The change in the correlation coefficient can be illustrated by systematically removing the observations with the highest registered income values (Figure 10.1).

The correlation between registered income and observed income is not very high, but it seems to be greater than 0.6, which may generate a potential precision gain when registered income data are used for the calibration of weights. Ten sets of calibration variables were created for study purposes (Table 10.2).

([68]) Derived as $9 \div (51 + 9) = 9 \div 60 = 15$ %.
([69]) There are some cases in which more than one household has been observed in one dwelling by EU-SILC. For these cases, the registered dwelling income is split equally between each household.

Figure 10.1: Correlation of registered household income and observed disposable household income (HY020)



NB: The estimate of the correlation coefficient changes as the observations with the highest registered income values are systematically removed.

Source: Author's calculations based on SRS data with reference year 2016: 2017 EU-SILC data for Latvia (wave 1).

Table 10.2: Calibration variables

Name	Variables	Number of variables
Demog	Demographic data (sex by age groups, region – NUTS 3, urbanisation) + constant + number of declared individuals in a dwelling. Demographic data are defined for R cases only. All demographic variables are zero for O cases. Constant and number of declared individuals are defined for R and O cases.	42
Inc	Demog + linearised total registered income	43
Inc20	Demog + registered income ventiles (*)	61
Inc10	Demog + registered income deciles	51
Inc05	Demog + registered income quintiles	46
EqInc	Demog + linearised equalised registered income	43
EqInc20	Demog + registered equalised income ventiles (*)	61
EqInc10	Demog + registered equalised income deciles	51
EqInc05	Demog + registered equalised income quintiles	46
ARPT	Demog + registered poverty thresholds (20 %, 40 % … 180 % from median equalised income)	51

(*) There are 19 ventiles that split the data into 20 equally sized parts.

Source: Author's calculations based on SRS data with reference year 2016: 2017 EU-SILC data (wave 1).

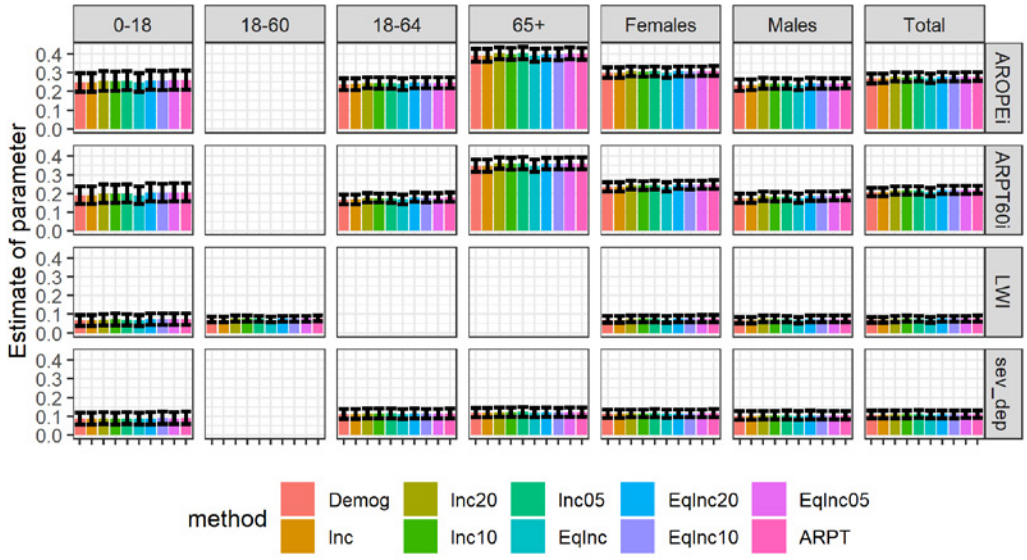
10.3.4. Results

The main EU-SILC estimates with corresponding confidence intervals have been estimated for each set of calibration variables (Figure 10.2). The estimates of standard errors sorted by the value of the standard error estimates in descending order are shown in Figure 10.3.

Several conclusions can be made.

- It is hard to tell which method performs best.
- Income variables do not improve the precision for the domain 'Age 0–18'. There is a slight improvement only for the 'Sev_Dep' indicator.
- The calibration set 'ARPT' shows the best performance for the indicator 'AROPE'.
- The calibration set 'EqInc05' shows the best performance for the indicator 'ARPT60'.
- Use of registered income variables for calibration has not improved the precision of the 'LWI' estimates.
- The calibration set 'EqInc10' shows the best performance for the indicator 'Sev_Dep'.

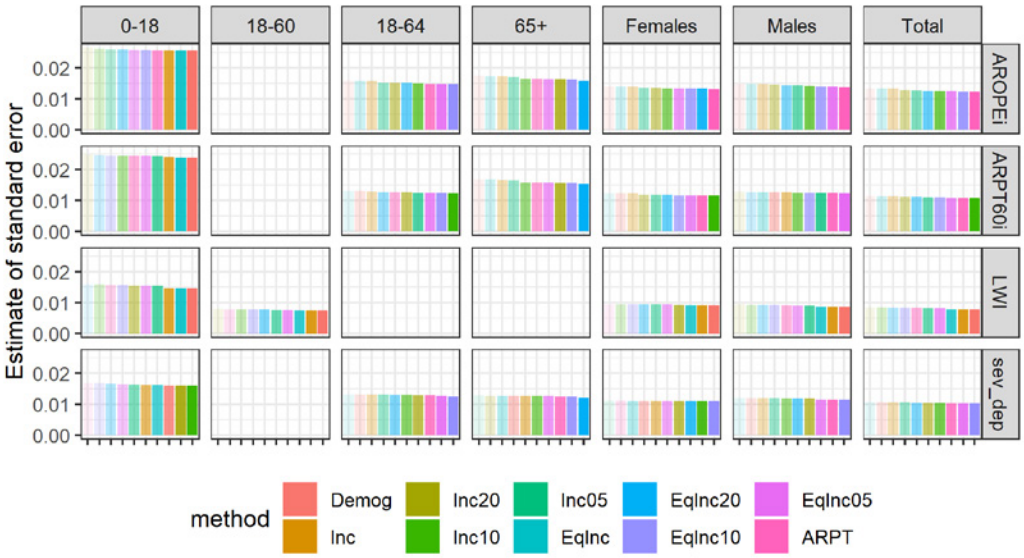
Figure 10.2: Estimates of the main EU-SILC indicators



NB: The indicators, as defined by Eurostat, are listed in Table 10.1. The figure presents estimates of the main EU-SILC indicators with confidence intervals using 10 different sets of calibration variables.

Source: Author's calculations based on SRS data with reference year 2016: 2017 EU-SILC data (wave 1).

Figure 10.3: Estimates of standard errors for the main EU-SILC indicators



NB: The figure presents estimates of the standard errors for the main EU-SILC indicators using 10 different sets of calibration variables. The bars are sorted by the estimated standard error in descending order. A smaller value (on the right-hand side) shows a higher precision of the corresponding estimate.

Source: Author's calculations based on SRS data with reference year 2016: 2017 EU-SILC data (wave 1).

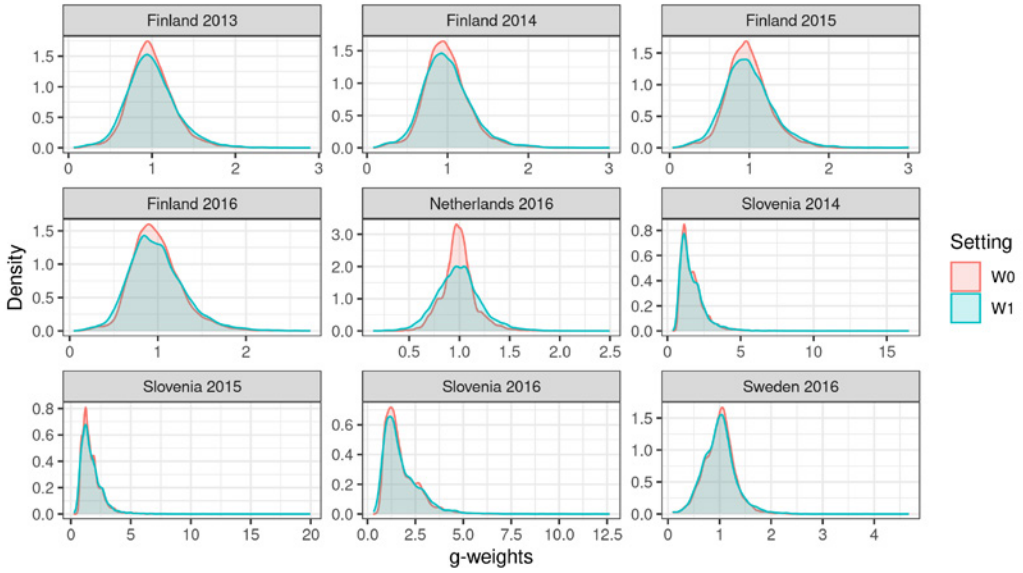
10.4. Results for the other countries

10.4.1. Weights

As expected, the dispersion of calibration factors (known as g-weights) and final (calibrated) weights increases when income variables are added to the calibration matrix. We can observe that both the range and the standard deviation of the calibration factors (g-weights) increase (Figure 10.4 and Table 10.3). The standard deviation of the calibrated weights also increases, but there are cases when the range of calibrated weights decreases (Figure 10.5 and Table 10.4).

At the same time, we can observe that the increase in range and standard deviation for the g-weights and final weights is small. This is an indication that calibration has not resulted in extreme values for g-weights or calibrated weights.

Figure 10.4: Density plots for calibration factors (g-weights)



NB: We observe lower peak values for the density of calibration factors (g-weights) using income data (green shading). A lower peak value is an indication of a higher level of dispersion of values.

Source: Author's calculations based on data from Statistics Netherlands, Statistics Finland, Statistics Sweden and Statistics Slovenia.

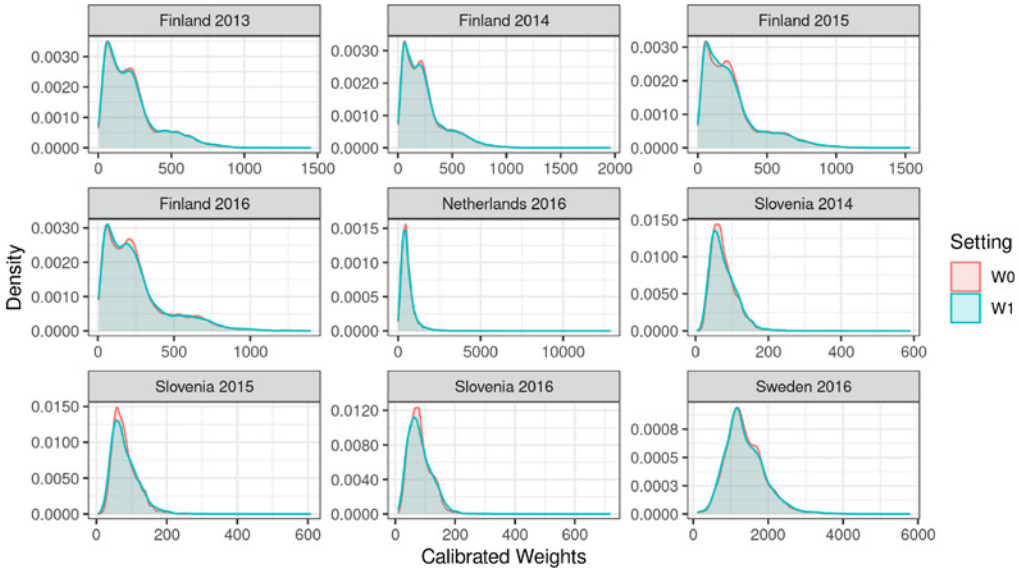
Table 10.3: Range and standard deviation of calibration factors

Country	Year	Setting	Mean	Minimum	Maximum	Range	Standard deviation
Netherlands	2016	W0	0.992	0.242	1.924	1.683	0.164
Netherlands	2016	W1	0.992	0.148	2.489	2.342	0.224
Slovenia	2014	W0	1.697	0.676	16.134	15.458	0.870
Slovenia	2014	W1	1.704	0.392	16.470	16.078	0.930
Slovenia	2015	W0	1.698	0.554	19.908	19.354	0.865
Slovenia	2015	W1	1.706	0.306	19.774	19.468	0.917
Slovenia	2016	W0	1.770	0.479	11.887	11.408	0.951
Slovenia	2016	W1	1.778	0.333	12.599	12.266	1.011
Finland	2013	W0	0.995	0.105	2.482	2.377	0.269
Finland	2013	W1	0.996	0.072	2.888	2.816	0.300
Finland	2014	W0	0.989	0.109	2.656	2.547	0.278
Finland	2014	W1	0.992	0.085	3.000	2.914	0.310
Finland	2015	W0	0.984	0.058	2.434	2.376	0.271
Finland	2015	W1	0.990	0.043	3.000	2.957	0.322
Finland	2016	W0	0.985	0.055	2.719	2.664	0.284
Finland	2016	W1	0.988	0.055	2.721	2.665	0.317
Sweden	2016	W0	0.997	0.100	3.832	3.732	0.298
Sweden	2016	W1	0.997	0.100	4.661	4.561	0.322

NB: The range and standard deviation increase when income variables are added to the calibration matrix. The highest values of the range and standard deviation for each country-year combination are indicated in red.

Source: Author's calculations based on data from Statistics Netherlands, Statistics Finland, Statistics Sweden and Statistics Slovenia.

Figure 10.5: Density plots for calibrated weights



NB: Some small changes are observable for densities for calibrated weights.

Source: Author's calculations based on data from Statistics Netherlands, Statistics Finland, Statistics Sweden and Statistics Slovenia.

Table 10.4: Range and standard deviation of calibrated weights

Country	Year	Setting	Mean	Minimum	Maximum	Range	Standard deviation
Netherlands	2016	W0	565.8	1.5	12 829.3	12827.8	453.9
Netherlands	2016	W1	565.8	1.8	12 370.1	12368.3	456.4
Slovenia	2014	W0	76.4	13.8	401.8	388.0	33.8
Slovenia	2014	W1	77.0	6.3	587.5	581.1	37.0
Slovenia	2015	W0	80.6	16.1	565.9	549.7	35.5
Slovenia	2015	W1	81.4	6.3	606.3	600.0	39.7
Slovenia	2016	W0	82.4	10.8	716.9	706.0	38.5
Slovenia	2016	W1	83.0	9.8	694.5	684.8	42.1
Finland	2013	W0	228.2	5.0	1 449.1	1 444.1	178.3
Finland	2013	W1	228.2	3.0	1 439.0	1 436.0	180.6
Finland	2014	W0	237.8	2.9	1 952.2	1 949.3	188.5
Finland	2014	W1	237.8	3.0	1 897.8	1 894.9	190.8
Finland	2015	W0	246.2	1.0	1 326.1	1 325.1	199.8
Finland	2015	W1	246.2	0.7	1 528.4	1 527.6	203.2
Finland	2016	W0	250.0	5.5	1 392.7	1 387.1	204.8
Finland	2016	W1	250.0	5.5	1 368.4	1 362.9	207.4
Sweden	2016	W0	1387.7	139.7	4 743.5	4 603.8	519.3
Sweden	2016	W1	1387.7	123.8	5 769.9	5 646.2	545.3

NB: The standard deviation of calibrated weights increases when income variables are added to the calibration matrix (W1 rather than W0). The range of calibrated weights increases in some cases and decreases in other cases. The highest values of the range and standard deviation for each country–year combination are indicated in red.

Source: Author's calculations based on data from Statistics Netherlands, Statistics Finland, Statistics Sweden and Statistics Slovenia.

10.4.2. The precision of estimates

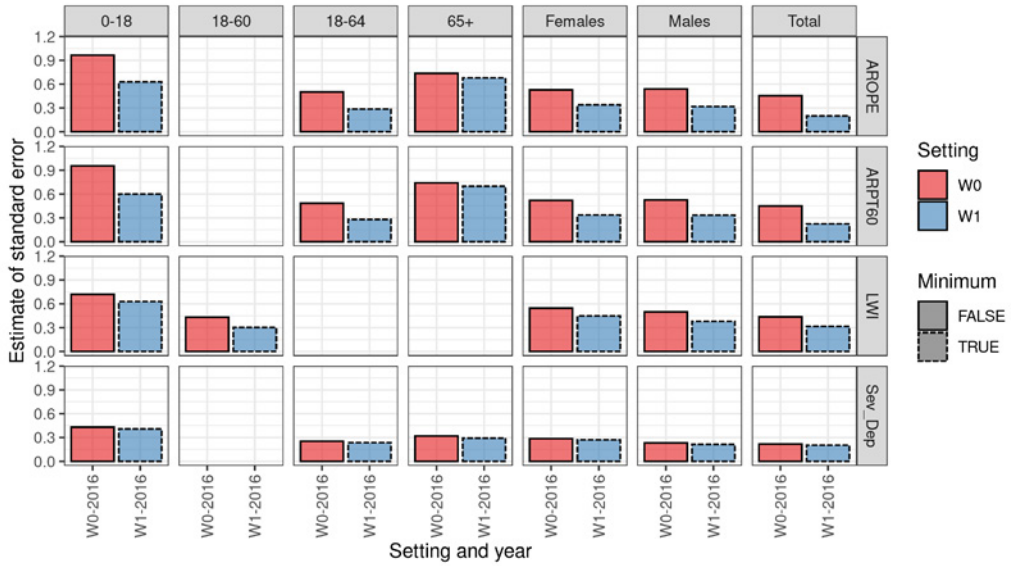
The effect of including income variables in the calibration process on the precision of the main EU-SILC estimates varies. In the case of the Netherlands, the standard errors are lower for all estimates under the W1 scenario (the decrease in the standard error for Sev_Dep is marginal). This is an indication that income variables have been very efficient regarding the improvement of precision (Figure 10.6).

In the case of Finland, most of the estimates have gained in terms of precision. The exceptions are three estimates of Sev_Dep for the domains 18–64 years, 65 years and over and males. However, the differences in the estimated standard deviations for Sev_Dep are marginal (Figure 10.7).

In the case of Sweden, calibration using income variables works quite well for the estimates of at risk of poverty or social exclusion and at-risk-of-poverty rate estimates (the exception is the domain 0–17 years). The gain in precision for the estimate of LWI is present only in one domain (18–60 years). Here, calibration using income variables has not improved the precision of the estimates of Sev_Dep (Figure 10.8).

In Slovenia, calibration using income variables improves the precision in most cases. However, the estimates of LWI are an exception: here, there is no gain in precision in almost all domains. Another exception is domain 0–18 years, where there is only a slight precision improvement for the AROPE estimate. However, it is noticeable that, for all other estimates, the gain in precision is not stable over the years observed (Figure 10.9).

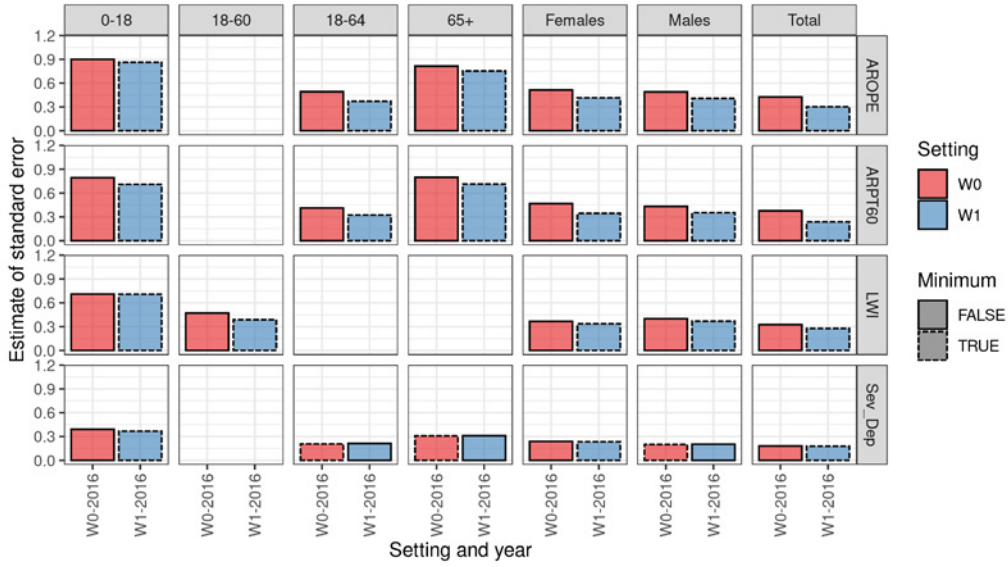
Figure 10.6: Change in precision using income variables for the Netherlands (2016)



NB: The precision improves for all estimates of the main EU-SILC indicators in the case of the Netherlands. The bars with dashed lines refer to the smallest values of the standard errors.

Source: Statistics Netherlands.

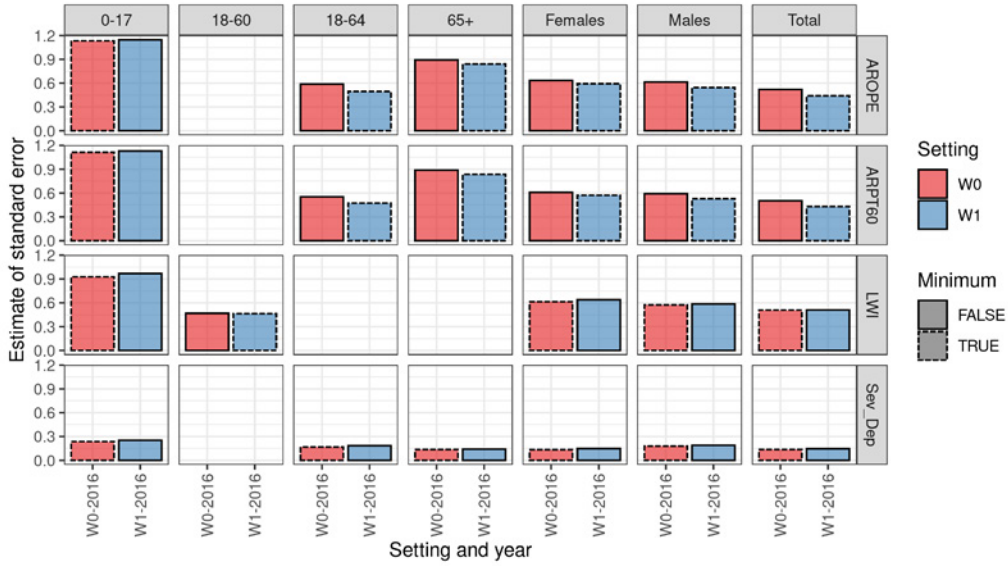
Figure 10.7: Change in precision using income variables for Finland (2016)



NB: Precision improves for almost all estimates (there are three exceptions) in the case of Finland. The bars with dashed lines refer to the smallest values of the standard errors.

Source: Statistics Finland.

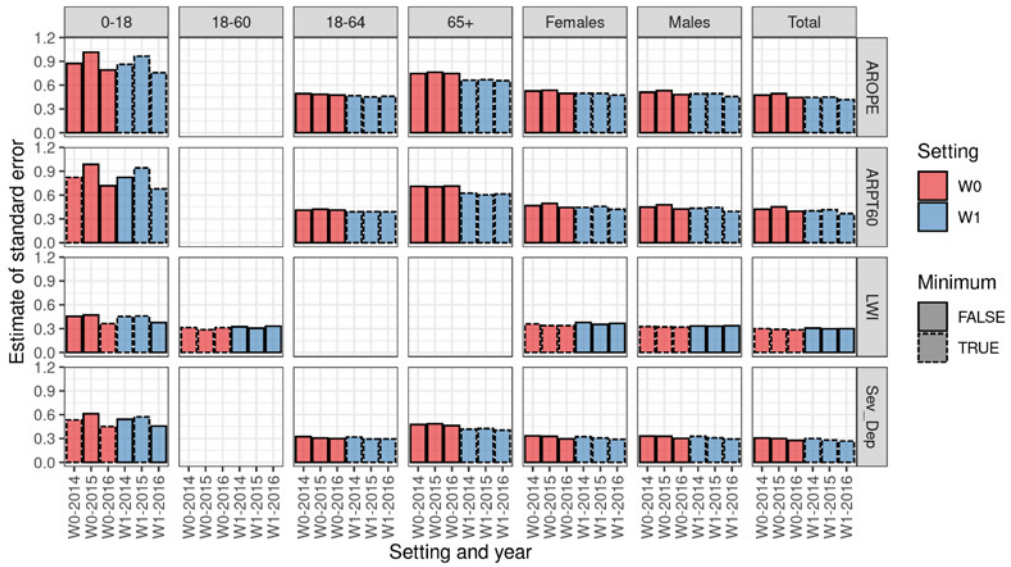
Figure 10.8: Change in precision using income variables for Sweden (2016)



NB: Precision improves for 11 estimates in the case of Sweden. There is no improvement in precision for the other 12 estimates. The bars with dashed lines refer to the smallest values of the standard errors.

Source: Statistics Sweden.

Figure 10.9: Change in precision using income variables for Slovenia (2014–2016)



NB: Precision improves in most cases for Slovenia. The exceptions are the estimates of LWI and the domain 0–18 years. The bars with dashed lines refer to the smallest values of the standard errors.

Source: Author's calculations based on anonymised EU-SILC microdata.

10.5. Conclusions

The gain in precision varies by country. For example, in the case of the Netherlands the precision increases for all 23 estimates when income variables are used for weight calibration. However, we do not see the same level of precision gain for other countries. In addition, the gain in precision varies by indicator. Usually, adding income variables to the calibration improves the precision of estimates of AROPE and ARPT60. The gain in precision for the estimates of LWI is usually smaller. There is no improvement or only marginal improvement for the estimates of Sev_Dep.

We can conclude that the inclusion of income variables does not guarantee the improvement of the survey estimates' precision. Precision improvement depends on several factors, for example the correlation of the study indicators with the register income variables. The precision gain will be higher for highly correlated indicators. The precision gain also depends on the correspondence between the register income data and the survey income data, which depends on record linkage quality, measurement errors in both sources and construction of the calibration variables.

There is no evidence that the inclusion of the income variables in calibration may be harmful: there are no cases observed in the study in which inclusion of income variables decreases the precision of the estimates. This is an expected property of the calibration estimator for estimates covering domains with a sufficient number of respondents observed.

Obviously, using income data in calibration has potential; however, a precision gain is not guaranteed. The general guideline is to test calibration using income variables if registered income data are available for linking with survey data. The precision gain must be evaluated. If significant precision gain is observed, the inclusion of income variables in the weight calibration should be strongly considered. If the gain is not significant, the less complex calibration setting without income variables should be chosen.

References

Breidaks, J., Liberts, M. and Ivanova, S. (2020), 'vardpoor – variance estimation for sample surveys by the ultimate cluster method', R package version 0.20.1 (https://csblatvia.github.io/vardpoor/).

Eurostat (2018), 'Reference metadata in ESS standard for quality reports structure' (https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/6f7191df-c72d-4537-aee4-81972188497d/details).

Hagenaars, A., de Vos, K. and Zaidi, M.A. (1994), *Poverty Statistics in the Late 1980s: Research Based on Micro-data*, Office for Official Publications of the European Communities, Luxembourg.

Jantti, M., Törmälehto, V. and Marlier, E. (2013), *The use of registers in the context of EU-SILC: Challenges and opportunities*, Publications Office of the European Union, Luxembourg (https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-TC-13-004).

Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, Wiley, Hoboken, NJ.

11 Weighting for a modular structure

Andy Fallows ([70])

11.1. Introduction

The objective of this chapter is to discuss and give examples of the use of two alternative options for weighting a survey with a modular design. The term 'modular design' is used to describe surveys that have a common (or core) set of questions that are administered in the same way, plus multiple modules of questions that are each administered to a separate subsample. The options considered for weighting are a composite method that allows consistent estimation of a common variable for the whole sample and the individual modules in one run of calibration, and a two-phase method whereby initially the whole sample is calibrated and an individual module is then calibrated to estimates obtained from the first phase.

The chapter demonstrates that a modular design can reduce standard errors for variables collected in only one module by drawing on the strength of the whole sample. No noticeable difference in precision is found between the two weighting methods, and so the choice of method becomes a more practical question in which the number of modules and complexity for the user should be considered.

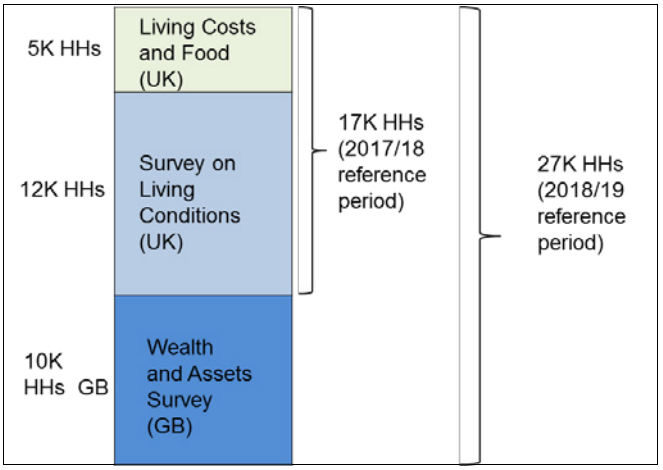
11.2. Background

Across the European Union (and indeed outside Europe), there is a move from traditional standalone surveys, whether cross-sectional or longitudinal, towards an integrated set-up whereby previously separate surveys become modules of a much larger survey with a core set of questions. A previous Office for National Statistics (ONS) project designed to implement this in the United Kingdom is described in the Integrated Household Survey user guide (ONS, 2012).

In the United Kingdom, the new Household Finance Survey (HFS) currently has a design with two modules (formerly separate surveys). A description is given in O'Neill and Webber (2020). One of these modules primarily covers household expenditure (Living Costs and Food Survey (LCF)), and the other covers household income (Survey on Living Conditions (SLC)). The SLC module has a longitudinal rotating panel design with six waves, with the same questions asked at each wave. Some European Union Statistics on Income and Living Conditions (EU-SILC) variables are in both the LCF and the SLC, whereas some (relating to change over time) are asked only in the SLC. There are plans to introduce an additional module that explores household assets (Wealth and Assets Survey) into this design (Figure 11.1). By bringing these previously separate surveys together, we can improve the precision of indicators for EU-SILC and help to improve understanding of household finance in the United Kingdom. The analysis presented in this chapter was conducted on the current two-module version of the survey.

([70]) Andy Fallows is with the UK Office for National Statistics.
The author would like to thank Paul Smith for his very useful comments. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Andy Fallows (andy.fallows@ons.gov.uk).

Figure 11.1: Sample design for the HFS

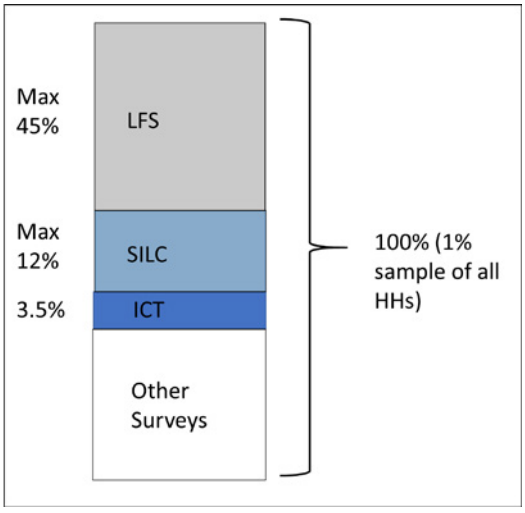


NB: HH denotes household.

The HFS sample of addresses is clustered by post-code sector, of which there are approximately 9 000 in the United Kingdom. Each sampled post-code sector is assigned to either module, with 15 households selected for the SLC and 18 for the LCF (in 2017/2018). It is not currently possible to assign addresses to both modules within the same sector, as a reduced number of addresses for the SLC leads to inefficiencies for the field interviewers (due to attrition).

In Germany, the micro-census (MZ) has been set up in a similar way. EU-SILC is just one of several possible modules that can follow on from the core set of questions (Figure 11.2). Marder-Puch (2018) gives more details.

Figure 11.2: Sample design for MZ2020



NB: HH denotes household, LFS denotes Labour Force Survey, and ICT denotes Information and Communication Technology survey.

The challenge from a weighting perspective is how to best draw strength from the entire sample to ensure the most accurate estimates possible while also ensuring that the estimates are consistent, regardless of whether the whole data set is used or one of the modules.

This chapter explores the possibility of using a composite calibration weighting method as suggested in Merkouris (2013) and contrasts it with a more traditional two-phase approach in which the entire sample is weighted and a follow-on module asked of a subsample is weighted to match selected resulting estimates (in practice each additional module would also need to be weighted). Calibration plays the key role here. These methods are described in more detail in Section 11.3.

The data were calibrated to standard age–sex and region population totals, as well as total number of households, so any core variables that are correlated with these should achieve similar estimates to those from the whole sample. Further to this, with the composite method we can calibrate to ensure consistent estimates on selected common variables. With the two-phase approach, we can calibrate a module to estimates from the whole sample. These approaches should achieve the result of ensuring that estimates are consistent on selected EU-SILC variables. This analysis explored which of these methods is the most successful in achieving the aims of consistency and reduced variance.

This initial analysis used wave 1 of the UK HFS (around 9 000 households) and thus did not require the attrition weighting that is used to weight the whole survey. At the time of writing, only the first year of data from the HFS were available, and so the only way to consider the longitudinal nature of the weighting would have been through simulation. This is something that could be considered further in the future. (Some general considerations in longitudinal weighting are discussed in Lynn and Watson, 2021.)

This cross-sectional weighting involved calculation of design weights through the inverse of the probability of selection, a non-response adjustment and, finally, calibration to known population totals, as is standard practice for ONS surveys, but using the composite method and the two-phase approach to ensure consistent employment estimates. Estimates were produced at the household level for a key demographic variable not used for the weighting (housing tenure) as well as the main EU-SILC poverty indicators.

SAS and the Canadian generalised estimation system (GES) were used for the analysis.

11.3. Approach

Weights for the SLC module were produced in three different ways: using the composite calibration method, the two-phase method and a direct method that served as a benchmark against which to assess the performance of the first two methods. All three methods are described below.

The composite calibration method allows calibration of the entire sample and modules in one go. The application of GES was modified slightly to take account of the respective sample sizes of the modules – utilising the model variance functionality. The data were calibrated to the standard age–sex, region and household totals (in which each module has its own set of totals), but the calibration matrix was expanded to include a variable (or variables) that is common to both modules, for example employment (labour force status). In the matrix, this variable was coded in a different way for each module. An individual in the SLC who is employed was coded as 1; an individual in the LCF who is employed was coded as – 1. Both were coded as 0 if not employed. In the population totals, the total for this constraint was set as 0. Additional dummy variables can be added for each possible labour force status (i.e. unemployed and inactive; see Table 11.1 for a person-level example of this coding). An expanded version can be found in the appendix.

Calibration will then satisfy the age–sex, region and household totals for both modules jointly and separately, and, in addition, produce consistent distributions of employment regardless of whether the entire data set or one of the modules is used. If an additional module is brought into the survey, a set of dummy variables will be required for each possible pair of modules (nC2). Thus, three modules will require three sets of dummy variables, but

Table 11.1: Example of how labour force status is coded for composite calibration

Person	Module	Status	Employed	Unemployed	Inactive
1	SLC	Employed	1	0	0
2	SLC	Unemployed	0	1	0
3	SLC	Inactive	0	0	1
4	SLC	Employed	1	0	0
5	SLC	Unemployed	0	1	0
6	LCF	Employed	– 1	0	0
7	LCF	Unemployed	0	– 1	0
8	LCF	Inactive	0	0	– 1
9	LCF	Employed	– 1	0	0
10	LCF	Inactive	0	0	– 1

four modules, for example, will require six sets. This is something that should be considered further, as the number of calibration constraints will further affect the variance of the weights.

For the two-phase method, the entire data set was calibrated to the standard age–sex, region and household totals. A calibrated weight was obtained and used to produce estimates of employment status. The SLC module was then weighted (using the initial design weights) to not only the standard calibration variables but also the employment estimates obtained from the first phase.

Finally, the SLC module was weighted directly (not making use of employment) to population totals to provide a benchmark against which to compare these approaches. From this, three sets of weights were obtained and estimates were produced for several variables, which were then compared using standard errors calculated using a bootstrap approach.

The variables used for calibration (age–sex, region and household totals) are currently the only options available in the United Kingdom from a non-survey source. The choice of additionally using (survey-based) employment was because this is correlated with both housing tenure and the poverty indicators. This choice of variable is something that could also be considered further in the future.

11.4. Results

Estimates (proportions) were initially calculated for each category of housing tenure. Standard errors were calculated using a standard Taylor series approximation approach (as programmed in SAS) for each of the three methods. This takes account of the sample design but not the calibration. As can be seen in Table 11.2, there is almost no difference between the different approaches.

Table 11.2: Standard errors for housing tenure using the standard approach

Housing tenure	Composite		Two-phase		Direct	
	Estimate	SE	Estimate	SE	Estimate	SE
Missing	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
Owned outright	0.3290	0.0081	0.3289	0.0081	0.3296	0.0081
Mortgage	0.2990	0.0075	0.2994	0.0075	0.2970	0.0074
Mix rent–mortgage	0.0048	0.0012	0.0048	0.0012	0.0048	0.0012
Rented	0.3557	0.0092	0.3554	0.0092	0.3571	0.0092
Rent free	0.0112	0.0017	0.0112	0.0017	0.0112	0.0017

NB: SE, standard error.

Standard errors were then calculated using a bootstrap approach, whereby within each stratum a sample of primary sampling units ($m_h = n_h - 1$) was selected with replacement. Each replicate sample was drawn using proc surveyselect in SAS, whereby a different random number seed was used for each sample (and method). The design weights were appropriately adjusted and calibrated for each sample, using the specified method. Estimates were calculated for each of the 1 000 replicate samples (the number after which the estimates converged to four decimal places), and standard errors were calculated (by combining stratum variances with scaling to account for the sample size and number of replicates). These standard error estimates can be seen in Table 11.3 (the estimates themselves are the same as in Table 11.2).

Table 11.3: Standard errors for housing tenure using bootstrapping

Housing tenure	Composite		Two-phase		Direct	
	Estimate	SE	Estimate	SE	Estimate	SE
Missing	0.0003	0.0003	0.0003	0.0002	0.0003	0.0003
Owned outright	0.3290	0.0061	0.3289	0.0062	0.3296	0.0062
Mortgage	0.2990	0.0069	0.2994	0.0067	0.2970	0.0070
Mix rent–mortgage	0.0048	0.0012	0.0048	0.0012	0.0048	0.0011
Rented	0.3557	0.0076	0.3554	0.0074	0.3571	0.0077
Rent free	0.0112	0.0017	0.0112	0.0017	0.0112	0.0017

NB: SE, standard error.

As can be seen, both the composite and the two-phase methods result in a reduction in the standard errors for the largest categories compared with the standard direct method, which reflects the additional sample used for the weighting. There is almost no difference between the two methods under investigation, although the two-phase method gives marginally smaller standard errors for two of the three largest categories.

This analysis was then repeated using the EU-SILC poverty indicators. The severe material deprivation rate (Sev_Dep) and low work intensity (LWI) indicators are simple proportions, whereas at risk of poverty (ARPT60), which is the proportion below 60 % of the median income, and the summary variable (at risk of poverty or social exclusion (AROPE)), whether in poverty according to any of the indicators (including ARPT60), are more complex, since the variability around the median also needs to be accounted for in the variance estimation. Table 11.4 gives the estimates using the standard approach and Table 11.5 presents the estimated standard errors obtained through bootstrapping. With the standard approach, the figures are again extremely consistent across the three methods. The bootstrapped estimates show that the composite and two-phase methods both reduce the standard error of the estimates relative to direct estimation. This demonstrates that there is a benefit in precision from using a modular approach. There appears to be little difference between the two methods, although this time the composite method gives slightly smaller standard errors than the two-phase method for three of the four indicators.

11.5. Conclusion

The modular design reduces the standard errors for housing tenure and the key EU-SILC poverty indicators when compared with weighting the data directly as a standalone survey. This reduction is small but consistent. This gain is the result of the additional statistical information provided by the larger sample size, although it is probably partially offset by the additional variability introduced into the weights by calibrating twice (the two-phase method) or the additional calibration constraints (the composite method).

Table 11.4: Standard errors for EU-SILC poverty indicators using the standard approach

Indicator	Composite		Two-phase		Direct	
	Estimate	SE	Estimate	SE	Estimate	SE
Sev_Dep	0.0554	0.0041	0.0553	0.0041	0.0560	0.0041
LWI	0.0750	0.0045	0.0746	0.0045	0.0769	0.0047
ARPT60	0.1846	0.0067	0.1865	0.0067	0.1858	0.0067
AROPE	0.2395	0.0078	0.2411	0.0078	0.2420	0.0078

NB: SE, standard error.

Table 11.5: Standard errors for EU-SILC poverty indicators using bootstrapping

Indicator	Composite		Two-phase		Direct	
	Estimate	SE	Estimate	SE	Estimate	SE
Sev_Dep	0.0554	0.0040	0.0553	0.0039	0.0560	0.0041
LWI	0.0750	0.0041	0.0746	0.0042	0.0769	0.0046
ARPT60	0.1846	0.0063	0.1865	0.0066	0.1858	0.0068
AROPE	0.2395	0.0061	0.2411	0.0062	0.2420	0.0064

NB: SE, standard error.

Comparisons of the standard errors resulting from the two-phase and composite methods provide somewhat mixed findings, although for the EU-SILC poverty indicators the composite method appears preferable. In addition, the composite method requires only one run of calibration, resulting in a single set of weights, and thus is simpler for users, although the weights would need to be rescaled (reduced in size by a constant factor of 0.5) if using the entire sample to estimate totals.

This analysis has been run on a data set with two modules. In the United Kingdom, an additional module will shortly be added to the integrated survey, necessitating the addition of further calibration variables. GES gave a warning when calibrating ('weights are proximal'), and, although the weights gave the correct population totals, this is a slight cause for concern; given the increased complexity if a third module is added (in terms of set-up and the calibration converging), the two-phase method seems to be a more stable option.

There are a number of ways in which this analysis could usefully be extended. Alternative calibration variables could be considered, for example relating to household size, with some exploration of how these correlate with the variables of interest. The comparison of weighting approaches could be extended to the longitudinal weighting and to a situation in which an additional module is added to the design.

11.5.1. Recommendations

- Running EU-SILC as part of a modular design can help improve precision for key poverty indicators. It also provides potential for smaller standard errors through a reduction of the clustering effect.
- This improvement in precision can be realised by weighting the data using the composite or two-phase calibration method.
- If there are only two modules in the design, then the composite method may be slightly preferable.
- If there are more than two modules in the design, then the two-phase method is simpler to implement and is more likely to converge.

References

Lynn, P. and Watson, N. (2021), 'Issues in weighting for longitudinal surveys', in Lynn, P. (ed.), *Advances in Longitudinal Survey Methodology*, Wiley, Chichester, pp. 447–468.

Marder-Puch, K. (2018), 'The new German microcensus as an integrated survey for household statistics', paper presented at the European Statistical System Workshop on LFS Methodology, May, Reykjavik.

Merkouris, T. (2013), 'Composite calibration estimation integrating data from different surveys', in *Proceedings of the 59th ISI World Statistics Congress*, International Statistical Institute, The Hague, pp. 25–30.

O'Neill, J. and Webber, D. (2020), *Improving the Measurement of Household Income*, Office for National Statistics, London (https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/improvingthemeasurementofhouseholdincome).

ONS (Office for National Statistics) (2012), *IHS User Guide Volume 1: Background and methodology 2012* (http://doc.ukdataservice.ac.uk/doc/7061/mrdoc/pdf/ihs_user_guide_volume_1_2012.pdf).

Appendix: Expanded example of coding for composite calibration

Person	Module	Sex	Region	Status	SLC							LCF							Employed	Unemployed	Inactive
					England	Scotland	Wales	Northern Ireland	Female	Male		England	Scotland	Wales	Northern Ireland	Female	Male				
1	SLC	Male	England	Employed	1	0	0	0	0	1		0	0	0	0	0	0		1	0	0
2	SLC	Female	Scotland	Unemployed	0	1	0	0	0	0		0	0	0	0	0	0		0	1	0
3	SLC	Male	Wales	Inactive	0	0	1	0	1	0		0	0	0	0	0	0		0	0	1
4	SLC	Female	Northern Ireland	Employed	0	0	0	1	0	0		0	0	0	0	0	0		1	0	0
5	SLC	Male	England	Unemployed	1	0	0	0	0	1		0	0	0	0	0	0		0	1	0
6	LCF	Female	Scotland	Employed	0	0	0	0	0	0		1	0	0	0	1	0		−1	0	0
7	LCF	Male	Wales	Unemployed	0	0	0	0	0	0		0	1	0	0	0	1		0	−1	0
8	LCF	Female	Northern Ireland	Inactive	0	0	0	0	0	0		0	0	1	0	1	0		0	0	−1
9	LCF	Male	England	Employed	0	0	0	0	0	0		1	0	0	0	0	1		−1	0	0
10	LCF	Female	Scotland	Inactive	0	0	0	0	0	0		0	1	0	0	1	0		0	0	−1

12 Weighting panels together for cross-sectional estimation

Olena Kaminska ([71])

12.1. Introduction

Rotational panels such as EU-SILC, in which a new rotational group joins the sample at regular intervals and remains in the survey for a few waves, present a unique challenge for non-response weighting. Although the oldest rotational group has the highest attrition level at any point in time, the new rotational group that started in the current wave has no attrition at all. Combining these rotational groups for cross-sectional analysis presents the question of how best to control for attrition – the topic of this chapter.

Rotational panels have many advantages (Smith, Lynn and Elliot, 2009) in comparison with fixed panels, in which participants are selected at one point in time and followed thereafter. Rotational panels suffer from less attrition than fixed panels, because they follow participants for a shorter period of time (for example for 4 years). They are likely to have less influence on participants' attitudes or behaviours due to participation in the panel (fewer panel effects), as participants stay in the panel for a shorter time, and these effects are also balanced over all time periods. In addition, rotational panels include coverage of new entries to the population (e.g. immigrants) at all time points, enabling them to provide consistent cross-sectional estimates.

([71]) Olena Kaminska is with the Institute for Social and Economic Research at the University of Essex, Colchester, United Kingdom. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. All errors are the author's responsibility. Correspondence should be addressed to Olena Kaminska (olena@essex.ac.uk).

As with any survey, a rotational panel encounters non-response. This consists of two parts: each rotational group is subject to non-response at wave 1 and attrition at subsequent waves. Non-response at wave 1 is different in nature to subsequent attrition in a number of ways. A part of non-response at wave 1 may include ineligible elements (e.g. vacant houses or businesses). For households that were never contacted, no information is available except that from a register or a sampling frame, or information obtained from the interviewer's observations; usually, not much information related to the variables of interest is available. Non-response due to attrition is much more informative: a household or an individual has been interviewed at least at wave 1, so the answers to the questions from a previous wave are known. Some information is time invariant, such as sex, date of birth, the age that a person started their first job and the person's father's salary when said person was 16 years old. Other variables, although time variant, are often strongly autocorrelated. By knowing someone's salary, job and education level in a previous year, it is possible to better predict the salary in the current year than if such information is not available, for example. Thus, correction for non-response due to attrition is potentially more powerful, due to the informative auxiliary data.

In a fixed panel, in which a sample is selected at wave 1 and individuals are usually followed and interviewed at regular intervals, weights are used to correct for attrition (Lynn and Watson, 2021). At each wave, this correction can be done using one of three approaches.

- Cumulative adjustments can be made on a wave-on-wave basis, that is, predicting non-

response between two consecutive waves using information from the previous wave.
- Longer periods between waves can be used for cumulative non-response correction, for example predicting non-response at wave t conditional on response at wave $t-2$, regardless of response outcome at $t-1$, using auxiliary information from wave $t-2$.
- Adjustment can be made for non-response due to attrition since wave 1 in a single step by predicting the attrition using information from wave 1 only.

In Section 12.4, these three methods are extended to the rotational panel context and form the basis of the comparison of methods presented in Section 12.5. A cross-sectional estimate using such panels uses cross-sectional weights, which may be created by weight-sharing the longitudinal weights (Lavallée, 2007).

In a rotational panel, at any one point in time cross-sectional estimates are created by combining a number of rotational groups, each at a different point in their participation history and thus with a different level of non-response, and potentially different correlates of non-response. Figure 12.1 illustrates this sample structure for EU-SILC data released in survey year t, under the simplifying assumption that attrition is monotone. The green dashed line indicates the cross-sectional data set, and the red dashed line indicates the longitudinal data set. Group i is the rotational group for which data collection begins in year i: the participating sample size at wave 1 is denoted by n_i, whereas the proportion of wave j respondents of panel i also participating at wave $j+1$ is denoted by r_{ij} ($j = 1, 2, 3$). Note that, if we make the simplifying assumptions that neither the wave 1 sample size nor the wave-specific response rates vary between rotational groups, then the subscript i can be dropped. In this situation, the size of the responding sample in the cross-sectional data file is $((((r_3 + 1)\, r_2 + 1)\, r_1) + 1)n$.

Theoretically, one could use the most recent rotational group for a cross-sectional estimate, as this group does not suffer from attrition. However, using just one rotational group will provide much lower statistical power due to the much smaller sample size (n, rather than $(((r_3 + 1)\, r_2 + r_1) + 1)n$). Thus, the question arises of how best to correct for different levels of non-response across different rotational groups, in terms of bias and statistical efficiency, in order to be able to use the combined sample for cross-sectional estimation. Importantly, the three methods for weighting a single panel described above result in different bias reduction only if they differ in correlates of non-response. If the multivariate correlates are the same, the bias reduction will be exactly equivalent, and the method preference should be guided by variance reduction instead.

A peculiar feature of studying attrition in a rotational panel is that it is always possible to construct a cross-sectional estimator that is not subject to attrition bias by using only the most recent rotational group. This group, in the context of a rotational panel, is in its first wave and by definition has no non-response due to attrition. As the aim here is to assess the effectiveness of weighting methods at removing attrition bias, throughout the chapter this estimate based on only the most recent rotational group will be referred to as the 'target' estimate. Aside from the effect of immigrants (popula-

Figure 12.1: Sample structure for EU-SILC data released in survey year *t*

Group (*i*)	Survey year			
	***t* - 3**	***t* - 2**	***t* - 1**	***t***
1	n_1	$r_{11}n_1$	$r_{12}r_{11}n_1$	$r_{13}r_{12}r_{11}n_1$
2		n_2	$r_{21}n_2$	$r_{22}r_{21}n_2$
3			n_3	$r_{31}n_3$
4				n_4

NB: Cell entries are sample sizes. n_i is the first-wave responding sample size for panel i, whereas r_{ij} is the response rate for panel i at wave $j+1$, conditional on response at wave j. The green dashed line indicates the cross-sectional data set, and the red dashed line indicates the longitudinal data sets.

tion entrants between panels), it can be expected that all rotational groups will provide similar estimates to each other – and any differences found can be attributed to attrition (and sampling error, which is random in nature and can thus be ignored for our purposes). It is not unreasonable to ignore the effect on estimates of immigration, as the 1-year between-wave interval in EU-SILC means that the numbers of immigrants between panels are typically small in comparison with the whole population.

In this chapter, different methods to correct for non-response due to attrition in a rotational panel are compared with the aim of finding the best general method. The chapter starts with a description of the data used (Section 12.2), followed by attrition evaluation across countries (Section 12.3) and a discussion of how to compare different weighting methods (Section 12.4). The results are described in Section 12.5 and concluding remarks are provided in Section 12.6.

12.2. Data

This study uses data from the EU-SILC (see Chapter 2) user database (UDB) version 2016. Specifically, cross-sectional estimates are explored for 2016 using four rotational groups, which started in 2013, 2014, 2015 and 2016. Thus, cross-sectional estimates are based on sample members from each of the four sample subgroups (rotational groups): the 2016 group with no attrition as this is its first wave, the 2015 group at wave 2 with one wave of attrition, the 2014 group at wave 3 with two waves of attrition, and the 2013 group at wave 4 with three waves of attrition.

EU-SILC is a cross-national study, and in this analysis information from 27 countries is used. Four countries (Germany, Ireland, Lithuania and the United Kingdom) are excluded from this analysis, as information on the longitudinal data set was not available at the time of writing. Most of the countries in EU-SILC follow a four-wave rotational panel structure, interviewing individuals once per year four times before the rotational group terminates. A few countries have a longer period of interviewing: each rotational panel in France lasts for 8 years, and in Bulgaria each rotational panel lasts for 6 years. For consistency, the four most recent rotational groups in these countries, which started in 2013, 2014, 2015 and 2016, are included.

For the purpose of this chapter, response to the personal interview is studied, and for consistency information obtained from registers only is excluded, as it is not available for all countries. In addition, cases of full record imputation are dropped from the analysis. The attrition model is conditioned on response to the personal interview in the first wave of the rotational group. This first wave interview provides rich information for attrition correction that is relevant to the questions of interest in the 2016 wave (income and health – see below). Finally, households in which all members died between their first wave and 2016, those who moved out of the country, those who moved into an institution and households that do not contain any sample member are also excluded from the analysis. However, households that moved within the country are included.

The effect of different weighting on two estimates is explored: the lower quartile of the distribution of equivalised disposable household income (low income henceforth) and the proportion of those with self-reported poor health (poor health henceforth). The low-income measure is calculated in the following way: total disposable household income (HY020 as documented in Eurostat, 2017) is divided by the square root of the number of household members. The estimate used in the analysis is the country-specific lower quartile (25% mark) of the low-income measure. The measure of poor health has two categories: the first category represents those with the poorest health (bad and very bad general health as self-reported in PH010), and the second category indicates better health (very good, good and fair health as self-reported in PH010).

12.3. The nature of attrition in a rotational panel

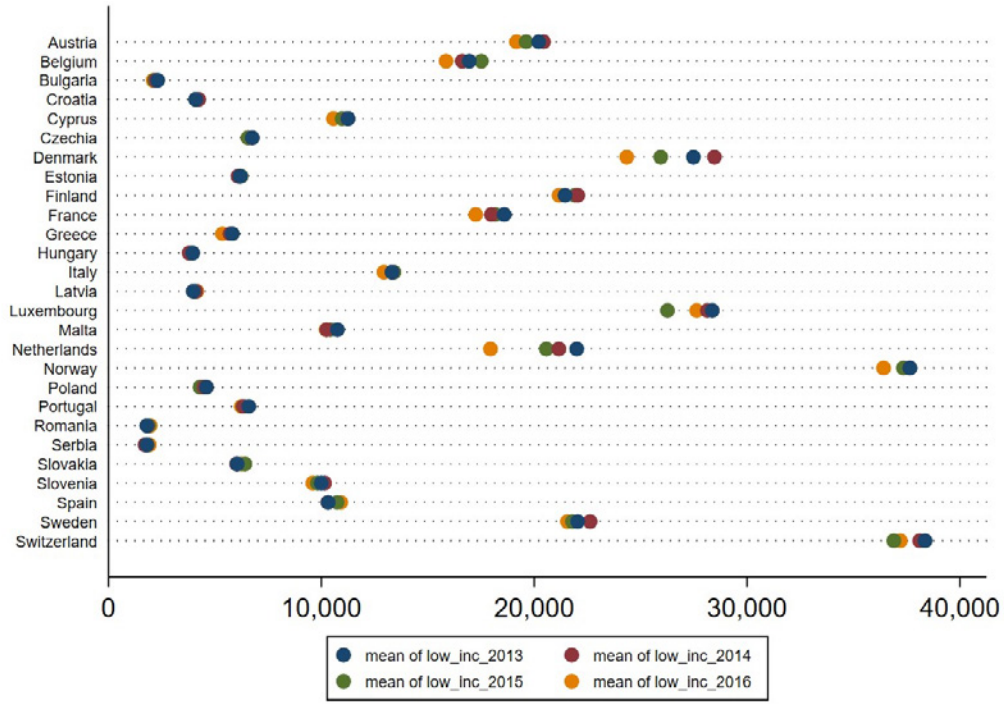
Imagine that an analyst wants to construct a cross-sectional estimate in 2016 using EU-SILC data.

For this estimate, a weight is needed that correctly adjusts for the non-response that arises from attrition in each of the rotational groups: 2013, 2014 and 2015. But how different is the bias due to attrition in each of the rotational groups? In this section, an estimate from the 2016 rotational group is considered a target value, with which estimates from the 2013, 2014 and 2015 rotational groups are compared. The difference between these estimates indicates attrition bias observed for different rotational groups at a particular historical moment in their panel life.

First, bias due to attrition for estimates of low income and health status is explored (Figures 12.2 and 12.3). To do this, base weights (PB050) are used that correct for unequal selection probabilities and non-response at wave 1 for rotational groups that started in 2013, 2014 and 2015, as well as an equivalent cross-sectional weight (PB040) for the rotational group that started in 2016. In six countries (Denmark, Finland, the Netherlands, Norway, Slovakia and Sweden), only the selected respondent, rather than all household members, is asked about health status. For these countries, the appropriate base weight (PB080) is used for the older rotational groups, and a cross-sectional weight (PB060) is used for the most recent rotational group when looking at health status. Essentially, these weights include everything but the correction for attrition.

Figure 12.2 presents estimates in euro for the first quartile of disposable household income per country estimated separately for each rotational group. All estimates are of low income in 2016, and if there were no attrition all four estimates would be expected to be the same in each country. Given that the rotational group that started in 2016 has not yet encountered attrition, it is possible to treat the estimate based on this group as a target value (marked in orange) and compare other estimates with it.

Figure 12.2: Estimates of low income (first quartile) across four rotational groups



NB: The *x*-axis indicates income in euros, with the dots representing the 25th percentile of income estimated separately for each of the four rotational groups . Thus, if the orange dot (estimate for the 2016 panel) is left-most on a row and the blue dot (estimate for the 2013 panel) right-most, this implies that people with higher incomes are more likely than people with lower incomes to have dropped out of the survey between the first and fourth waves, so the sample is becoming progressively poorer across the waves (e.g. in the Netherlands), and vice versa.

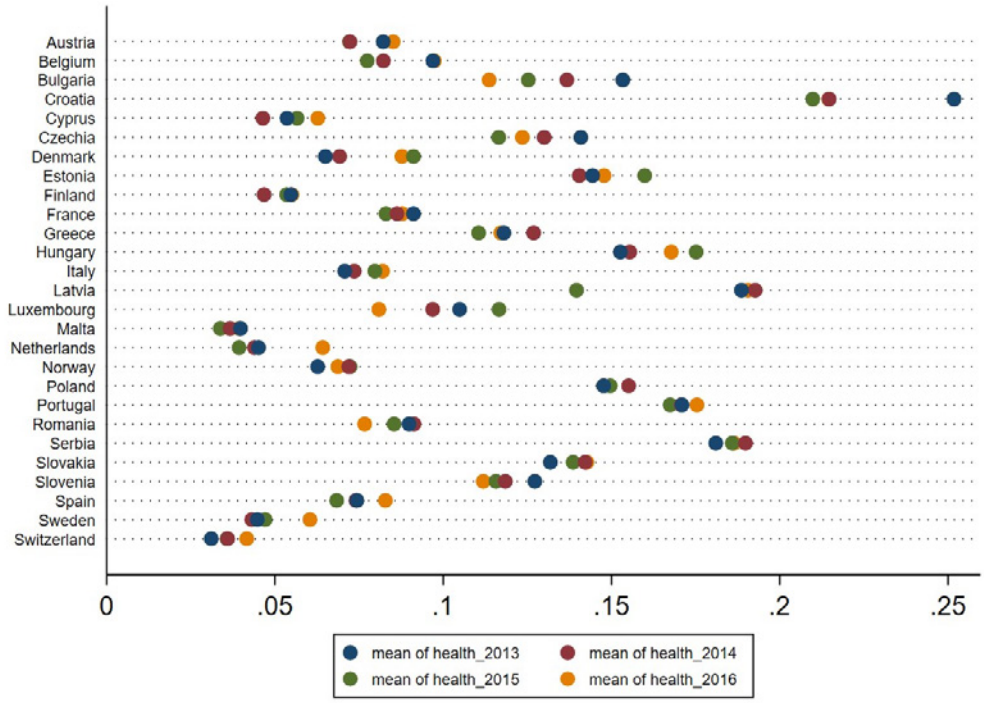
Source: EU-SILC cross-sectional 2016 UDB.

A few points can be observed from Figure 12.2. First, attrition bias is relatively small for many countries – as the dots are often grouped together. Second, for the countries with a lower value of low income the absolute difference is smaller, and therefore the dots are closer. Third, for countries where a pattern is visible, the value of the first quartile of income tends to be higher for older rotational groups – suggesting that with time those with lower incomes have a higher tendency to drop out.

Looking at poor health, namely the proportion of those who self-reported that their general health is bad or very bad according to PH010 (Figure 12.3), bias seems to be present for almost all the countries to some extent, as the dots do not align with each other and the orange dot representing the value without attrition is sometimes far from other dots. Nevertheless, there is no clear pattern of dropout across the countries (Banks, Muriel and Smith, 2014). It could be expected that people with worse health might tend to drop out from the panel (Chatfield, Brayne and Matthews, 2005), which would result in a smaller proportion of people with poor health in older rotational groups. This pattern is observed in Cyprus, the Netherlands, Spain, Sweden and Switzerland. In a handful of countries, for example Bulgaria, Luxembourg, Romania and Slovenia, people with bad health tend to stay in the panel, resulting in a higher proportion of people with bad health in older rotational groups. For other countries the pattern is mixed.

The cross-country variability in the extent and nature of bias due to attrition and the patterns across rotational groups suggest that the bias is very much country specific and specific to each wave. It is therefore likely to be important to adjust for attrition based on models that tailor the selection of predictors to each country's context.

Figure 12.3: Proportion of those with poor health across four rotational groups



NB: The *x*-axis indicates the proportion of respondents reporting poor health. Thus, if the orange dot (estimate for the 2016 panel) appears to the left of the blue dot (estimate for the 2013 panel), this implies that people in good health are more likely than people in poor health to have dropped out of the survey between the first and fourth waves, so the sample is becoming progressively healthier across the waves (e.g. in Bulgaria), and vice versa.

Source: EU-SILC cross-sectional 2016 UDB.

12.4. Comparison of weighting methods: method

The previous section showed that the estimates from older rotational groups for many of the countries would suffer from attrition bias if no adequate adjustment were made. To adjust for this a user can use weights. A weight used in analysis is a product of several different parts: a sampling weight that corrects for differential selection probabilities, correction for household non-response at wave 1, correction for within-household non-response at wave 1 if appropriate, adjustment for new population entrants (immigrants since the beginning of the oldest rotational group and 16-year-olds) and, finally, correction for attrition. It is the last part that this chapter explores, specifically how best to correct for attrition in a rotational panel. All the other parts of the weight are the same across the weighting methods below and would be multiplied by the attrition correction to create a final weight for user analysis.

First, the 'naïve' estimate that has no correction for attrition is explored. For this, a base weight (PB050) for the 2013, 2014 and 2015 rotational groups is used, and an equivalent cross-sectional weight (PB040) for the 2016 rotational group is used, as in the analysis presented in Section 12.3. Again, for poor health, on which only selected respondents are interviewed (in Denmark, Finland, the Netherlands, Norway, Sweden and Slovenia), a base weight (PB080) and cross-sectional weight (PB060) are used. In this way, the base weight for people interviewed in the 2016 survey year is obtained. This excludes any attrition correction but includes all other weight components described above.

The target estimate is obtained from the most recent rotational group only – the 2016 group, which has not yet encountered attrition. For this, a cross-sectional weight (PB040 or PB060 for health status in the six specified countries) is used.

We compare estimates obtained using each of three adjustment methods introduced below with estimates using the EU-SILC current weights, which will be referred to as 'original' weights. The original weights were derived using wave-on-wave prediction of attrition as follows. For the 2013 rotational group, attrition is predicted in three steps: (i) response in 2014 is predicted using predictors from the 2013 interview (the model is conditional on response in 2013); (ii) in the next model, response in 2015 is predicted conditional on response in 2014 using predictors from 2014; and, finally, (iii) in a third model response in 2016 is predicted conditional on response in 2015 using predictors from 2015. Using each of the above models, conditional predicted probabilities of responding at each wave (2014, 2015 and 2016) are obtained. The product of these three probabilities indicates a total probability of continually responding in all waves conditional on participation in the 2013 wave. The inverse of this total probability indicates the weighting correction for attrition in the 2013 rotational panel. A similar process is then followed for the 2014 group (with two models reflecting two opportunities to drop out) and the 2015 group (with one model for drop-out in 2016). In the end, each respondent has an attrition adjustment, which after multiplying by the base weight and correcting for new population entrants (immigrants and those who have turned 16 since 2013) results in a combined weight. This combined weight may also be post-stratified (calibrated against external measures).

Section 7(4) of the annex to Commission Regulation (EC) No 1982/2003 on sampling and tracing rules (European Commission, 2003) provides the following general guidance on calculating weighting adjustments in EU-SILC:

> *Weighting factors shall be calculated as required to take into account the units' probability of selection, non-response and, as appropriate, to adjust the sample to external data relating to the distribution of households and persons in the target population, such as by sex, age (five-year age groups), household size and composition and region (NUTS [Nomenclature of Territorial Units for Statistics] II level), or relating to income data from other national sources where the Member States concerned consider such external data to be sufficiently reliable.*

It is important to note that EU-SILC current weights are calculated separately for each country by the national statistical institutes and for each rotational group prior to being combined at the last step.

More technical details on the weight construction guidelines for EU-SILC can be found in the *Methodological Guidelines and Description of EU-SILC Target Variables* by Eurostat (2017). Thus, for the 'original' estimate the provided cross-sectional weight PB040 (and PB060 if relevant) is used. The estimates for these come from the cross-sectional file.

Three additional methods to correct for attrition are developed in this chapter. For all three methods to correct for attrition, a standard set of predictors from previous waves are used. These include sex, age (three categories), marital status, degree of urbanisation, capacity to afford to pay for 1 week of annual holiday away from home, capacity to afford a meal with meat, chicken or fish (or vegetarian equivalent) every second day, capacity to face unexpected financial expenses, ability to make ends meet, dwelling type, tenure status, whether accommodation has a leaking roof, damp walls/floors/foundation, or rot in the window frames or floors, ability to keep the home adequately warm, and low income. Three more predictors were considered for inclusion in the models but were not included: self-defined current economic status (PL031), general health (PH010) and suffering from any chronic illness or condition (PH020). These questions are asked of only around a half to two thirds of household members in register countries where only one household member is interviewed in each household, namely Czechia, Denmark, Estonia, Finland, the Netherlands, Norway and Sweden. As this study aims for consistent weighting models across the countries, the three variables were excluded from all countries. Nevertheless, this does not have to be the case in practice. For predictors that have some item non-response, modal values (categorical variables) or mean values (continuous variables) within groups are imputed, where the groups are defined by the combination of country, rotational group and wave. For most variables this involves only a handful of cases, and the highest proportion imputed in any of the countries for any variable is under 5%.

The first method of correction for attrition (called the 'all' method) corrects for attrition in all rotational groups in one step. Because a user uses all rotational groups together in his or her analysis, the effect of attrition is shared across them – the total effect of a particular predictor consists of its effects in the oldest rotational group, middle rotational group and the most recent rotational group taken together. As this effect is shared in the analysis, it may also be efficient to predict it in one step. This can be done, because many questions asked in the first wave of any rotational group are the same – so the same predictors across the groups are kept. The model therefore uses predictors obtained from the first wave interview for each rotational group and predicts response in the 2016 wave conditional on response in the first wave (which can be 1, 2 or 3 years ago) in one step for all three groups combined. The group that started in 2016 is assigned an attrition adjustment of 1. To obtain an all method weight, this attrition correction is multiplied by the two base weights.

The second method developed in this chapter (called here the 'each' method) is to predict attrition between wave 1 and the 2016 wave in a single step, but separately for each rotational group. The model therefore corrects for non-response between 2015 and 2016 for the 2015 rotational group, between 2014 and 2016 for the 2014 group, and between 2013 and 2016 for the 2013 group. For each rotational group, predictors are obtained from the first wave. The predicted values of response are then reciprocated to obtain attrition adjustments. Again, the last rotational group requires no attrition adjustment. To obtain an each method weight, this attrition correction is multiplied by the base weight. Note that two sets of weights are created – one to estimate low income and another for health status – as the base weights differ between them.

Finally, the third method, the wave-on-wave method, is considered. This is a similar method to the one used to create the original weights but with a standardised set of predictors (same variables as the ones for the each and all methods) and standardised procedures across all countries. For the wave-on-wave method, non-response is predicted within each rotational group on a wave-on-wave basis. For example, for the 2013 rotational group three models for attrition are used: prediction of response in 2014 conditional on response in 2013 and using predictors from 2013; response in 2015 conditional on response in 2014 and using predictors from 2014; and response in 2016 conditional on

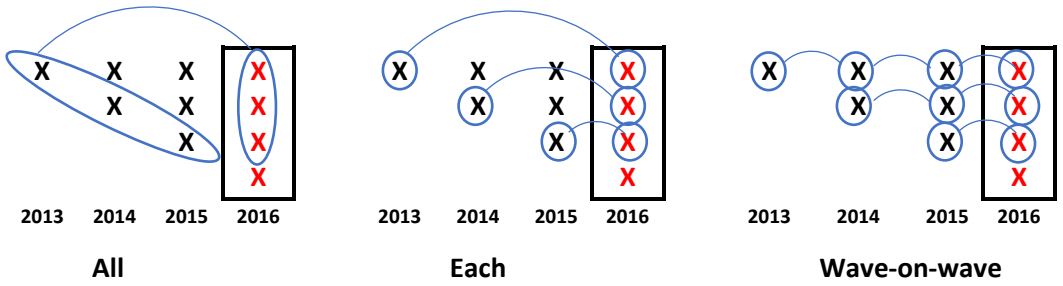
response in 2015 and using predictors from 2015. The final attrition correction is the reciprocal of the product of the response probabilities obtained from the three abovementioned models. A similar procedure is used for other rotational groups. The 2016 group obtains an attrition correction of 1, and all the attrition corrections are then multiplied by the base weights.

All the prediction models were run in Stata 15.1. The same set of variables was included across all countries and across the each, all and wave-on-wave methods. Separate models were fitted for each country. Backward stepwise logistic regression was used if only those variables or the categories of categorical variables that were significant at the p = 0.05 level were included in the model. Thus, although the original predictors specified for the model were the same for all countries and rotational groups, the variables that were retained in the final models, and the coefficients, were tailored to the country (for all three methods), specific rotational group (for the each and wave-on-wave methods) and wave-on-wave combinations (for the wave-on-wave method). No trimming of the weights – the procedure in which a number of the highest weights are capped at a set maximum value to reduce variance – was performed.

In addition, for each of the three methods, each, all and wave-on-wave weights were adjusted for new age-related entrants: 16-year-olds had the weight multiplied by 4, 17-year-olds by 2 and 18-year-olds by 1.333. This was to reflect the fact that, for example, 16-year-olds could be selected only through the 2016 rotational group – and thus had a 25 % chance of being in this sample. This procedure follows the current EU-SILC recommendations. Finally, as weights from longitudinal and cross-sectional data sets are combined, the final weights are scaled to a mean of 1 within each rotational group.

When comparing the three methods of interest, it is important to note that the amount of work involved in creating them differs: although the wave-on-wave method requires six separate models per country, the each method needs only three models per country, and the all method calculates the attrition correction in a single model for the whole panel for each country. See Figure 12.4 for a graphical depiction of this. In the figure, each arch denotes a separate model.

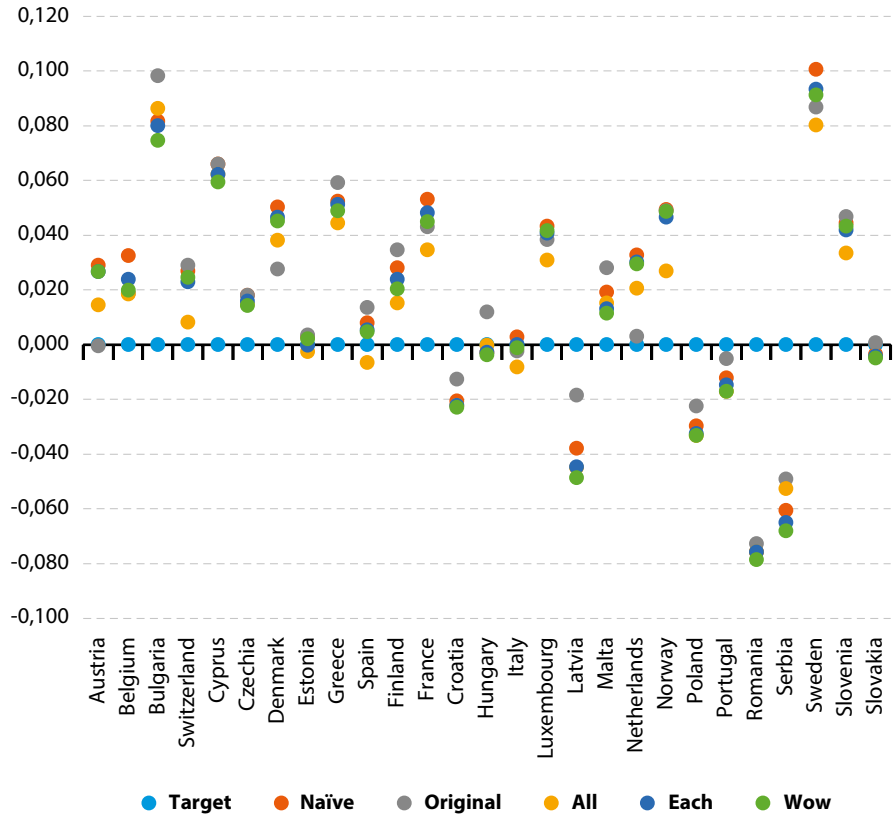
Figure 12.4: Alternative weighting approaches for a rotational panel



12.5. Results

Figures 12.5 and 12.6 present comparisons between estimates produced using each of the five different methods to correct for attrition and the target estimates. Both figures present standardised bias, which represents a relative difference between an estimate and the target estimate. It is obtained by subtracting a particular estimate from the target estimate, after which the difference is divided by the target estimate. The target estimate is therefore placed at zero (light-blue dots), and the difference between the target estimate and an estimate constitutes relative bias. The naïve bias is depicted with red dots, and this represents the relative bias if no attrition correction is applied. The aim with attrition correction is to bring this bias close to zero, and methods are compared based on how much bias reduction they provide.

Figure 12.5: Relative bias by attrition correction method for the low-income measure across countries

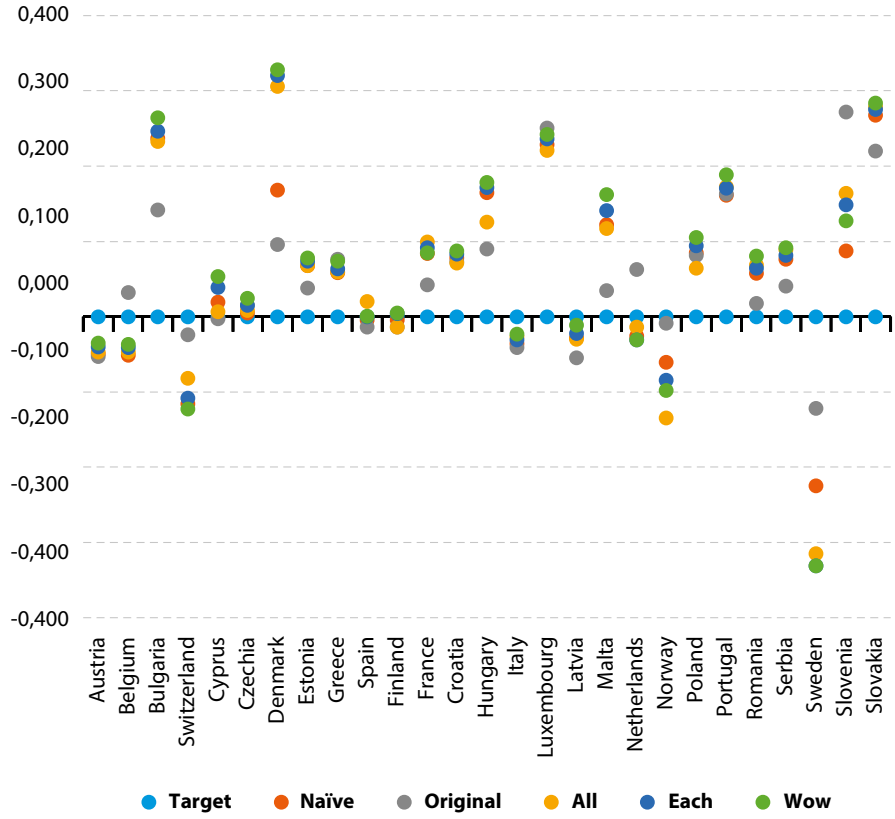


There are two particular questions that are of most interest. First, which method among those explored here performs best? Second, how do the three weighting adjustment methods presented in this chapter (all, each and wave-on-wave) compare with the current EU-SILC methods (original)?

Looking at low-income bias (Figure 12.5), the overall picture is mixed: for some countries the original method performs worse than the weighting methods developed in this chapter, for example for Bulgaria, Greece, Hungary and Malta, whereas for other countries the reverse is true, for example for Denmark, Croatia, Latvia, the Netherlands, Austria, Poland and Portugal. A more consistent pattern is observed for poor health (Figure 12.6): the original method performs the best in most of the countries. This is a very curious finding, suggesting that the calculation method may not be as important as the variables used in weighting models. As noted in the previous section, predictors used in the models developed in this chapter are highly related to income but not to health (the main general health variable was considered but not included in the models). It is therefore not surprising that the final weights perform worse for the poor health measure than for the low-income measure.

When comparing the three methods developed in this chapter, the best performance is observed consistently for the all method for low income (yellow dots in Figure 12.5). For poor health (Figure 12.6), little difference is observed across the three methods, although the all method performs just slightly better than the wave-on-wave and each methods in Bulgaria, Cyprus, Hungary, Malta, Poland and Switzerland. Noteably, for a number of countries the three methods do not correct for bias at all (the yellow, blue and green dots overlap with the orange dots), and in some cases these meth-

Figure 12.6: Relative bias by attrition correction method for health status across countries



ods increase bias by mistakenly moving the correction in the opposite direction to the one required (e.g. in Denmark, Slovenia and Sweden, where the yellow, blue and green dots are further from the zero line than the orange dots).

Finally, it is important to note that the overall relative bias due to attrition is often not large – mostly within 5–10 % of the low-income estimate and within 10–15 % of the poor-health estimate.

12.6. Conclusions

Attrition adjustment remains one of the crucial steps of inference to the population from a panel survey. A rotational panel with its complex structure offers several options for calculating this adjustment. This chapter aimed to compare bias reduction across different methods of calculating this attrition adjustment. In addition, it explored whether any of the methods may have an advantage over the current weighting methods used in EU-SILC.

The main finding is that the choice of predictor variables may be more important than the choice of modelling method in determining how well an attrition adjustment performs. This conclusion is consistent with Lavallée and Beaumont (2015) and Little and Vartivarian (2005). The three methods that were developed in this chapter performed considerably better in estimating low income than in estimating poor health. This is most likely a result of the predictors chosen for the models: they include several that are highly relevant to income but none that are highly relevant to general health. Importantly, it is possible to include the general health question in the models, as this question is asked at each wave. For comparison reasons, as this question is asked

in different ways across countries, it was excluded from the models with devastating results for poor-health estimation – all three methods developed here typically either did not improve bias in the health measure or sometimes even increased it.

It is therefore crucial for any weighting model to have predictors that are as highly related to the variables of interest as possible. For an attrition correction within a panel context in which questions are repeated at each wave, the statistician is at an advantage: the predictors from the first (or a previous) wave are typically strongly correlated with the variables in question. Even if a country has high-quality external information for post-stratification purposes, using more detailed information from earlier waves should be at least explored and potentially encouraged, as it can be more relevant to the variables of interest. This is possibly why in some countries the methods developed in this chapter outperformed the original method currently used in EU-SILC for the low-income measure, and this may also be why the original method outperformed the other three methods for the health measure.

In comparing the methods to calculate an attrition correction within a rotational panel context, there is some suggestion that the all method outperforms the each and wave-on-wave methods. It is not immediately clear why this should be the case and how generalisable this finding is to other contexts. However, two other advantages of the all method should be mentioned: first, in contrast to the other methods, which require multiple models, this method requires only one model and is therefore much more labour efficient; second, one could expect lower variance of weights (and therefore a narrower confidence interval of estimates) with the all method given that it is generated with fewer variables (only from wave 1) and with only one model (as each model has its own associated uncertainty). Further exploration and development of the all method should therefore be encouraged.

Two practical recommendations follow from this study.

- More attention should be paid to the predictors in the EU-SILC weighting models, with an attempt to include those that cover all the main research topics of EU-SILC.
- The all method can be added to the repertoire of standard methods recommended for attrition correction.

References

Banks, J., Muriel, A. and Smith, J. (2011), 'Attrition and health in ageing studies: evidence from ELSA and HRS', *Longitudinal and Life Course Studies*, Vol. 2, No 2, pp. 101–126, doi:10.14301/llcs.v2i2.115.

Chatfield, M. D., Brayne, C. E. and Matthews, F. E. (2005), 'A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies', *Journal of Clinical Epidemiology*, Vol. 58, No 1, pp. 13–19, doi:10.1016/j.jclinepi.2004.05.006.

European Commission (2003), Commission Regulation (EC) No 1982/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the sampling and tracing rules, OJ L 298, 17.11.2003, p. 29.

Eurostat (2017), *Methodological Guidelines and Description of EU-SILC Target Variables – DocSILC065 – 2017 operation (version September 2017)* (https://ec.europa.eu/eurostat/documents/203647/203704/Guidelines+SILC+2018/).

Lavallée, P. (2007), *Indirect Sampling*, Springer, New York.

Lavallée, P. and Beaumont, J.-F. (2015), 'Why we should put some weight on weights', *Survey Methods: Insights from the Field* (https://surveyinsights.org/?p=6255).

Little, R. J. and Vartivarian, S. (2005), 'Does weighting for nonresponse increase the variance of survey means?', *Survey Methodology*, Vol. 31, No 1, pp. 161–168.

Lynn, P. and Watson, N. (2021), 'Issues in weighting for longitudinal surveys', in Lynn, P. (ed.), *Advances in Longitudinal Survey Methodology*, Wiley, Chichester, pp. 447–468.

Smith, P., Lynn, P. and Elliot, D. (2009), 'Sample design for longitudinal surveys', in Lynn, P. (ed.), *Methodology of Longitudinal Surveys*, Wiley, Chichester, pp. 21–33.

13 Current best practice in weighting for a rotating panel

Gareth James, Mārtiņš Liberts and Peter Lynn ([72])

13.1. Introduction

In this chapter, we discuss weighting and calibration of European Union Statistics on Income and Living Conditions (EU-SILC) data sets, the options available, how choices may be determined and what represents good practice. We include a discussion on a range of situations, including weighting for both cross-sectional and longitudinal analyses.

As this is an overview and discussion of approaches, we do not seek to provide step-by-step instructions and calculations for the production of weights. Regarding EU-SILC, Eurostat provides such instructions in its *Methodological Guidelines and Description of EU-SILC Target Variables* (DocSILC065; Eurostat, 2019). Rather, we concentrate on aspects of weighting and calibration that are particular to the context of a rotating panel.

However, we summarise here some of the more general aspects of weighting, as this gives an introduction to the rest of the chapter (for a more detailed introduction, see Lynn, 2005). The broad aim of weighting is to make the responding sample representative of the target population in some way – specifically to derive population parameter estimates from responding sample data with expected precision and to achieve consistency with population statistics. In practice, that means calculating a weight and assigning it to each responding sample unit (person or household) in any given data file. The weight may then be regarded in a fairly general sense as the number of units in the target population that that responding sampled unit represents.

The calculation of a weight is complex and will depend on analysis objectives (whether the unit is being used for longitudinal or cross-sectional estimation, for example). Although presented as a single number, the weight is typically calculated as a product of at least three components, which reflect different aspects of the survey process and include the survey design, an adjustment for non-response and a calibration adjustment, each of which we examine in greater detail later in this chapter.

We assume that each EU-SILC sample has been drawn according to a specified random probability sample design, which includes the use of a sampling frame. That frame may list people, households, addresses or some other units, but it is important that a probability design is used for EU-SILC and not a non-probability design such as a quota sample. The probability that a unit is included in the sample is reflected in the design weight.

Non-response is an ever-present and growing phenomenon in social surveys. It presents a risk to the quality of survey outputs through non-response bias, which occurs when the characteristics of interest differ systematically between those who responded and participated in the survey and those who did not. That risk is mitigated by modifying the design weights to account for differential non-response patterns. In a similar way, attrition – essentially accumulated non-response or dropout between waves of a survey – may be regarded

([72]) Gareth James is with the UK Office for National Statistics; Mārtiņš Liberts is with the Central Statistical Bureau of Latvia; and Peter Lynn is with the Institute for Social and Economic Research at the University of Essex, Colchester, United Kingdom. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Gareth James (gareth.james@ons.gov.uk).

analogously, although it is usually treated with a separate adjustment.

Calibration to external sources is the final stage of weighting. Its aim is to modify the weights further so as to be able to recover known (or possibly estimated) population totals. That means that, if summed, the totals of the weights of the responding sample in particular domains will equal the known population totals in those domains.

13.2. Selection probabilities

The design used for an EU-SILC sample must be reflected in the calculation of inclusion probabilities for the selected sample and include all stages of selection. That includes cases in which, for example, addresses are selected from a frame and it is only when interviewers arrive at the address and establish who lives there that a final selection of households can be drawn. The reciprocal of the inclusion probability gives the design weight.

Calculation of design weights specific to a rotational group is generally straightforward and reflects the probability of a case being included in the wave 1 gross sample ([73]). Subsequent waves attempt to follow up all previous-wave respondents, so no new design weight calculation is needed. The final weights calculated in any given wave will form the base weights for the same cases in the next wave, which are then modified to account for attrition, changes in the eligibility of cases for EU-SILC and changes in the population over time.

It should be noted that, for an analysis that combines multiple rotational groups, the selection probability of an individual is the combined probability of being selected for any of the rotational groups in question. For example, cross-sectional EU-SILC analysis is based on combining four rotational groups – see Chapter 12 of this book. A person who has been continuously in the resident population over the past 4 years will have had four chances of selection (although not necessarily equal chances, as both sample size and population size may have fluctuated over these 4 years), whereas a person who, say, became eligible for the survey only 2 years ago (through immigration or by turning 16) would have had only two chances. For each person, then, the relevant selection probability is the sum of four component probabilities, some of which could be zero. The recommended EU-SILC approach to producing design weights is based on an approximation that assumes that the selection probability for each sampled unit remains constant over the 4 years. Although best practice is to base design weights on the actual selection probability of each case, the EU-SILC approximation is generally reasonable if the size of each new panel is the same.

([73]) The term 'gross sample' is used here to indicate the full sample of selected units, regardless of whether or not response to the survey is successfully obtained.

13.3. Adjustments for non-response and attrition

Non-response to the survey means that the number of cases in the responding data set (net sample) will be smaller than the number selected (gross sample) and that the gross and net samples may not have the same distribution in terms of important survey variables. The next stage of the weighting helps to account for this by inflating the design weights of the responding cases; it does so by dividing them by an estimated propensity to respond. This exercise depends on being able to correctly distinguish those households that are non-responders but eligible for EU-SILC from non-eligible households or addresses – a distinction that is not always obvious. We consider eligibility further in Section 13.4. This stage also requires auxiliary variables to be available to classify both responding and non-responding cases. These may be available on the sampling frame or may be linked to the EU-SILC sample from other sources.

Various approaches may be taken to estimate the propensity to respond, and all rely on being able to identify auxiliary variables that are associated (correlated) with EU-SILC indicators and also with variation in response rates (Brick, 2013). Examples of such variables may be those that relate to the households themselves – perhaps household size or composition, income levels, economic-activity status, and so on – or the geographical area in

which the household is located. Appropriate analysis should be undertaken to determine suitable variables, and decisions should be made on how frequently such analysis is undertaken, trading off the benefits of updating models frequently to get the best results for that year with greater consistency across years.

Construction of 'homogeneous groups' is currently the approach used by most EU Member States (see Chapter 9 of this book). Such homogeneous groups are defined around combinations of (suitably grouped) variables of the sort described above, and it is necessary that each responding household can then be assigned to precisely one of the groups defined. A response propensity then needs to be estimated for each group, which is usually calculated as the ratio of the design weights of the responding units to those of the selected ones, although it may be reasonable to use propensities estimated previously if current data are not available. In defining homogeneous groups, there is a balance to be struck between bias reduction and variance inflation: it is advisable not to construct groups that are too small and not to have too many groups.

The usual alternative approach to adjustment is to develop non-response models using regression modelling (logistic or similar), such that propensity to respond is predicted using a household's characteristics. As for the homogeneous groups approach, the variables are needed for both responders and non-responders. If keeping the same model for a number of years, the coefficients should be re-estimated each year if possible.

A regression-based approach seems to be the next most common form of adjustment used by Member States, with the remaining countries using combinations of the two or employing no explicit non-response adjustment at this stage, instead leaving the final calibration adjustment to mitigate potential non-response bias.

For attrition subsequent to wave 1, similar approaches may be used to estimate the propensity to remain in the sample but it is likely that different formations of the homogeneous groups or different regression models will be employed. The primary reason for using different groups or models is that, as all households in subsequent waves responded in a previous wave, we have much richer information available about both the responders and the non-responders at the current wave and can make good use of that in the attrition adjustment. For later waves, it is possible to make either incremental one-wave-at-a-time adjustments or a one-step adjustment instead. Kaminska (see Chapter 12 of this book) found that there was little difference between these approaches in the effects of the weighting in her application. The choice of auxiliary variables to include in the models may be more important. If the responding sample at a particular wave includes 're-entries' (people who responded initially but subsequently missed at least one wave before returning to respond at a later wave), the incremental approach needs to be adjusted to allow for these cases (see Chapter 9 of this book).

13.4. Unknown eligibility

Sampled households (addresses) that do not respond present a challenge, as it is not known whether they are within the scope of EU-SILC, or, if they were within the scope at a previous wave, whether they remain within the scope for the current wave. Thus, they are cases of unknown eligibility, and an explicit decision must be taken regarding how to handle such cases in non-response weighting. A simple solution, used in EU-SILC by most Member States, is to impute eligibility status for each case. Most countries either regard all such cases as eligible or regard them all as ineligible. A slightly more sophisticated form of imputation would involve taking into account address-level information, for example imputing eligibility if an address appears to be occupied and ineligibility if it appears to be unoccupied. Statistical imputation is also possible, by fitting a model to predict eligibility among the cases in which eligibility is known and then applying the model to the cases of unknown eligibility.

An additional issue, unique to longitudinal surveys, is that cases that were eligible initially (at wave 1) can become ineligible during the course of the survey, as a result of death, emigration or moving out of the household population and into the institutionalised population (Lynn and Watson, 2021). If such moves out of eligibility are not correctly identified, the at-

trition models used for non-response weighting can produce biased results. It is possible, for example, that people who have died or who have moved into residential care settings had worse living conditions, on average, than others. Incorrectly assuming that all non-responents have remained eligible could therefore result in overestimation of poor living conditions. Longitudinal surveys often go to considerable lengths to identify deaths or emigrants or to model the eligibility status of non-respondents whose status is uncertain (Lynn and Watson, 2021).

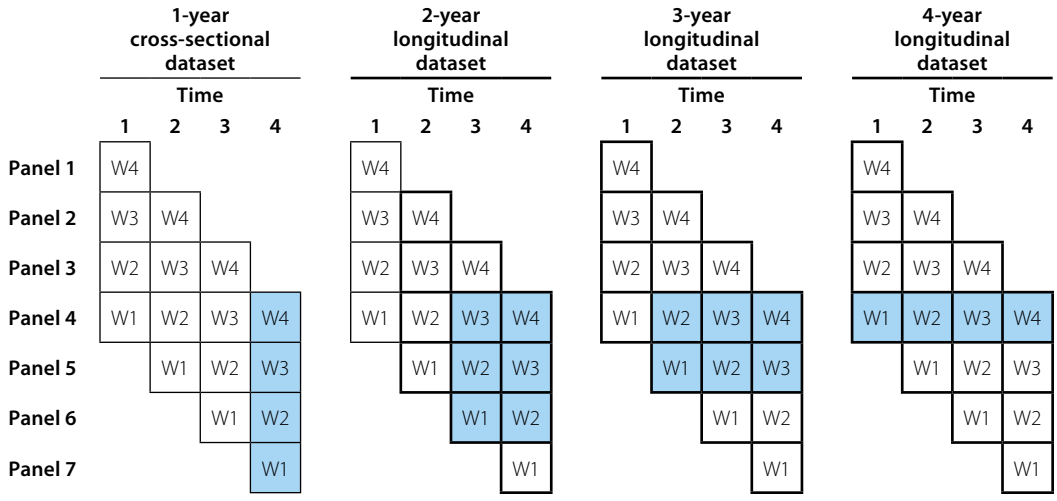
13.5. Combining panels and calibration

13.5.1. Combining panels

A 1-year cross-sectional data set comprises data from households that have responded in that year. Those households will have been in EU-SILC for varying lengths of time. For one panel (also known as a rotational group), it will be their first interview; for others, it may be their second, third or fourth, and thus four panels are aggregated to form the cross-sectional data set. In a similar way, the 2-year longitudinal data set (containing households that have responded in the two most recent consecutive years) will contain households from three panels, the 3-year longitudinal data set will contain two panels, and the 4-year longitudinal data set will contain just one panel. This is illustrated in Figure 13.1.

When combining panels, scale factors need to be used to adjust the weights of the constituent panels, each of which may have been calibrated separately, but this depends on the chosen order of operations. It is necessary to scale the weights of the constituent panels, and DocSILC065 recommends using factors of ¼, ⅓ and ½ (as appropriate for the data set being constructed); however, this would be optimum only in the case of equally sized panels with equal variance of weights. That may not be the case after a sample redesign has been implemented, and some countries instead scale to actual sample size or the number of responding households in a panel to determine the scale factors (see Chapter 9 of this book).

Figure 13.1: Illustration of how panels are combined to form cross-sectional and longitudinal data sets under a four-wave EU-SILC structure



NB: The four examples show the panels that are combined at time (year) = 4 to produce:

- a cross-sectional (1-year) data set (four panels), which is referenced to year = 4;
- a 2-year longitudinal data set (three panels), which is referenced to year = 3;
- a 3-year longitudinal data set (two panels), which is referenced to year = 2;
- a 4-year longitudinal data set (one panel), which is referenced to year = 1.

A further option is to scale each panel to the effective sample size, n_{eff}, defined as the sample size divided by the design effect; that approach should minimise the variance of estimators in the combined data set. Kish (1992) provides an approximation to n_{eff} that accounts for variation in the weights:

$$n_{\text{eff}} = \frac{\left(\sum_{i \in s} w_i\right)^2}{\sum_{i \in s} w_i^2}$$

where w_i is the weight of the i-th respondent in the sample, s, in this case of the particular panel. This quantity should be calculated separately for each panel, and the weights scaled in proportion. Note that, in the case of equal weights for all respondents (that is $w_i = w$ for all i in the panel sample), $n_{\text{eff}} = n$, and thus the scaling then reduces to be in proportion to the responding sample size.

When combining panels, as discussed in Section 13.2, the selection probability for each case is the joint probability of selection for any of the panels. To allow for this, the EU-SILC methodological guidelines suggest a simple adjustment, whereby units that could have been selected in m of the four panels (in the case of the 1-year cross-sectional data set) have their panel-specific selection weight scaled by $4/m$. A more precise approach would be to use the actual joint selection probabilities as the basis for the relevant design weight. Alternatively, an approximation that additionally takes into account variation in the size of each panel should remove the majority of the bias inherent in the current approach. With this approximation, each panel-specific weight would be scaled by $4/m \times n/n_i = 4n/mn_i$, where n_i is the size of the panel for which the case was selected and n is the total size of the four panels being combined (and analogously for the longitudinal data sets).

It is usual practice to develop non-response-adjusted weights separately for each panel, before combining the panels as described above. However, it is unclear whether this represents best practice. Kaminska (see Chapter 12 of this book) investigated the potential to fit a single non-response model after combining four panels and found that this approach could be as effective as approaches that model the non-response separately for each panel. This approach would also be simpler to implement, as it requires just a single model, but further research is needed to establish the generalisability of the method.

13.5.2. Calibration

Calibration to external sources, implemented as the final step in the weighting process, can ensure that population totals can be recovered from the sample weights, improve precision, mitigate the potential for non-response bias and ensure a degree of consistency between estimates from different sources (Lundstrőm and Särndal, 1999). All Member States use calibration, with a large majority (see Chapter 9 of this book) seeming to use the integrative approach described in Eurostat (2019).

In terms of when the calibration takes place, there are choices about the order of operations, and it is possible to calibrate EU-SILC data sets several times, although it is advisable to always recalibrate the final data set once all panels have been combined. From the point of view of variance estimation, it is advisable to calibrate weights in one step. Extra calibration steps make the estimator much more complex, and it becomes much harder to derive an appropriate variance estimator. The solution would be to use resampling variance methods such as bootstrap or jackknife, in which population parameter estimator complexity is not an issue theoretically. A general recommendation would be to use a standard generalised regression variance estimator (with regression residual estimation) if calibration is done in one step (Särndal and Lundström, 2006), but to consider resampling variance estimation methods if calibration is done in several steps. Variation certainly exists between Member States as to which weights (base weights, cross-sectional weights or longitudinal weights) are calibrated, and when that occurs in the order of processing, and whether calibration takes place on combined samples or separately by waves. There is also variation in the calibration methods employed and the software used.

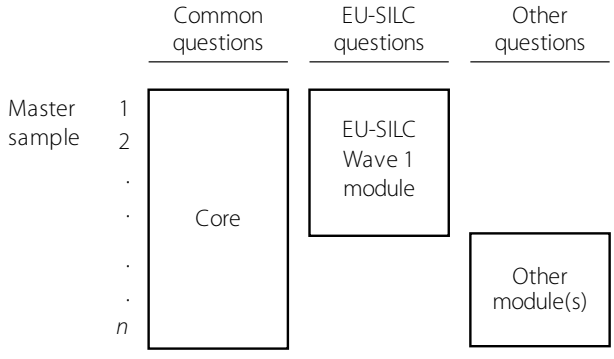
The same arguments and recommendations about choice of variables for calibration apply as for the

development of non-response and attrition models. The variables need to be available for the responding sample. There are three possible sources for calibration totals: (i) the total is extracted from an external data source; (ii) the total is derived as a total of a sampling frame variable; (iii) the total is derived as a design-unbiased estimate using sample data and design weights (Särndal and Lundström, 2006). It is possible to combine totals from multiple sources in a vector for use in single-step calibration. Totals should be numerically coherent, for example providing an equal population size from any breakdown. The best variables to improve accuracy will be the ones most strongly associated with the main EU-SILC indicators, whereas others – such as age, sex and location (geography) – may be chosen more for consistency purposes. It is difficult to be prescriptive about what variables may prove useful, but use of any available administrative data or registers, for example to provide information about income, is certainly promising. However, empirical evidence suggests varied success in terms of results and that case-by-case investigations are required (see Chapter 10 of this book).

A further option for calibration could be to use a modular design structure for the survey, in which all respondents of a larger master survey are asked a core set of common questions, whereas other question topics (modules) are asked only to particular subsamples. Such a structure has been introduced in the United Kingdom, where EU-SILC is now one of two modules within a larger survey (see Chapter 25 of this book). This design is depicted in Figure 13.2 and is a particular case of what has been variously referred to in the literature as the split questionnaire design (for example Chipperfield and Steel, 2009; Peytchev and Peytcheva, 2017), matrix sampling (for example Gonzalez and Eltinge, 2008) and modular design (Peytchev et al., 2020).

For weighting, a modular design presents the opportunity to benefit from the bigger sample size of the larger master survey. One way to achieve this benefit is to first weight the master sample to the population using external sources and then weight the subsample responding to the EU-SILC module to the (weighted) master sample through calibration on common variables – a simple two-phase approach. An alternative is composite calibration (Merkouris, 2013). These two approaches are compared by Fallows (see Chapter 11 of this book), who notes the practical limitations around the number of (common) variables that can be estimated consistently between the master sample and the subsamples, and the detrimental effect this could have on the precision of module-specific variables (for example EU-SILC indicators). Only minor differences in estimates were found between the two methods.

Figure 13.2: Schematic diagram of a modular design with two modules



NB: All cases in the master wave are asked a core set of questions. The sample was then partitioned, with each part asked questions relating to a single module; EU-SILC would be one such module.

13.6. Recommendations

Some improvements to the EU-SILC guidelines would be simple to implement and have little or no resource implications.

- Change the scale factors by which panel-specific weights should be scaled, from $4/m$ (where m is the number of panels for which population members could have been selected) to $4n/mn_i$ (where n_i is the size of the panel for which the case was selected and n is the total size of the four panels being combined), to produce cross-sectional weights. See Section 13.5.1.
- Recommend a standardised simple approach to imputation of eligibility status. Current practice includes the use of extreme opposite assumptions in different countries, which is likely to damage between-country comparability of estimates. See Section 13.4.
- Recommend a regression-based approach to non-response modelling in preference to the homogeneous groups approach (while recognising that the two can be rather similar, depending on the methods used to identify the homogeneous groups).

Other possible improvements require further consideration, as the implications would be more considerable.

- Greater consistency between countries in the choice of auxiliary variables for non-response adjustment would seem desirable. Research – including that presented in Chapters 11 and 12 of this book – has shown that estimates are generally more sensitive to the choice of adjustment variables than the choice of adjustment method. It could be desirable to specify a minimum core set of variables for inclusion in attrition adjustments.
- It could be desirable to standardise the approach to non-response adjustment for an analysis that combines multiple panels, although further work would be required to identify the best approach for the multipurpose EU-SILC weights.
- A better method of estimation of eligibility status could be recommended, based on modelling using available covariates.

13.7. Conclusion

Current practice in weighting in EU-SILC includes all the main elements that would be expected (design weighting, non-response adjustments, combining panels, calibration) and represents best practice in some respects. Overall, the weighting procedures are of a good standard, but some elements of current procedures do not represent best practice and could be improved. In particular, for a comparative survey it is particularly important for weighting procedures to be comparable between countries. In some respects, such as the choice of auxiliary variables and the assumptions made about eligibility, there is a concerning level of variation between countries. Some of the potential improvements would be relatively easy to implement. These are set out in Section 13.6.

References

Brick, J. M. (2013), 'Unit nonresponse and weighting adjustments: a critical review', *Journal of Official Statistics*, Vol. 29, No 3, pp. 329–353.

Chipperfield, J. and Steel, D. (2009), 'Design and estimation for split questionnaire surveys', *Journal of Official Statistics*, Vol. 25, No 2, pp. 227–244.

Eurostat (2019), *Methodological Guidelines and Description of EU-SILC Target Variables – DocSILC065 – 2018 operation (version July 2019)* (https://circabc.europa.eu/sd/a/e9a5d1ad-f5c7-4b80-bdc9-1ce34ec828eb/DOCSILC065%20operation%202018_V5.pdf)

Gonzalez, J. M. and Eltinge, J. L. (2008), 'Adaptive matrix sampling for the Consumer Expenditure Quarterly Interview survey', in *Proceedings of the American Statistical Association, Survey Research Methods Section,* American Statistical Association, Alexandria, VA.

Kish, L. (1965), *Survey Sampling,* Wiley, New York.

Kish, L. (1992), 'Weighting for unequal P_i', *Journal of Official Statistics*, Vol. 8, No 2, pp. 183–200.

Lundström, S. and Särndal, C.-E. (1999), 'Calibration as a standard method for treatment of non-

response', *Journal of Official Statistics*, Vol. 15, No 2, pp. 305–327.

Lynn, P. (2005), 'Weighting', in Kempf-Leonard, K. (ed.), *Encyclopedia of Social Measurement*, Elsevier, Amsterdam.

Lynn, P. and Watson, N. (2021), 'Issues in weighting for longitudinal surveys', in Lynn, P. (ed.), *Advances in Longitudinal Survey Methodology*, Wiley, Chichester, pp. 447–468.

Merkouris, T. (2013), 'Composite calibration estimation integrating data from different surveys', in *Proceedings of the 59th ISI World Statistics Congress*, International Statistical Institute, The Hague, pp. 25-30.

Peytchev, A. and Peytcheva, E. (2017), 'Reduction of measurement error due to survey length', *Survey Research Methods*, Vol. 11, No 4, pp. 361–368.

Peytchev, A., Peytcheva, E., Conzelmann, J. G., Wilson, A. and Wine, J. (2020), 'Modular survey design: experimental manipulation of survey length and monetary incentive structure', *Journal of Survey Statistics and Methodology*, Vol. 8, No 2, pp. 370–384.

Särndal, C. and Lundström, S. (2006), *Estimation in Surveys with Nonresponse*, Wiley, Chichester.

14 Item non-response in EU-SILC income variables

Richard Heuberger ([74])

14.1. Introduction

This chapter discusses non-response in income variables in EU-SILC. In a perfect world, there would be no survey non-response – all the answers that researchers attempt to obtain would be available within the data set. However, in the real world research is required not only to detect and evaluate non-response after a survey is completed but also to identify methods to avoid – or at least minimise – non-response at the data collection stage (Lynn, 2008). Thus, this chapter aims to place the impact of item non-response in the context of the total survey error framework (Section 14.2). The chapter then identifies potential sources of item non-response (Section 14.3) as one step in an overall strategy to prevent non-response in income variables in EU-SILC.

In addition to identifying the sources of non-response, this chapter analyses item non-response in income variables in EU-SILC from a comparative perspective. To do so, the flag variables available for income variables are discussed (Section 14.4) and used in the analysis. The following section (Section 14.5) deals with empirical findings on non-response for income variables in EU-SILC. Here, the non-response is analysed using the user database (UDB) ([75]) data sets. Unfortunately, not all countries participating in EU-SILC provide data for research as UDB data. Since this particularly refers to Germany ([76]) and therefore a large share of the population of the EU, it impedes the calculation of European indicators.

The conclusions (Section 14.6) summarise the chapter in three perspectives. The first perspective focuses on the analysis of the level and nature of non-response in income variables in EU-SILC in recent years. The second perspective reflects experiences of other (income-related) surveys in other (non-European) countries with non-response. Finally, the third perspective concentrates on the implications for future EU-SILC operations and for imputation (Durrant, 2009) of income variables in EU-SILC. The main aim of these conclusions is to formulate suggestions about how to deal with item non-response from the perspective of data producers and from the perspective of the user of the data. Chapter 15 addresses imputation methods as one approach to cope with item non-response.

14.2. Item non-response and total survey error

Item non-response affects almost all social surveys. It is hard to identify any social survey with a reason-

([74]) Richard Heuberger works at Statistics Austria. The author would like to thank Peter Lynn, Eric Marlier, Matthias Till and participants of the Net-SILC3 International Best Practice Workshop in February 2019 for their very useful comments. All errors are the author's responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Richard Heuberger (richard.heuberger@statistik.gv.at).

([75]) The EU-SILC data are available in two formats: the production database, including all available variables for responding and non-responding households, and the UDB, which excludes non-responding units and variables that could potentially allow identification of households.

([76]) The data set provided to national statistical institutes does not include data for Germany.

able sample size that has no missing information. Item non-response denotes missing information within the set of answers of a response unit (individual or household, or a company). Unit non-response denotes the missingness of all answers of a response unit or the dropout of a response unit. Non-response results in the recorded sample being somewhat smaller than the originally selected sample.

For a better understanding of the causes and consequences of item non-response in income questions and the context of the imputation process, it may be useful to contextualise these issues within the so-called total survey error framework (see also Chapter 1):

> *Total survey error (TSE) refers to the accumulation of all errors that may arise in the design, collection, processing and analysis of survey data. In this context, a survey error can be defined as any error arising from the survey process that contributes to the deviation of an estimate from its true parameter value.* (Biemer, 2016, p. 122)

Within the theoretical or conceptual framework of total survey error, item non-response is one element among other sources of bias. As described by Groves et al. (2009), non-response error, together with coverage error, sampling error and measurement error, add up to the total survey error. Hence, the aim is to discuss item non-response in the context of or in relation to other errors in the EU-SILC survey. This highlights that we have to consider the different contexts and settings in which EU-SILC is conducted in different countries when thinking about item non-response.

14.3. Sources of item non-response

Item non-response can be discussed from various perspectives. One perspective is to differentiate between the sources from which data are obtained: the reasons for and the mechanisms leading to item non-response and the means by which item non-response can be avoided differ depending on whether the data come from a survey questionnaire or from administrative sources. In EU-SILC, item non-response in administrative data may occur through imperfections in the process of linking administrative data to survey interviews.

If data are obtained in interviews, the mode of the data collection plays a role in the item non-response process. In addition, item non-response can be discussed with regard to the actor causing the missingness, be it the respondent, the interviewer or the data editor.

Looking at item non-response from various perspectives may help to evaluate the processing of the data, to identify problems causing non-response, and therefore helps to reduce non-response.

For income variables collected during the survey through a questionnaire, a missing value can arise for any one of several reasons.

- **Reasons connected with the respondent.** The respondent is not willing to provide an answer, the respondent does not know the answer or the respondent does not give an answer for other reasons (e.g. simple error in the case of self-completion surveys).
- **Reasons connected with the interviewer.** The interviewer fails to ask a question, or the interviewer does not record an answer (although an answer was given) either intentionally or unintentionally (for example an operating error with a laptop).
- **Reasons connected with question or questionnaire design.** The answer does not fit into the format or categories provided for answers, the question is too cognitively demanding, or the question requests information unknown to the respondent.
- **Reasons connected with the handling of the data.** Missingness can be a result of problems with the questionnaire or processing of the data after data entry. The missingness in this case can also be produced by mistake (an error during the editing process) or be a result of a deliberate decision by the data editor (for example when values are suspected to be implausible).

These four types of reasons for item non-response are not independent; rather, they interact. For example, a respondent is more likely to choose not to answer a survey question if the question is poorly designed or the questionnaire is too long or burdensome.

For variables from administrative sources, missing values can arise due to:

- problems with the variable that links administrative data and survey data;
- problems with the variable within the administrative data set (missing entries, errors in processing the data, etc.);
- problems connected with editing the administrative data that have been matched to the survey data.

To link administrative data to survey data, the existence of a suitable matching variable(s) ([77]) is crucial, aside from other legal and technical requirements for using administrative data. This linking variable may be the social security number or another pseudonymised personal identifier number (PIN) used in the administration of the country in question. Linking necessitates that it is possible to link administrative data to survey data at an individual level. In most cases, the provider of the data has to ensure that the administrative data are identifiable using this linking variable. Missingness of this linking variable on the side of the administrative data is connected to the processes of the administrative data in the specific country and the specific data provider. For example, processes to find the correct identification number for an individual record may differ from provider to provider and depend on the legal situation of the country ([78]).

If a link to administrative data is not achieved for a survey record that includes a PIN, it is not possible to determine whether a link was not possible because the PIN was missing in the administrative data, or because the person is not included in the administrative records. In other words, it is not known in this case whether a missing value is correctly absent or missing due to a missing PIN.

Missingness can be specifically connected to the mode of the interview. It could be related to technical problems, for example with the programming of the interview software, or problems connected with transferring the data after the interview. Technical problems may also arise when paper and pencil interviews are used, for example problems connected with scanning questionnaires for further electronic processing. Computer-assisted modes also enable integration of routing in the questionnaire and checks that can reduce item non-response. These features are not available for paper modes. The presence of an interviewer may also influence the occurrence of item non-response – for example the interviewer may assist the respondent in giving a correct answer, or the presence of the interviewer may impede the respondent from providing an answer (which is more likely for sensitive questions). The processing of the data either after the interview or after linking administrative data may produce missingness in income variables, either unintentionally or intentionally.

With regard to income variables, checks for extreme values may produce missingness. This depends on whether extreme values are replaced by the threshold determining the extreme cases (cut-off point) or kept missing until the imputation process. The reason for controlling extreme values can be either that these extreme values are regarded as wrong or somehow implausible, or that these extreme values would allow the identification of individuals.

In some cases, information provided in the interview ([79]) indicates that an income component is missing (albeit stated as not received in the questionnaire) or that the given income information has to be considered erroneous. The data editing process entails having to solve these contradictions

([77]) Dependent on whether it is a statistical matching process (mapping on the basis of probability) or a distinct matching process (distinct mapping of data records for individual units).

([78]) For the use of administrative data in EU-SILC, see Jäntti, Törmälehto and Marlier (2013).

([79]) In EU-SILC in Austria, for each question the respondent, or rather the interviewer, can give additional information on the given answer (for example additional information about the structure of the received income or details about the exact nomenclature of the received income).

and determine which information in the questionnaire is considered more plausible. Data editing may sometimes involve having to determine an income component as missing. In EU-SILC, information from prior waves may also cast doubt on the correctness of an answer in the current year. Here, the editor of the data has to judge whether the current information or the information from prior waves is correct.

At the limits of questions of item non-response are cases in which it is not possible to record the correct answer: 'Information on an item may be incomplete simply because *it is not feasible to seek it exactly or in full detail* in an interview survey; these errors are akin to "conceptual errors"' (Atkinson and Marlier, 2010, p. 61, italic in the original text).

14.4. Construction of flag variables

Flag variables are used to record information about the characteristics and qualities of the information recorded in a variable. The information provided in these flag variables can be used to better understand item non-response. As there are different types of income variables in EU-SILC, there are also different types of flag variables for the income variables. Flag variables are constructed by the national statistical institutes. These data are referred to as the production database (PDB) data. Data provided to (scientific) users are provided with slightly different variables ([80]); these differences mainly concern the anonymisation of the data. However, the flag variables are also different in the UDB data. The information provided in the flag variables of the PDB data is somewhat reduced and provided in two variables in the UDB data.

Construction of the income flag variables in the PDB data is described in the *Methodological Guidelines and Description of EU-SILC Target Variables* (Eurostat, 2016). Separate flag variables are calculated for the total household income, gross incomes and net incomes. Generally, the income flag is the concatenation of different digits referring to different elements of the survey design and the imputation:

- first digit:
  - information on the type of collection – whether the value has been collected net or gross,
  - the type of the net/gross recorded value;
- second digit:
  - information on the type of imputation method applied;
- third digit:
  - information on the imputation factor – this is the proportion of the value that is not imputed.

The type of net recorded value is recorded only in the flag variable for each net income variable – for these flag variables, this information is recorded in an additional digit (added to the first two digits). Thus, the flag variable of each net income variable consists of four digits. For the total household income flag variable and the gross income flag variable, only the digits for the level of collection (net or gross), the imputation method and the imputation factor are used. For the three different types of flag variables, the codes for the level of collection differ; the codes for the imputation method are the same for all three flag variable types. Table 14.1 provides an overview.

([80]) The differences between PDB data and UDB data are described in two documents provided by Eurostat on the platform of the Communication and Information Resource Centre for Administrations, Businesses and Citizens. One document describes the cross-sectional data (https://ec.europa.eu/eurostat/documents/203647/203704/C_differences) and the other the longitudinal data (https://ec.europa.eu/eurostat/documents/203647/203704/L-differences).

Table 14.1: Digits of income flags in EU-SILC (as defined for the PDB)

Indicator	Total household income	Gross income	Net income
Collected net or gross (first digit)	1 – Net 2 – Gross 3 – Net and gross 4 – Unknown	1 – Net of tax on income at source and SICs 2 – Net of tax on income at source 3 – Net of SICs 4 – Gross 5 – Unknown 6 – Mix	1 – Net of tax on income at source and SICs 2 – Net of tax on income at source 3 – Net of SICs 4 – Gross 5 – Unknown
Imputation method (second digit)	0 – No imputation 1 – Deductive imputation 2 – Statistical imputation 3 – Gross/net conversion	0 – No imputation 1 – Deductive imputation 2 – Statistical imputation 3 – Gross/net conversion	0 – No imputation 1 – Deductive imputation 2 – Statistical imputation 3 – Gross/net conversion
Type of net recorded value			1 – Net of tax on income at source and SICs 2 – Net of tax on income at source 3 – Net of SICs 5 – Unknown 6 – Mix
Imputation factor (third digit)	Collected/recorded value	Collected/recorded value	Collected/recorded value

NB: SIC denotes social insurance contributions.

The imputation method differentiates between no imputation, deductive imputation, statistical imputation and gross/net conversions. When more than one imputation method is applied for one variable (which can be the case if the income target variable is an aggregate), it is unclear which imputation method should be specified. It seems reasonable that the imputation method for the component with the largest contribution should be given.

The imputation factor gives the share of the non-imputed information of a value and is calculated as the collected value divided by the recorded value, where collected refers to the non-imputed value from the survey or the administrative records and recorded refers to the final value after imputation. If the collected value is identical to the recorded value, the imputation factor equals 100; if the value of the income target variable is fully imputed, the imputation factor equals zero. The imputation factor cannot then be correctly calculated if positive and negative values are summed up in one variable (for example if family-related benefits consist of both repayments and benefits). However, the imputation factor provides quantitative information on the extent of imputed information. It may be assumed that a higher share of non-imputed information indicates better data quality. However, this is not really the case, since this would imply that the share of the imputed information should be kept low regardless of the extent of item non-response. Even in cases in which the income variable is fully imputed, assessment of the imputation process cannot be based on this information alone ([81]).

Most income variables in EU-SILC consist of more than one income component. Thus, it may be difficult to mirror different practices for different income components within one income flag.

The flag variables provide some insight into the non-response process and the imputation procedures for income variables. However, this insight is limited. These limitations are due to the imperative of comparability (of variables) in EU-SILC: country-specific information on non-response and imputation cannot be fully provided within this format of the flag variables.

([81]) See Chapter 15 on imputation procedures for information on imputation in the quality reports.

14.5. Analysis of non-response using user database flag variables

In this section, the extent of item non-response in EU-SILC income variables is discussed. The aim is to examine differences between countries and their use of the UDB flag variables. The income flag variable information in the PDB data set is separated into two different variables (for each income variable) in the UDB files: one includes information on the level of collection (net or gross, the first digit – this variable is identifiable by the suffix '_F') and the other includes information on the imputation procedure and the imputation factor (identifiable by the suffix '_I'). For this exercise, the 27 countries for which data are available in the data set of the 2015 EU-SILC are included. The countries that are missing in the data set despite participating in EU-SILC are Germany, Iceland, Ireland, Macedonia, Switzerland and Turkey.

The main challenge of working with the income flag variables of EU-SILC is that different countries interpret the specifications differently. For example, in Austria the value 2100 denotes a non-imputed value in the variable HY020_I, in Belgium a non-imputed value is denoted by 100, and in the Slovakian data set a non-imputed value is denoted by 1. The logic behind the interpretation of how to fill flag variables is possibly justified (and consistent within one country) but makes cross-country comparisons difficult.

Table 14.2 presents the percentage of households without income information and the percentage of households with no imputed income by country. Households with no household income (HY010 = 0 or HY020 = 0) are not an indication of an error in the survey in every case. It is possible for a household to not have any income in that particular year and solely meet their expenses using assets or savings. The share of households with no income information at all is generally low: it is below 1 % in all participating countries providing UDB data. The share of households with no net household income (HY020) is, in most of the countries, smaller than the share of households with no gross income in HY010 – if there are any differences at all between these shares.

In contrast, the share of households with no imputed income information differs greatly between countries – from below 1 % to 100 % (for the variable HY010). A share of 100 % means that there was no item non-response to income questions in these countries, and therefore no imputation was necessary for the calculation of household income or any income component contributing to this household income. Some countries using administrative data for the calculation of the household income fall into this category (Denmark, the Netherlands and Norway), but this group also includes, for example, Czechia, Hungary and Portugal. It is noteworthy that, in these countries, even for income components such as private transfers between households, either there is no missing information from respondents or there is sufficient information from a register. For traditional register countries such as Denmark, Finland and Norway, these variables (HY080, HY130) are zero.

Some countries, such as Belgium, Bulgaria, France and Italy (but also Croatia, Latvia and Luxembourg), exhibit a comparatively low share of households with no imputed income information. For these countries particularly, a look at the income components is enlightening. In countries where the share of households with no imputed information for HY010 is lower than the share of households with no imputed information for HY020 (or vice versa), the difference most probably stems not from statistical imputation but from gross–net/net–gross imputation. This is the case, for example, in Estonia, France, Italy, Hungary and Slovenia. For Romania, the flag variables for household income (HY010 and HY020) were missing.

The discussion of non-response here deals with total household income, an aggregation of many specific income (sub)components. Below, the item non-response of all other income variables of EU-SILC is discussed. The focus will be on income variables that contribute to household income. The question is whether there are any country-specific trends in item non-response in income variables or any variables suffering particularly from non-response. This information should be included in the quality report, but only some countries address this topic. For example, the Slovakian report states:

Table 14.2: Percentage of households without income information and without imputed income

Country	Number of households	% of households with zero income (HY010)	% of households with no imputed income (HY010)	% of households with zero income (HY020)	% of households with no imputed income (HY020)
Austria	6 045	0.0	72.0	0.0	89.7
Belgium	6 006	0.1	46.5	0.0	17.3
Bulgaria	4 965	0.2	16.3	0.2	15.3
Croatia	6 562	0.6	31.0	0.6	4.7
Cyprus	4 357	0.0	99.5	0.0	99.7
Czechia	7 914	0.0	100.0	0.0	100.0
Denmark	6 025	0.0	100.0	0.0	100.0
Estonia	5 728	0.4	23.9	0.3	82.3
Finland	10 726	0.0	93.4	0.0	93.8
France	11 390	0.0	0.3	0.0	22.5
Greece	14 096	0.5	99.5	0.3	99.7
Hungary	7 770	0.2	99.8	0.2	79.5
Italy	17 985	0.9	9.7	0.7	41.5
Latvia	6 113	0.2	31.6	0.1	9.6
Lithuania	4 849	0.7	97.4	0.6	96.0
Luxembourg	3 474	0.3	38.3	0.2	0.5
Malta	4 233	0.0	85.3	0.0	85.3
Netherlands	9 806	0.0	100.0	0.0	100.0
Norway	6 393	0.0	100.0	0.0	100.0
Poland	12 183	0.1	18.9	0.0	56.7
Portugal	8 740	0.0	100.0	0.0	100.0
Romania	7 415	0.0	NA	0.0	NA
Slovakia	5 637	0.2	99.8	0.2	99.8
Slovenia	8 685	0.0	54.8	0.0	44.8
Spain	12 367	0.3	83.7	0.2	87.9
Sweden	5 859	0.2	99.8	0.1	99.9
United Kingdom	9 312	0.2	37.9	0.0	37.2

NB: NA, no information available.

Source: EU-SILC UDB, 2015.

> *Regarding data on income obtained during interviews, household members have the tendency to underestimate individual sources of income or data on some income components is missing (item non-response). The elimination possibilities of this survey data underestimation are limited. In the presented survey, only such adjustments were done, where there was sufficiently reliable external statistical source or which can be based on the legislation.* (Statistical Office of the Slovak Republic, 2017)

The focus will also be on two income components in particular: family-related benefits (HY050N/G) in Table 14.3 and Table 14.4, and employment income (PY010N/G) in Table 14.5. These components were selected mainly because of their significance and importance for household income.

Overall, the approach to fill these (flag) variables seems to be quite different between countries. Croatia coded cases with no imputed values with zero; the Netherlands coded non-imputed values in the gross variable with the code '1 000.00', and Norway used the code '0'. Romania provided no information on the imputation of HY050N and HY050G. Sweden coded the variables HY050N_I and HY050G_I with the code '31' for all cases –

Table 14.3: Flag variables of family-/children-related allowances, HY050N_F and HY050G_F

	HY050N_F (1st and 2nd digits)							HY050G_F (2nd digit)					
	00 – No Income	11 – Net of income tax and SICs	21 – Collected net of income tax; net of tax and SICs	22 – Net of income tax	31 – Collected net of SICs; net of income tax and SICs	33 – Net of SICs	41 – Collected gross; net of income tax and SICs	0 – No income	1 – Net of income tax and SICs	2 – Net of income tax	3 – Net of SICs	5 – Gross	Total – number of households
Austria	4 232	1 813	0	0	0	0	0	4 232	1 813	0	0	0	6 045
Belgium	3 954	2 052	0	0	0	0	0	3 954	0	0	0	2 052	6 006
Bulgaria	4 121	844	0	0	0	0	0	4 121	0	0	0	844	4 965
Croatia	5 750	812	0	0	0	0	0	5 750	0	0	0	812	6 562
Cyprus	3 372	985	0	0	0	0	0	3 372	0	0	0	985	4 357
Czechia	7 161	753	0	0	0	0	0	7 161	0	0	0	753	7 914
Denmark	0	0	0	0	0	0	0	4 249	0	0	0	1 776	6 025
Estonia	3 875	0	1 853	0	0	0	0	3 875	0	0	0	1 853	5 728
Finland	0	0	0	0	0	0	0	7 510	0	0	0	3 216	10 726
France	8 366	0	0	0	0	3 024	0	8 366	0	0	3 024	0	11 390
Greece	12 659	1 437	0	0	0	0	0	12 659	1 437	0	0	0	14 096
Hungary	5 518	0	0	0	0	2 252	0	5 518	0	0	0	2 252	7 770
Italy	14 020	3 965	0	0	0	0	0	14 020	3 965	0	0	0	17 985
Latvia	4 596	1 517	0	0	0	0	0	4 596	0	0	0	1 517	6 113
Lithuania	4 527	0	0	0	0	0	322	4 527	0	0	0	322	4 849
Luxembourg	2 330	1 144	0	0	0	0	0	2 330	0	0	0	1 144	3 474
Malta	0	0	0	0	0	0	0	3 054	0	0	0	1 179	4 233
Netherlands	0	0	0	0	0	0	0	6 826	0	0	0	2 980	9 806
Norway	0	0	0	0	0	0	0	4 168	0	0	0	2 225	6 393
Poland	10 655	1 528	0	0	0	0	0	10 655	1 528	0	0	0	12 183
Portugal	7 418	0	0	1 322	0	0	0	7 418	0	1 322	0	0	8 740
Romania	5 647	1 768	0	0	0	0	0	5 647	1 768	0	0	0	7 415
Slovakia	0	0	0	0	0	0	0	3 488	0	0	0	2 149	5 637
Slovenia	6 482	2 203	0	0	0	0	0	6 452	0	0	0	2 233	8 685
Spain	12 023	344	0	0	0	0	0	12 023	32	0	0	312	12 367
Sweden	4 169	0	0	0	1 690	0	0	4 169	0	0	1 690	0	5 859
United Kingdom	0	0	0	0	0	0	0	6 753	0	0	0	2 559	9 312
Total	130 875	21 .165	1 853	1 322	1 690	5 276	322	166 893	10 543	1 322	4 714	31 163	214 635

NB: In Denmark, Finland, Malta, the Netherlands, Norway, Slovakia and the United Kingdom, only the gross variable is filled. NA, no information available. SIC, social insurance contributions.

Source: EU-SILC UDB, 2015.

which was perhaps a mistake with regard to the flag variables. In Slovenia, the variable HY050N_I was filled only for 2 203 cases, 29 cases fewer than for the gross variable.

Overall, countries seem to fill flag variables and the imputation flag variables quite individually. Comparative analysis is consequently cumbersome and time-consuming. In some cases, the meaning of what is included in the files is quite unclear. A better shared understanding and shared practices would make it easier to work with these variables.

If a gross value is available but not a net value, or vice versa, one can assume that countries use net–gross or gross–net conversion methods to calculate the missing corresponding value. In more than 3 500 cases, the net value is higher than the gross value – which may be a sign of an error in the conversion, in the data editing or in filling the variables. Several countries have cases with higher net values than gross values. For some countries, only net values are filled, and the corresponding gross variable is filled with '– 5', which is correct and comprehensible.

Table 14.4: Flag variables of family-/children-related allowances, HY050N_I and HY050G_I

	Total households	Filled	HY050N_I		HY050G_I	
			Not imputed	% not imputed	Not imputed	% not imputed
Austria	6 045	1 813	1 807	99.7	1 807	99.7
Belgium	6 006	2 052	1 942	94.6	1 873	91.3
Bulgaria	4 965	844	717	85.0	712	84.4
Croatia	6 563	812	762	93.8	762	93.8
Cyprus	4 357	985	985	100.0	985	100.0
Czechia	7 914	753	753	100.0	753	100.0
Denmark	6 025	4 249	NA	NA	NA	NA
Estonia	5 728	1 853	1 816	98.0	1 636	88.3
Finland	10 726	7 510	NA	NA	NA	NA
France	11 390	3 024	524	17.3	26	0.9
Greece	14 096	1 437	1 437	100.0	1 437	100.0
Hungary	7 770	2 252	1 981	88.0	655	29.1
Italy	17 985	3 965	3 667	92.5	3 667	92.5
Latvia	6 113	1 517	1 445	95.3	1 445	95.3
Lithuania	4 849	322	190	59.0	313	97.2
Luxembourg	3 474	1 144	926	80.9	891	77.9
Malta	4 233	1 179	NA	NA	1 139	96.6
Netherlands	9 806	2 980	NA	NA	2 980	100.0
Norway	6 393	2 225	NA	NA	2 225	100.0
Poland	12 183	1 528	1 383	90.5	1 258	82.3
Portugal	8 740	1 322	1 322	100.0	1 322	100.0
Romania	7 415	1 768	NA	NA	NA	NA
Slovakia	5 637	2 149	NA	NA	2 149	100.0
Slovenia	8 685	2 233	2 203	98.7	2 232	100.0
Spain	12 023	344	330	95.9	312	90.7
Sweden	5 859	1 690	1 690	100.0	1 690	100.0
United Kingdom	9 312	2 559	NA	NA	1 859	72.6

NB: NA, no information available.

Source: EU-SILC UDB, 2015.

Table 14.5: Flag variables of employee cash or near-cash income, PY010N_F

		PY010N_F											
	Total households	0	– 5	– 3	– 2	– 1	1	11	22	31	33	41	61
Austria	10 935	4 557	0	0	1 801	1	14	4 562	0	0	0	0	0
Belgium	11 364	5 978	0	0	1 433	0	72	3 881	0	0	0	0	0
Bulgaria	10 402	5 097	0	0	1 547	0	77	3 681	0	0	0	0	0
Croatia	14 840	9 493	0	0	1 687	15	40	3 273	0	0	0	330	2
Cyprus	9 984	0	8 720	0	1 218	0	41	5	0	0	0	0	0
Czechia	15 139	7 534	0	0	3 039	0	49	2 698	0	0	0	1 819	0
Denmark	11 708	0	10 817	0	849	0	42	0	0	0	0	0	0
Estonia	12 054	4 761	0	0	1 872	0	252	5 046	0	0	0	123	0
Finland	21 189	0	17 877	1 603	1 632	0	77	0	0	0	0	0	0
France	21 292	9 159	0	0	2 079	0	117	0	0	0	9 937	0	0
Greece	29 405	21 370	0	0	1 955	0	266	5 814	0	0	0	0	0
Hungary	15 689	7 994	0	0	2 293	0	144	0	5 258	0	0	0	0
Italy	36 602	20 001	0	0	5 364	3	575	10 659	0	0	0	0	0
Latvia	11 726	5 000	0	0	1 407	0	261	5 058	0	0	0	0	0
Lithuania	9 783	4 771	0	0	1 363	0	69	3 443	0	0	0	137	0
Luxembourg	7 204	3 027	0	0	899	0	30	3 248	0	0	0	0	0
Malta	9 557	0	8 284	0	1 254	0	19	0	0	0	0	0	0
Netherlands	18 597	0	17 084	0	1 509	0	4	0	0	0	0	0	0
Norway	12 662	0	11 586	0	1 025	0	51	0	0	0	0	0	0
Poland	27 997	16 056	0	0	3 442	0	321	8 178	0	0	0	0	0
Portugal	18 702	10 504	0	0	2 049	0	430	4 401	0	0	0	1 318	0
Romania	15 730	10 502	0	0	1 911	0	169	3 148	0	0	0	0	0
Slovakia	13 769	0	11 411	0	2 257	0	101	0	0	0	0	0	0
Slovenia	22 366	9 330	0	2 457	1 484	0	14	9 081	0	0	0	0	0
Spain	27 215	13 960	0	0	3 423	0	195	9 637	0	0	0	0	0
Sweden	11 398	3 347	0	0	970	1	55	0	0	7 025	0	0	0
United Kingdom	16 717	0	14 115	0	2 489	0	113	0	0	0	0	0	0
Total	444 026	172 441	99 894	4 060	52 251	20	3 598	85 813	5 258	7 025	9 937	3 727	2

NB: No information available.

Source: EU-SILC UDB, 2015.

In some countries (the Netherlands and Norway), the income variables and the imputation flag variables do not correspond; it seems that the imputation flag variables include more cases. For Romania, it was not possible on the basis of the flag variables to determine how many cases were available without imputation.

It is often hard to understand what the imputation flag indicates. Assumptions must be made about which value(s) of the imputation flag variables denotes (no) imputations. However, the share of imputations clearly differs significantly between countries. Most countries employed net–gross or gross–net conversion methods or statistical imputation procedures to impute values. The statistical imputation methods featured varied from use of standard statistical packages such as imputation and variance estimation software (IVEware) to individual approaches using linear regression models.

The key problem in using the flag variables in EU-SILC to understand non-response is that a shared understanding about how to fill these flag variables is missing.

14.6. Conclusions

14.6.1. The level of non-response and structural effects

In general, the level of non-response is not particularly high for income variables in EU-SILC. However, what is of more concern is whether or not the non-response is selective.

The first apparent structural or systematic effect is the difference between countries that use register data for the income target variables and countries that use survey questions: 'register countries' generally feature a lower level of non-response than non-register countries. In every case, information on non-response is available at an aggregate level, since income target variables consist of (in most cases) more than one variable. The information on non-response is a composite index and does not identify which income component is affected by non-response.

The aim should be to develop a shared understanding of how to fill the flag variables. As it is, countries follow different types of logic in filling these flag variables. These different types of logic seem to be understandable and, by and large, consistent within one country, but they make the analysis and interpretation of flag variables difficult. A common understanding can be facilitated by projects that make use of these variables and thereby prove their importance.

Is the level of item non-response for income questions in EU-SILC problematic? Overall, one would be tempted to answer this in the negative: the levels of item non-response are not dramatically high (as seen in the tables and as documented in the quality reports ([82])); quality reports and analysis do not indicate significant biases in the distribution of income variables. Nevertheless, it is possible that there may be some questions/variables in some countries in some years where the level of non-response is unignorable. What is of more interest here, in the context of a comparative project, is the difference in the level of non-response between countries – which leads (again) to the discussion of comparability between different data sources (register data or other) and different data collection practices in EU-SILC.

There is a trade-off between item non-response and other errors: attempts to reduce unit response could result in adding more respondents of the kind who tend to produce item non-response to the sample. Furthermore, efforts to lower the level of item non-response (more checks within the interview, more specific questions with detailed explanations) may make the interview more burdensome and lead to a higher level of unit non-response (but maybe not until the following waves of the panel). Pushing respondents for answers may also lead to measurement error, as respondents are urged to give any (possibly guessed or otherwise incorrect) answers.

To provide more information on the selectivity of non-response, quality reporting should be extended, for example by adding more and detailed income comparisons with externals sources (ideally

([82]) The European Statistical System standard quality report structure includes a table on item non-response of income variables.

administrative resources). These comparisons are included in some quality reports (e.g. Austria, Slovakia and Switzerland) but not in all.

14.6.2. Experiences from other surveys

All (or nearly all) surveys in the social sciences suffer from item non-response in one way or the other. The mechanisms that produce item non-response are common to all surveys (refusal, ignorance of the answer, unintended skip, corrections during the data editing, etc.). An element that differs when discussing the question of non-response is the repertoire used to avoid non-response. The American Survey of Income and Program Participation (SIPP), for example, uses dependent interviewing: 'Dependent interviewing is the process in which information from a previous interview(s) carries forward into the current survey instrument in order to streamline the interviewing process and maximize data quality' (US Census Bureau, 2016, p. 8). The US Census Bureau argues that this is particularly important for a panel survey:

> *Additionally, dependent data has been shown to mitigate the negative effects of seam bias on data quality. Seam bias is a common ailment of longitudinal surveys, where event changes are reported disproportionately at the "seam" between waves. The key to alleviating seam bias is to create overlapping periods where one wave's interview period includes a portion of the next year's reference period. Through dependent interviewing, the instrument already possesses data for the early part of the next wave's reference period. With these data from the previous wave, the Computer Assisted Personal Interviewing (CAPI) instrument can tailor question wording to remind respondents of their situation during the previous wave. Therefore, the recall window shrinks and respondents are less apt to report changes at the transition between two reference periods.* (US Census Bureau, 2016, p. 9)

Proactive dependent interviewing may reduce non-response, since the burden of response is reduced. One danger connected with dependent interviewing is that the burden is reduced in a way that respondents provide an answer by approving the suggested answer. Thus, dependent interviewing may lead to undetected changes in the living conditions of participating individuals and households or even agreement with incorrect answers (see Eggs and Jäckle, 2015); however, this does not appear to happen often (Lynn et al., 2012). Proactive dependent interviewing could improve data quality for second and subsequent EU-SILC interviews, although it could not be used for households and individuals in the first wave – this is about one quarter of the sample in most of the countries.

14.6.3. What is to be done?

This chapter provides an overview of non-response in income target variables in EU-SILC. In short, it discusses three differences related to non-response between countries:

1. differences in the level of non-response dependent on the data collection practices,
2. differences in the filling of flag variables,
3. differences in the available information on non-response in the quality reporting.

Differences in data-recording practices could be remedied if countries employed the same data-recording practices with regard to interview modes, the use of register data, the handling of contacting practices, incentives, and so on. This is rather unrealistic and would be contrary to the output harmonisation approach of EU-SILC. However, further efforts towards input harmonisation and a broader understanding of common practices may in the long run lead to a lower level of non-response and more consistency between countries. Any steps towards standardisation may increase comparability of the data as well as understanding of item non-response. Documentation of the item non-response in the quality reports could also be improved, for example by adding (standardised) comparisons with external sources to identify systematic errors.

The analysis of item non-response of income target variables is not easy, since practices of filling the variables documenting the item non-response – the flag variables – are not comparable between countries. This challenge when working with the flag variables has informed plans to overhaul these

flag variables in the new framework of EU-SILC. A suggestion was presented at the meeting of the task force on the legal revision of EU-SILC in October 2017. As suggested above, further work on EU-SILC item non-response and imputation practices could foster awareness of the importance of these issues and the importance of a common understanding.

References

Atkinson, A. B. and Marlier, E. (eds) (2010), *Income and Living Conditions in Europe*, Publications Office of the European Union, Luxembourg.

Biemer, P. P. (2016), 'Total survey error paradigm: theory and practice', in Wolf, C., Joye, D., Smith, T. W. and Fu, Y.-C. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London, pp. 122–141.

Durrant, G. B. (2009), 'Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates', I*nternational Journal of Social Research Methodology*, Vol. 12, pp. 293–304.

Eggs, J. and Jäckle, A. (2015), 'Dependent interviewing and sub-optimal responding', *Survey Research Methods*, Vol. 9, No 1, pp. 15–29.

Eurostat (2016), *Methodological Guidelines and Description of EU-SILC Target Variables – 2015 operation (version March 2016)*, Eurostat, Luxembourg.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009), *Survey Methodology*, 2nd edition, Wiley, Hoboken, NJ.

Jäntti, M., Törmälehto, V.-M. and Marlier, E. (eds) (2013), *The use of registers in the context of EU-SILC: Challenges and opportunities*, Publications Office of the European Union, Luxembourg.

Lynn, P. (2008), 'The problem of nonresponse', in de Leeuw, E. D., Hox, J. and Dillman, D. (eds), *International Handbook of Survey Methodology,* Taylor & Francis, Hove, pp. 35–55.

Lynn, P., Jäckle, A., Jenkins, S. P. and Sala, E. (2012), 'The impact of questioning method on measurement error in panel survey measures of benefit receipt: evidence from a validation study', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 175, No 1, pp. 289–308, doi:10.1111/j.1467-985X.2011.00717.x.

Statistical Office of the Slovak Republic (2017), *National reference metadata in ESS Standard for Quality Reports Structure (ESQRSSI), 2015 data* (https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/a7216fba-4b8d-4f8a-97d0-c9f56a2d8e5a/details).

US Census Bureau (2016), *Survey of Income and Program Participation – 2014 panel user's guide*, 1st edition, US Census Bureau, Washington DC.

15 Imputation for income variables in EU-SILC

Sophie Psihoda, Nadja Lendle, Richard Heuberger and Thomas Glaser ([83])

15.1. Introduction

This chapter deals with the question of imputation practices in EU-SILC. EU Member States are legally obliged to apply appropriate imputation strategies in cases of item non-response. As quality criteria, two aspects are mentioned in the EU-SILC regulation on imputation ([84]). First, the procedure applied to the data should preserve the variation of and correlation between variables. Methods that incorporate 'error components' into the imputed values should be used, rather than those that simply impute a predicted value. Second, methods that take into account the correlation structure (or other characteristics of the joint distribution of the variables) should be used, rather than the marginal or univariate approach.

The regulation does not refer to the distribution of the imputed variable and does not prescribe a certain method or family of imputation methods. Thus, Member States may choose from a wide range of methods. Recommended methods are described in other documents, for example the *Methodological Guidelines and Description of EU-SILC Target Variables* (Eurostat, 2016a).

Building on Chapter 14, which deals with item non-response, this chapter describes the different missing data mechanisms and imputation techniques. It then lays out which methods are used in EU-SILC in different countries and illustrates different outcomes of the income distribution by method and country clustering. This is done by applying simulation techniques to 2016 EU-SILC data. Different approaches may reflect country-specific differences in the sources of non-response. However, there is also an apparent need for harmonisation of methodologies to ensure comparability.

15.2. Theoretical considerations

15.2.1. Reasons for imputation

For EU-SILC, there are two main reasons why imputing data is considered the only feasible solution for dealing with incomplete cases.

First, using only complete cases would have a drastic effect on the number of data records that are available for analysis. Income questions in particular are affected by non-response behaviour and therefore show a lot of missing information. In some countries, the share of complete cases of household income in the 2016 EU-SILC amounts to only 30 %. Deleting all cases with incomplete income information would considerably shrink the

([83]) Sophie Psihoda, Thomas Glaser and Richard Heuberger work at Statistics Austria, and Nadja Lendle is with the Institute of Applied Statistics, Johannes Kepler University Linz, Austria . We would like to thank Peter Lynn, Eric Marlier, Matthias Till, Marlene Blüher and participants of the Net-SILC3 International Best Practice Workshop in February 2019 for their very useful comments. All errors are the authors' responsibility. This work was supported by the Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Richard Heuberger (richard.heuberger@statistik.gv.at).

([84]) Commission Regulation (EC) No 1981/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the fieldwork aspects and the imputation procedures, OJ L 298, 17.11.2003, p. 23.

sample size of data to be analysed and also lead to a tremendous loss of information recorded in other income components that are not missing. The very low proportion of complete cases is related to the fact that household income is calculated from many different income sources (income from employment, self-employed income, social transfers, etc.). This provides a more accurate picture of the composition of income. However, the high number of variables involved also increases the chance of missing information in one of the variables and therefore of incomplete observations of the household income.

Second, non-response to income questions is usually not completely random. The number of income components a person receives can itself be predictive of the propensity for incomplete income information. Response behaviour may reflect the fact that people earning very little or a lot may not feel comfortable in providing (complete) information about their financial status. Hence, using complete cases potentially excludes only particularly low-income and particularly high-income households. Finally, non-randomness may also be related to the receipt of specific income components. For example, social desirability may result in missing information on capital incomes or social transfers.

As touched upon in the last paragraph, there are distinct patterns of missing data. Missing completely at random (MCAR) means that the missingness of cases is completely random and that the probability of an observation being missing does not depend on observed and unobserved measurements. Missing at random (MAR) describes a situation in which the missingness process depends on observed data and the probability of surveying an item does not directly depend on that item. Given observed data for all cases, the value of a missing item can be predicted. In other words, between different combinations of variables known for all cases the probability of an observation varies. However, for the same set of these characteristics known prior to the observation, the probability is assumed to be equal. Missing not at random (MNAR) describes a systematic pattern of missingness in which the reason for observations being missing depends on the missing values themselves.

15.2.2. Different imputation techniques

A possible way of categorising imputation techniques is by the mechanism producing the imputed value (Little and Rubin, 1987). On the one hand, deductive imputation utilises specific rules to impute a missing value. On the other hand, statistical imputations are based on a random process or stochastic model. In the work presented here only the latter method is used.

The main distinction between different imputation techniques presented here is focused on single and repeated imputation. Single imputation replaces a missing value with one exact estimate. In repeated imputation (Rubin, 2004), for every missing value more than one value is estimated, mostly by using simulations of one or more distributions. The resulting value for the case is then calculated on the basis of these values. Guided by this distinction, there are numerous statistical imputation methods or families of methods, which will be explained briefly in Sections 15.2.3–15.2.5.

15.2.3. Single imputation methods

Deductive and mean imputation

The simplest single imputation methods are deductive methods and mean imputation. If a deductive method is applied, the missing value is imputed by using logical relations between variables. The value for the missing item is then derived with high probability. Mean imputation is a method in which the missing value of a certain variable is replaced by the mean of the available cases. This method is easy to use; however, it has serious drawbacks, such as the reduction of variability in the data and ignoring of variables that are correlated with the missing values and may help estimate a plausible imputed value. Furthermore, it also does not incorporate characteristics that may explain the missingness mechanism and assumes that a completely random process causes missing values (MCAR).

Regression imputation

In regression imputation, the imputed value is predicted from a regression equation. For this method, the information in the complete observations is used to predict the values of the missing observations. Regression assumes that the imputed values fall directly on a regression line with a non-zero slope, so it implies a correlation between the predictors and the missing outcome variable. In terms of income, it will produce a more even distribution with fewer outliers with very high or very low income. After imputation, the distribution is more even and leads to an underestimation of the variance of the variable of interest (e.g. income from employment) as well as an overestimation of the strength of association between dependent variables and predictor variables (e.g. between income and household size). Regression imputation and all further methods presented that make use of variables available for all observed data assume a missing value mechanism that can be explained by given data (MAR). This means that the propensity of a missing value is conditional not on the missing value itself but on other variables in the data set (and thus can be estimated on the basis of these variables).

Donor imputation methods: hot-deck and nearest-neighbour imputation

In donor imputation methods, the missing value of a variable is replaced with an observed value from a respondent who is similar to the non-responding case with regard to characteristics observed for both cases. This rationale can also be applied to impute a set of more than one variable simultaneously in order to maintain the multivariate distribution of these variables. In hot-deck imputation, the donor case is taken from the same data set (in contrast to cold-deck imputation). Donor imputation methods differ with respect to how similarity between cases is defined and how the donor case is determined. If only a single donor case is identified as the most similar case, mostly on the basis of a function defining this similarity, the methods are called deterministic hot-deck methods. If the donor is selected by using a distance function that minimises the specified distance between donor and recipient, the method is called nearest-neighbour imputation.

Predictive mean matching imputation

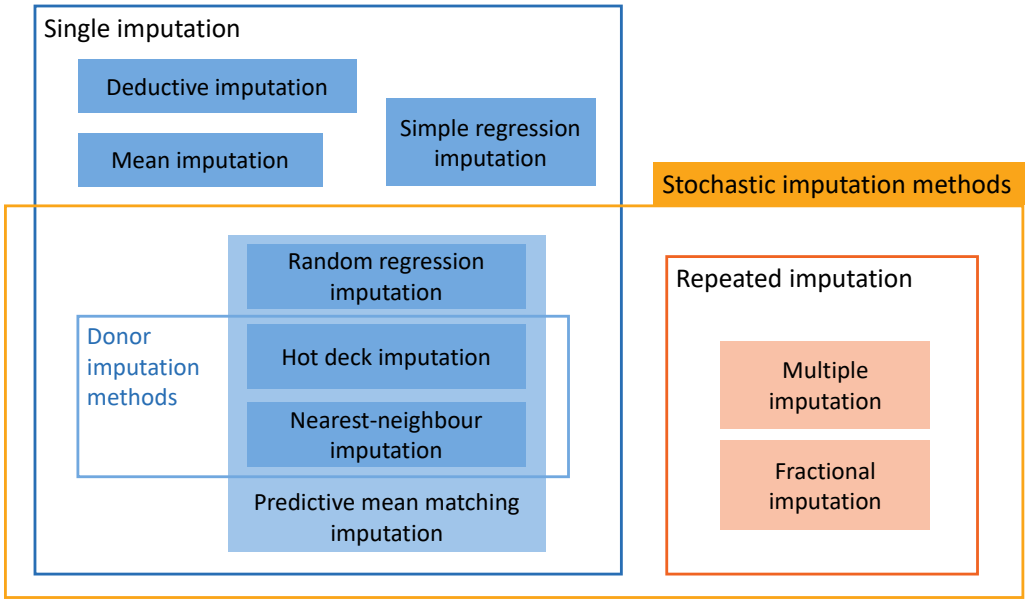
Predictive mean matching imputation is a composite method of nearest-neighbour, hot-deck and regression imputation. A linear regression is fitted for the observed values, and based on the resulting model predicted values for the missing values are obtained. Then, a small number of values (usually three) predicted by the model for cases that are closest to the case with the missing value (according to a certain distance function) are chosen. From this small number of predicted values, one value is chosen randomly and set as the imputed value (Vink et al., 2014). Predictive mean matching imputation is a semi-parametric method making use of the imputation model but not fully relying on it and aiming to reduce potential bias caused by statistical models. Therefore, Schenker and Taylor (1996) assume it to be less sensitive to misspecifications of the underlying model than, for example, regression imputation alone.

15.2.4. Repeated imputation methods: multiple imputation and fractional imputation

In multiple imputation, missing values are replaced by M simulated values, where $M > 1$. For this purpose, a posterior distribution of the missing data, conditional on the observed data, is constructed. Then, M independent random draws are made from this posterior distribution, and M multiply imputed data sets are created, reflecting the uncertainty about imputation. Separately utilising each of the M multiply imputed data sets, statistical analyses are carried out and results are combined to produce a joint point estimate (Takahashi and Ito, 2013). Estimators based on multiple imputation have the potential to have nearly optimal statistical properties (Allison, 2012). However, multiple imputation requires large sample sizes.

In practice, there are different ways to implement multiple imputation. The two most popular and widely used imputation methods are Markov chain Monte Carlo and fully conditional specification methods. For detailed technical explanations, see Schafer (1997), Little and Rubin (2002) and Gill (2008).

Figure 15.1: Schematic overview of imputation methods discussed



Several statistical software packages allow implementation of multiple imputation. One of the most common software packages is imputation and variance estimation software (IVEware), which is part of other statistical programs (SAS, Stata, IBM SPSS Statistics and R). More information on IVEware can be found in Raghunathan, Berglund and Solenberger (2018).

Fractional imputation is another form of repeated imputation that relies on repeating a stochastic imputation method several times, such as repeated random hot-deck or repeated predictive mean matching imputation. The resulting estimator is then viewed as a weighted estimator with fractional weights that reflects the conditional probability of the imputed value given the non-missing observed values in the data set (Yang and Kim, 2016). The main aim of repeated imputation is to improve the efficiency of the resultant point estimator applied in imputation.

To summarise, Figure 15.1 shows the different imputation methods presented in the previous sections. This overview is based on the distinction between single imputation and repeated imputation, but also shows that only three procedures discussed here are not stochastic methods.

15.2.5. Combination of different methods and error terms

All methods described above feature different characteristics and drawbacks. Therefore, a practice to compensate for the different weaknesses and strengths of methods is the combination of methods. There are several ways of combining imputation procedures, which cannot be disentangled from the flag variables currently used in EU-SILC. First, different methods can be applied for the same variable. Second, different imputation procedures can be used for different income types. If the imputation is carried out at the level of income subcomponents, a combination of methods is possible within an income target variable. Third, different imputation methods can also be combined for different subgroups.

Deterministic imputation methods, including mean imputation and simple regression imputation, misleadingly reduce the variance of the data. In order to correct this shortcoming, deterministic methods can be made stochastic by the addition of an appropriate error term, that is, adding 'a "stochastic" random element to the imputed value to

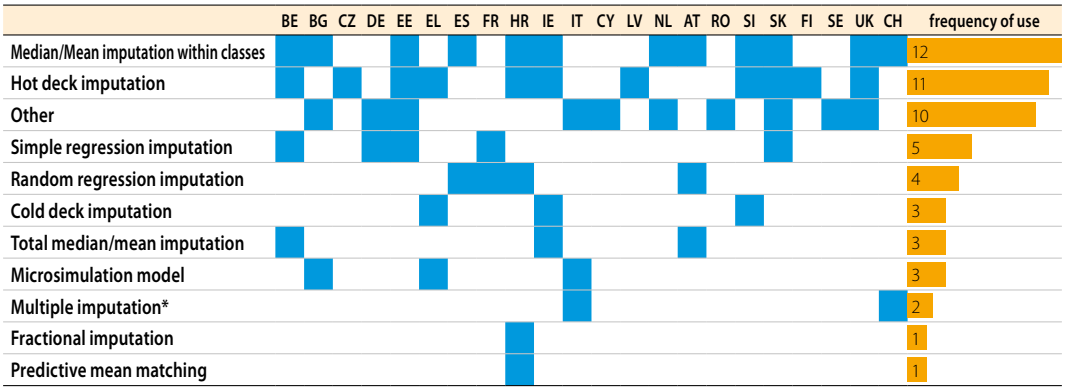
reflect the known variability of estimates' (Groves et al., 2004, p. 331). Most importantly, such an error term must have an expectation of zero in order to not introduce additional bias. Furthermore, the variability of the observed data should also be preserved by the error term. Therefore, a stochastic error term added to an imputed value should be sampled from a distribution with an expectation of zero and variance equal to the empirical variance in the observed data. It is also necessary that the stochastic error term must meet the assumptions of the imputation method applied. For example, in the case of regression imputation, the error term must be sampled from a normal distribution. It also makes sense to restrict the added value with upper and lower bounds in order to allow only the imputation of values within the spectrum of the empirical distribution of the variable that gets imputed. It is important to note that, for donor imputation methods (e.g. hot deck), an error term is already included, because the observed value used as the imputed value includes a realised error.

15.3. Country-specific descriptions of imputation

This section aims to describe the methods for income imputation used by the 31 participating countries in EU-SILC. Therefore, several sources have been used: 22 countries took part in a survey on imputation and weighting (sent out by the Belgian partners of Net-SILC3 at the University of Antwerp), in which the countries were asked to specify the imputation method as well as the statistical software they use for imputations. In addition, the 2016 quality reports were used to analyse the imputation methods applied. On the basis of these sources, an overview of country-specific methods is outlined in Figure 15.2.

According to the survey on imputation and weighting, at least 11 different methods of imputation are currently used. The most common are median/mean imputation within classes and hot-deck imputation, which are used by 12 countries and 11 countries, respectively (Figure 15.2). Two countries use variations of multiple imputation methods, although the results are single values. The difference is the method used (how the single value is obtained from the multiple imputation): the single value is either a random value from the multiple imputation (Italy) or the average of the multiple imputation (Switzerland). Most countries use a combination of imputation procedures: 16 of the 22 countries participating in the survey on imputation and weighting use two or more methods of imputation; and 11 use three or four methods (Figure 15.2).

Figure 15.2: Imputation methods by frequency of use and country (22 EU-SILC countries)



(*) IT – a randomly selected value from multiply imputed values is included in the user database; CH – an average of multiply imputed values is included in the user database.

Sources: Survey on imputation and weighting and 2016 quality reports.

In terms of the software used for imputation, the countries indicated using seven different types of software. The most frequently employed is SAS, which is used by seven countries, followed by five countries using IVEware. Other specific software programs used are the Siena micro-simulation model (SM2) (two countries), SPSS (two countries) and VIM (one country). Four countries use another imputation software program, and three countries use another software program. Almost all of the 22 countries carry out the imputation primarily at the level of subcomponents of the income components; only one country carries out the imputation primarily at the level of income components (target variables), and one country did not indicate at which level the imputation takes place. Since the 2010 EU-SILC, 2 of the 22 countries have undertaken changes in their imputation procedure or imputation strategy: Slovakia started using the VIM package to carry out pre-imputation analysis, nearest-neighbour imputation and regression models in 2014; and Latvia carried out major changes between 2014 and 2016 and substituted the last value carried forward hot-deck method with the nearest-neighbour hot-deck method.

For the remaining nine countries that did not participate in the survey, information on their imputation methods was retrieved from the quality reports. In Norway, only missing values for rent and company car are imputed; no income variables are imputed. In Denmark only rent is imputed, and this information is taken from the income statistics register. Income variables are imputed using tax register data in Iceland and, in the case of the few income variables for which values cannot be obtained from the tax register, data are imputed using a regression method. Hungary uses regression imputation. Malta and Poland use a combination of regression imputation and hot-deck imputation; Poland also uses deterministic methods such as regression deterministic imputation and deduction imputation. Deterministic and deductive imputation methods in the form of mean/median imputation are also used in Lithuania. Luxembourg uses a combination of regression imputation and predictive mean matching. In Portugal, the net series of income data is obtained by the application of a specific gross-to-net micro-simulation model. In addition, for HY025 (within-household nonresponse inflation factor), values from the preceding wave are used, and if they are not available the hot-deck method is applied.

In summary, almost all of the countries impute income variables in one way or another. Only Denmark, the Netherlands, Norway and Sweden do not carry out any imputation, as they use administrative data (apart from target variables for rent and company car). For the other countries, information on how the imputation is carried out and which concrete methods are used is not documented in a detailed way. At best, the methods are named in the quality reports, but a detailed description of how and on which variables the imputation methods are used is often lacking. Nevertheless, countries can be clustered. Most countries use a combination of simple imputation (either deductive or mean/median imputation), regression imputation and donor imputation (either hot-deck or nearest-neighbour imputation). This applies to Belgium, Estonia, Ireland, Croatia, Luxembourg, Poland and Slovakia. Germany, Spain and Austria use a combination of simple imputation and regression imputation; Slovenia and the United Kingdom use a combination of simple imputation and donor imputation; and Malta uses a combination of regression imputation and donor imputation. Nine countries rely on only one of the three methods for their imputation procedure: Bulgaria, Cyprus and Latvia rely on simple imputation (deductive and mean imputation); France and Luxembourg rely on regression imputation; and Czechia, Greece, Latvia, Portugal and Finland rely on donor imputation. Repeated imputation is used by three countries: Italy and Switzerland use multiple imputation, and Croatia uses predictive mean matching imputation. Iceland uses primarily administrative data but refers to regression imputation for three income variables. In addition, there is a group of six countries (Bulgaria, Estonia, Ireland, Spain, Austria and Finland) that use administrative data to a moderate degree but need to rely on imputation methods for missing information. In order to improve cross-national comparisons using EU-SILC data, an initial step would be to further develop the documentation of the imputation procedure in the quality reports. As the following section will show, a second step could be the harmonisation of imputation strategies.

Apart from country-specific applications of imputation methods, it is also relevant to consider which imputation methods are used for which income components. In the quality reports, the description of imputation practices relies on an open field in the section on data compilation (section 3.5.2, 'Estimation and imputation'). Although information on the imputation of rent and the company car has to be provided, there are no guidelines for the description of the imputation of income components. Hence, the documentation varies in detail and extent. The description of the imputation methods in the quality reports is often rather sparse. An exception here is Estonia, whose quality reports, in the annex, provide a detailed description of the income target variables as well as the income variables that are aggregated into the income target variables. Thus, it is impossible to give a comprehensive overview of the imputation methods that are used for each income component.

15.4. Outcomes

Having discussed the methods used for imputation in EU-SILC in the previous section, we now assess the effect of different imputation techniques on the distribution of the total disposable household income (HY020).

In the first part of this section (Section 15.4.1), user database (UDB) data from a few countries are used to compare imputed household income with non-imputed household income. In the subsequent section (Section 15.4.2), 2016 EU-SILC data from Austria are used for simulations of the effect on household income if there are different patterns of missing values and different methods for imputation are used.

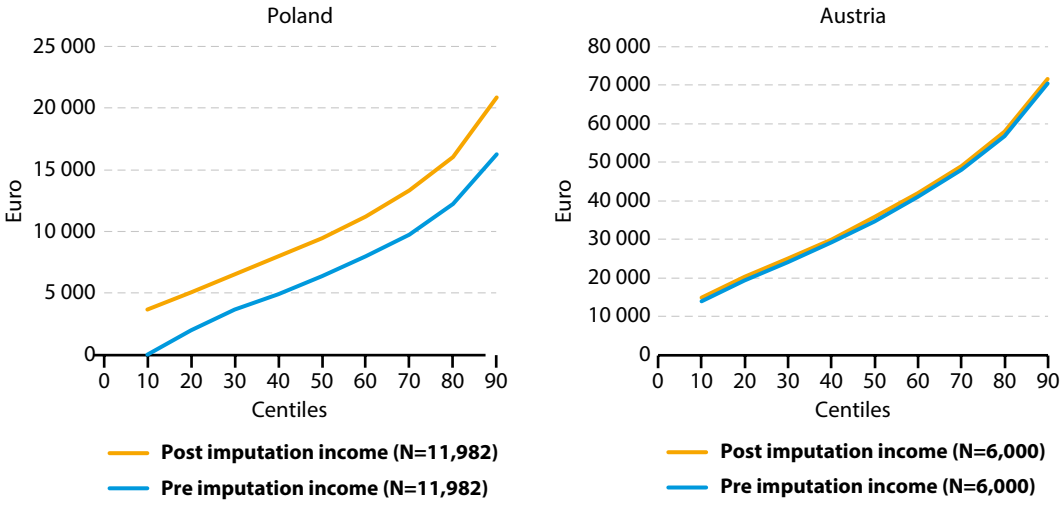
15.4.1. Empirical part

Imputation of household income in EU-SILC is documented by two flag variables in the UDB. The variable HY020_F contains information on whether the data on income were collected in terms of net income or gross income or both. The variable HY020_I should consist of four digits. The first digit should indicate which imputation method has been used, and the second, third and fourth digits should indicate the imputation factor, namely what percentage of the household income has been imputed (Eurostat, 2016a,b). Despite the guidelines of the European Commission, not all country-specific income flag variables in the UDB data sets follow this uniform logic. Instead, quite different formats can be found, which makes working with income flag variables a challenging task (see Chapter 14 of this book).

Nevertheless, it is possible to distinguish the share of non-imputed data from the share of imputed data. In addition, information on the imputation methods may allow conclusions to be drawn on the effect of the imputation techniques used on household income. For this purpose, two analyses are pursued. In the first analysis, two countries (Austria and Poland) with different shares of imputed data and different imputation techniques have been chosen as exemplars to compare the differences in household income between imputed data and non-imputed data. In the second analysis, the household income of the total sample of Austria is analysed in terms of the distribution across income deciles before and after imputation.

In Figure 15.3, the difference in household income before and after imputation is shown for Poland, where data are imputed fully or to some extent for 45 % of households. In the first income decile, the household income was raised from EUR 0 to EUR 3 684 after imputation. The difference is a little smaller for the fifth income decile, in which the household income was raised by about EUR 3 000 after imputation. A considerable increase in household income is shown for the last income decile, with a difference of EUR 4 372 between pre-imputation income and post-imputation income. The comparison between pre-imputation income and post-imputation income for Austria, where data are imputed for 11 % of households, shows much smaller differences, ranging from EUR 886 in the first income decile to EUR 1 114 in the fifth income decile and EUR 1 010 in the last income decile.

Figure 15.3: Pre- and post-imputation household income in Austria and Poland in the 2016 EU-SILC



Source: Authors' own calculations based on data from the Eurostat EU-SILC UDB, 2016.

Apart from comparing the distribution of household income before and after imputation, another important issue concerns the mobility of households between income deciles. In this regard, we investigated how many households move into higher or lower income deciles after imputation. In the case of Austria, the household income of 654 households (11 % of all households) was imputed (1–100 % imputation) in the 2016 EU-SILC data. In total, 48 % of these households with imputed household income remained in the same income decile after imputation, 6 % of the households moved down one income decile after imputation, and 1 % moved down more than one income decile after imputation. The remaining 45 % of households with imputed household income moved to higher deciles in the income distribution: 17 % moved up only one decile, 20 % moved up two to four income deciles, and 8 % moved up more than four income deciles.

However, imputation does not only affect households with imputed income; it can also impact the income deciles of households whose household income has not been imputed. In Austria, of the 5 346 households (89 % of all households) with non-imputed household income, 86 % remained in the same income decile. The remaining 14 % moved down one income decile as a result of the upward mobility of the 45 % of households with imputed household income.

Altogether, 18 % of the total net sample of 6 000 Austrian households in the 2016 EU-SILC changed their income decile after imputation: 5 % moved into lower income deciles, 13 % moved into higher income deciles.

The mobility of households between income deciles is particularly relevant when analysing the at-risk-of-poverty (AROP) rate, which is defined as being below 60 % of the median equivalised income. If households change their relative position in the distribution of household income, the individuals in such households may be reclassified as AROP if their equivalised income falls below the AROP threshold after the imputation of household income. Of course, this consequence of imputation can also have the opposite effect of lifting some households above the AROP threshold.

In comparing pre-imputation income and post-imputation income, it is not only relevant to consider the share of fully or partly imputed income components compared with the share of non-imputed components. The share of imputed income for each income component is also important; this too can impact the decision on which imputation method(s) to use.

Lastly, as well as imputation methods and shares of imputed data, there is another important issue. The deviation of imputed data from non-imputed data is also influenced by the point in time within the process of data preparation when data are imputed, because stochastic imputation methods in particular need additional or auxiliary information for estimating missing values. In the case of Austria, an initial data set is produced after plausibility checks and linking of administrative records. Using this initial data set, target variables are built and missing values are imputed. However, in some countries, imputation may occur early in the process of data preparation, which may have a considerable or negligible effect on the deviation of imputed data from non-imputed data.

15.4.2. Simulation

In this section, the effects of imputation methods on estimates of the AROP rate ([85]) are analysed. This is done using a simulation study based on the complete (imputed) 2016 Austrian EU-SILC data ([86]). The fact that the data are already imputed with some of the tested models should not have any influence on the results based on the simulation. First, the share of imputed data is very small. For example, for employee cash (PY010) only 4.2 % of the data were imputed. For income from agriculture 15.6 % had to be imputed – but for 8.2 % the range of income was known. Therefore, the influence of imputation on the analyses presented here should be negligible (see Statistics Austria, 2017).

Some (originally known) observations are removed and assumed to be item non-response. These are then imputed and the AROP rate is calculated for this artificial data set. Different imputation methods are assessed using different non-response structures and non-response shares. For each of these scenarios (combination of imputation method, non-response structure and non-response share), 1 000 different data sets (repetitions) are analysed (only 500 sets for longitudinal methods). The analysed (cross-sectional) methods are linear models, total median imputation and total mean imputation. The methods are applied separately to each of the 18 income components. The components are subsequently summed to produce an estimate of total household income, on which estimates of AROP rates are based. In addition to the three cross-sectional methods, three longitudinal methods are assessed: last value carried forward, uprating imputation and row-and-column imputation. For a more detailed explanation of longitudinal methods, see Chapter 17 of this book. No combinations of methods are included; all 18 income variables are imputed using the same method. The linear models are optimised in terms of the Akaike information criterion using the function 'stepAIC' of the R package 'MASS'. Finally, the number of covariables is limited. For household income, the variables for the (equalised) household size (HX050), the number of rooms available (HH030), the number of children living in the household (HN13) and region (DB040) are used. Furthermore, the variables for sex (RB090), age (RX020), general health (PH010), region (DB040) and hours worked (PL060) are used for personal income. No covariates are used for the total mean and total median imputation.

For this simulation, the item non-response has to have a certain structure. Four different structures are used: MCAR, MAR and two forms of MNAR (dependent on low income and dependent on low and high income).

Table 15.1 shows the risks of item non-response for each quintile (of income or age) and for different non-response structures when the overall item non-response rate is 5 % or 25 % . The missingness share concerns only those individuals who have an income of this kind or are at least eligible to have it. In total, there are 6 (imputation methods) × 4 (non-response mechanisms) × 2 (non-response shares) = 48 scenarios.

For each item of each observation, a Bernoulli (0/no, 1/yes) distributed non-response flag is created with the mentioned risk. The observations with a 'yes' flag are removed. For example, for dependency on low income and an item non-response share of 25 %, this will lead to approximately 80 % missing data in the first quintile. As using the real quintiles will lead

([85]) The threshold is set to be 60 % of the median equivalised household net income in the country.

([86]) It should be noted that the data are viewed as the 'truth'. This does not indicate how well the data represent Austrian society. For this report, these data stand for a representative sample of a society for which the AROP rate will be calculated.

Table 15.1: Risks of item non-response for the simulation exercise (for overall non-response rates of 5 % and 25 %)

Structure	First quintile	Second quintile	Third quintile	Fourth quintile	Fifth quintile	Item non-response
Dependent: low and high income (MNAR)	0.1	0.025	0.0	0.025	0.1	0.05
	0.4	0.225	0.0	0.225	0.4	0.25
Dependent: low income (MNAR)	0.175	0.075	0.0	0.0	0.0	0.05
	0.8	0.45	0.0	0.0	0.0	0.25
Random (MCAR)	0.05	0.05	0.05	0.05	0.05	0.05
	0.25	0.25	0.25	0.25	0.25	0.25
Dependent: age (MAR)	0.175	0.075	0.0	0.0	0.0	0.05
	0.8	0.45	0.0	0.0	0.0	0.25

NB: For example, the risk of item non-response for an observation in the fourth quintile is 2.5 % for the missingness structure dependent on low and high income when overall non-response is 5 %. For the same scenario with an overall non-response rate of 25 %, the risk is 22.5 %.

to high variances in the share of item non-response, the data are sorted and the first fifth of the applicable data is set to be the first quintile, the second fifth the second quintile, and so on. It should be noted that item non-response affects only the 2016 data – the other years are neither used nor imputed.

First, box plots of the 1 000 repetitions and independent estimations within the scenario median imputation for random missingness are analysed. The box plots of the estimated AROP rate (Figure 15.4a) confirm all expectations: the variance of the rate increases with increasing item non-response. In this case, there is no bias. It should be noted that the absence of bias cannot be deduced for other countries or other years. The one general conclusion is that variance is added by non-response (and the consequent necessity to use imputation). Therefore, imputation should be used only if unavoidable and all opportunities to collect data have been exhausted.

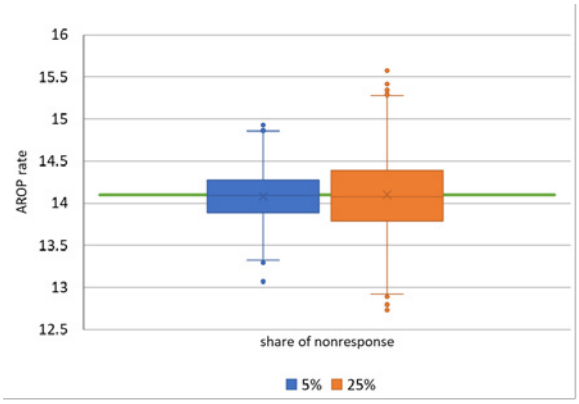
Next, income distributions are analysed for missingness depending on low and high income (assuming a non-response rate of 25 %). Figure 15.4 displays one example of income distribution for each of the imputation methods. The green horizontal line indicates the 'true' income distribution in the survey in the case of no artificial non-response in the simulation and serves as a basis for comparison. Most methods underestimate both tails and overestimate the share with an income slightly above the median. For the same scenario, a box plot is presented with the results of all runs (Figure 15.4c). Which method overestimates and which method underestimates the AROP rate depends on the scenario. Longitudinal methods generally tend to overestimate. As 18 subincomes are imputed instead of the total household income, there is only a very slight increase in the area of mean/median values for the mean/median imputation. A clear spike is visible only when analysing one subincome.

Another result of the simulation study is a dependency between the item non-response share and the 'best method'. For a particular non-response structure, there are several cases in which the best method for 5 % missingness does not show a good performance for 25 % missingness (and vice versa). This may be explained by the impact of the imputation on the new income median (which defines the income poverty threshold). For example, if a household income is below the median both before and after imputation, then the imputation does not change the threshold. However, if the income exceeds the poverty threshold after imputation but not before, then the AROP rate decreases. This outcome may appear quite often for 5 % missingness but would seem less likely for the scenario of 25 % missingness, as the imputation would be more likely to affect the median.

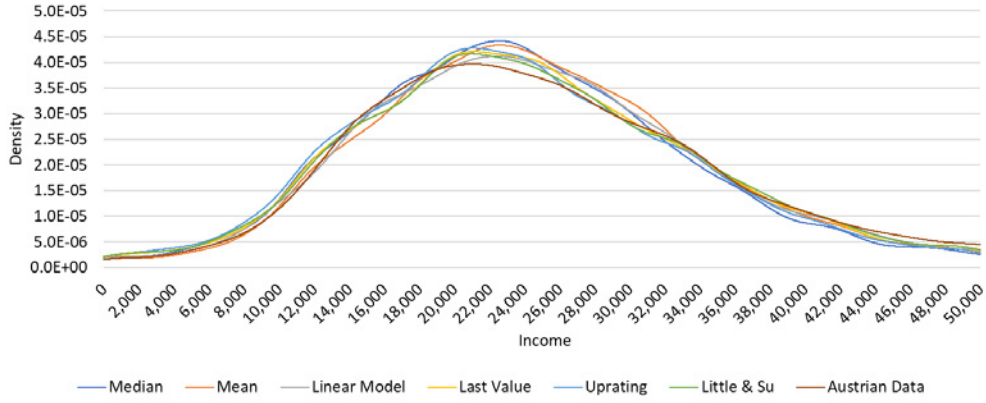
Almost all methods in Austria result in a mean AROP rate varying between 13 % and 15 %, ± 1 percentage point around the true value. Here, the main problem concerns the bias and variance of the AROP rate. The European AROP rates are used for ranking the 27 Member States and the United

Figure 15.4: Distribution of estimated AROP

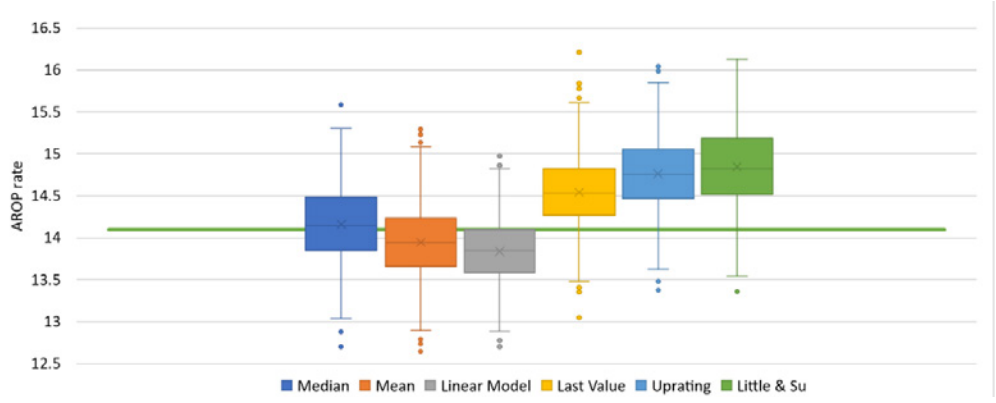
a) Estimated AROP rates (median imputation, random missingness)



b) Density of (Estimated) Income (missingness dependent on high & low income, 25% nonresponse)



c) Estimated AROP rates (missingness dependent on high & low income, 25% nonresponse)



NB: For 5 % missing data, the median and mean of the estimated AROP rates for median imputation and random missingness are close to the 'true' value of 14.1 % (green line in the box plot diagrams). The minimum and maximum differ by approximately 1 percentage point from the 'true' value. In the case of missingness dependent on high and low income and 25 % non-response, last value, uprating and Little and Su (1989) imputation are biased estimates, as their mean and median differ from the 'true' value of 14.1 %.

Kingdom. In 2016, Romania had the highest AROP rate, at 25.3 %, and Czechia the lowest, at 9.7 %. For the 28 countries to be ranked in between those 15 percentage points in a reliable order, the results have to have smaller variance.

Further investigations could usefully consider other missingness mechanisms, such as those that allow missingness to depend on the extent of change in annual income. Do people with a large change in income report as faithfully as people with a steady average income? Furthermore, it could be worth considering other imputation strategies for dealing with the 18 component income variables in the Austrian EU-SILC. Each component has a different missing mechanism, and currently imputation strategies vary between the components. However, it is unclear whether this bottom-up approach improves the post-imputation composite measure of household income or simply imposes volatility.

15.5. Conclusions

In summary, several conclusions can be drawn. First, carrying out analyses with imputed and non-imputed data highlighted the need for better documentation (conclusion I). More transparent documentation enables users to interpret the data in a more holistic way. Second, the analyses show that the use of imputation techniques has an impact on household income and the relative position of households in the distribution of the household income. This may suggest the need for policy considerations in terms of the definition of the AROP rate (conclusion II). Third, the effects of imputation methods on estimates of the AROP rate were investigated in a simulation study, which showed that there was not much difference between most methods (conclusion III). Lastly, the study presented in this work shows that imputation procedures vary substantially between countries (conclusion IV).

As a consequence of these conclusions, two main recommendations can be derived: there appears to be a need for more detailed documentation of imputation procedures (recommendation I), which should facilitate the development of best methods (recommendation II).

Conclusion I: Incomplete or inconsistent documentation

- Documentation of imputation procedures is not always complete or consistent.
- Furthermore, this study has shown that a thorough (qualitative) description of the imputation process (from a 'holistic perspective') helps to improve the understanding of what a national statistical institute (NSI) does.

Conclusion II: Empirical part – household income and the relative position of households

- Comparing the household income of the total sample before and after imputation in deciles shows that household income rises after imputation. This occurs to a larger extent when more information is missing, as is the case for Poland, where data are imputed fully or to some extent for 45 % of households. In the case of Austria, where data are imputed fully or to some extent for 11 % of households, the difference is rather small.
- After imputation, a considerable number of households transition into higher income deciles, which affects the median and the threshold value of the AROP rate. Moreover, it also affects the number of households in poverty, which decreases after imputation.

Conclusion III: Simulation – effect on the AROP rate

- There is little difference between most methods with regard to the AROP rate. There is no method without bias. There is also no method with exceptionally small variance. Mean imputation is a slight exception, as effects are highly dependent on the missing mechanism and are therefore hard to evaluate for real surveys. The biggest exception is multiple imputation with random selection. This method suffers from increasing bias with increasing item non-response. However, fractional imputation can counter this effect.
- With all models, the estimators vary within a range of 1–2 percentage points. The difference

between the lowest variance (Czechia) and the highest variance (Romania), however, amounts to only 15 percentage points. As such, the choice of the imputation method is very important, since it can define the ranking of poverty rates among the participating countries. However, the ranking should not be overinterpreted, as similar rates should just be seen as similar instead of higher and lower. Because of the survey error, there will always be minor deviations from the true value.

- The results of the simulation study raise the question of whether there is a better method with a smaller variance. Longitudinal methods such as the row-and-column method have the advantage of using known information. Therefore, in a future study these longitudinal methods will be analysed in the same way as the methods in this study were analysed.
- If no method is found that is able to resolve this problem, the effects of the methods on the country-specific data should be analysed. Although there is no solution to improve the estimation for Austrian data, there may be one for other countries due to country-specific effects.

Conclusion IV: Different imputation procedures

- A variety of methods are used for imputation across countries. The applied procedures differ in the statistical methods applied as well as in the software used.

Recommendation I: The need for better documentation

- There is a need for a flag variable or several flag variables that are filled in the same way by all countries.
- There is also a need for better documentation of the imputation procedure, including documentation of the process that also covers the level on which the imputation takes place. Ideally, this description would be available for each income and income target variable. This would require a template for the quality reports, with guidelines on how to describe imputation in more detail.

Recommendation II: The need for different approaches and/or the need for harmonisation

- Different reasons for non-response and different patterns of non-response require balanced and customised procedures for an integrated imputation approach. Each NSI should therefore develop such an approach. The country-specific availability of auxiliary information and additional sources allow country-specific approaches and strategies.
- Nonetheless, it also seems necessary to develop a mutual understanding of imputation practices between countries and work towards converging approaches of imputation strategies. Comparative studies can outline the methods used, share best practices, show different effects of the imputation methods used and foster mutual learning. Insights gained by such comparisons could be used to formulate best practices for each imputation procedure.

References

Allison, P. D. (2012), 'Handling missing data by maximum likelihood', *SAS Global Forum 2012 – Statistics and data analysis*, paper 312.

Eurostat (2016a), *Methodological Guidelines and Description of EU-SILC Target Variables – 2016 operation*, Eurostat, Luxembourg.

Eurostat (2016b), *Cross-Sectional Data – Differences between original database (as described in the guidelines) and the anonymised user database*, Eurostat, Luxembourg.

Gill, J. (2008), *Bayesian Methods – A social sciences approach*, 2nd edition, Chapman & Hall / CRC, London.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2004), *Survey Methodology*, Wiley, Hoboken, NJ.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd edition, Wiley, Hoboken, NJ.

Little, R. J. A. and Su, H.-L. (1989), 'Item non-response in panel surveys', in Kasprzyk, D., Duncan, G. and Singh, M. P. (eds), *Panel Surveys*, Wiley, New York, pp. 401–425.

Raghunathan, T. E., Berglund, P. and Solenberger, P. W. (2018), *Multiple Imputation in Practice: With examples using IVEware*, CRC Press, Boca Raton, FL.

Rubin, D. B. (2004), *Multiple Imputation for Nonresponse in Surveys*, Wiley, Hoboken, NJ.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall / CRC, London.

Schenker, N. and Taylor, J. M. G. (1996), 'Partially parametric techniques for multiple imputation', *Computational Statistics and Data Analysis*, Vol. 22, pp. 425–446.

Statistics Austria (2017), *Standard-Dokumentation Metainformationen (Definitionen, Erläuterungen, Methoden, Qualität) zu EU-SILC 2016*, Statistics Austria, Vienna.

Takahashi, M. and Ito, T. (2013), 'Multiple imputation of missing values in economic surveys: comparison of competing algorithms', in *Proceedings of the 59th World Statistics Congress of the International Statistical Institute*, 25–30 August, Hong Kong, pp. 3240–3245.

Vink, G., Frank, L. E., Pannekoek, J. and van Buuren, S. (2014), 'Predictive mean matching imputation of semicontinuous variables', *Statistica Neerlandica*, Vol. 68, No 1, pp. 61–90.

Yang, S. and Kim, J. K. (2016), 'Fractional imputation in survey sampling: a comparative review', *Statistical Science*, Vol. 31, No 3, pp. 415–432.

16 Net–gross conversion in EU-SILC

Richard Heuberger ([87])

16.1. Introduction

This chapter deals with a specific question on item non-response of income variables: the question of how to deal with missing net or gross values if only the counterpart (the respective gross or net value) is available. The task is then to calculate a replacement for the missing value on the basis of the non-missing counterpart: to convert the net value into the missing gross value (or the other way around). This task is executed during the editing of the income data, as one step in the procedure to replace a missing value with an adequate value. The aim of this chapter is to clarify where data producers can influence the net–gross relationship, evaluate the methods used for the net–gross conversions in EU-SILC, discuss alternative options and advise – if possible – whether specific changes in net–gross conversion procedures should be considered.

Generally speaking, in the context of income variables the difference between net income and gross income is taxes and social insurance contributions (SICs). Thus, if these taxes and SICs are known, the calculation of the corresponding value, given that either the gross or the net value is known, should be easy. Consequently, it can be said that this kind of item non-response is favourable, since there may be more information available about the missing value than in other cases.

What makes the question of net–gross/gross–net (NG/GN) conversion specific? Net–gross conversion is a kind of imputation: a set of rules that is used to find a replacement for a missing value. What is specific in this situation is the fact that, in principle, it is possible for everyone to calculate this replacement, since the set of rules defining the relationship between net and gross values is generally known and in a more or less transparent way is applied in the daily routines of administrations. The rules regulating the relationship between net and gross values are defined by taxation laws and laws regulating SICs; there is, in principle, no element of randomness involved – the relationship between net and gross values can be determined in a strict sense. Therefore, NG/GN is not a typical case of imputation, since the imputed value need not be a statistical estimate but may instead be calculated using a deterministic algorithm.

Tax systems and systems of SICs are complex – and different in different countries. So even if in principle it is possible to calculate the missing value, data producers need to have a very good and deep understanding of the national taxation and SIC system. However, even if they have this deep understanding, they may not have sufficient information about the individual case in the data. This is possible, since taxation or SICs may depend on things that are not measured in the survey, such as types of employment contract or employment history.

Only income components that are taxable or subjected to SICs are relevant for net–gross conver-

([87]) Richard Heuberger works at Statistics Austria. The author would like to thank Peter Lynn, Lars Lyberg, Eric Marlier, Matthias Till, Gianni Betti, Sophie Psihoda, Marlene Blüher and participants of the Net-SILC3 International Best Practice Workshop in February 2019 for their very useful comments. All errors are the author's responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Richard Heuberger (richard.heuberger@statistik.gv.at).

sions. These can be income from work or income from capital or social benefits including pensions. Thus, net–gross conversions are important for all income components in EU-SILC, given that they are taxable or subjected to SICs in particular countries. Net–gross conversions can have an impact on the ratio of gross income to net income. If the ratio is not well estimated, the net–gross conversion can influence results regarding any income-based indicators, depending on whether an indicator is based on the gross or the net income.

In particular, when considering the impact of social transfers on income levels and/or income-based poverty, it is important whether gross or net variables are considered and whether NG/GN conversions are done in a way that produces the correct ratio of net income to gross income. If gross variables are considered where net variables should have been considered ([88]), the effect of social transfers is overestimated and thus also the poverty-reducing or redistributive effect of policies. Therefore, the proper application as well as the specific design of NG/GN conversions is relevant. Net–gross conversions are particularly important for countries where only net or only gross variables are collected from respondents. For other countries, conversions are only relevant for those cases in which the respondents were not able or willing to report one of the two corresponding amounts.

This chapter is structured as follows. Section 16.2 discusses the methods used in countries participating in EU-SILC for the NG/GN conversion. Here, the answers that the national statistical institutes (NSIs) provided in a questionnaire on weighting and imputation procedures and the quality reports are used to document the different practices. We are dependent on the information available in quality reports, income flag variables and other information sources to identify the specific features of the NG/GN conversion procedure in participating countries. Section 16.3 deals with the differences between the methods and the empirical evidence on gross–net relations in the participating countries. Differences between the methods are of interest if they lead to different outcomes. Here, the aim is to discuss the applied methods in greater detail. As net–gross conversions are a rather specialised, complex topic and of particular importance to data producers and analysts (although this importance is not self-evident), the concluding section of this chapter (Section 16.4) aims to provide an extensive discussion on why net–gross conversion practices should be further developed and how they could be developed in individual countries and at the European level. It discusses the results and whether a harmonised framework for net–gross conversion would be preferable.

16.2. Which methods are used in EU-SILC and how many cases are concerned?

16.2.1. Methods of net–gross and gross–net conversion in EU-SILC

The Third Network for the Analysis of EU-SILC (NetSILC3) team set up a survey of NSIs conducting EU-SILC to record the methods used in the context of imputation and item-non-response for income questions. Twenty-two NSIs participated in this survey. This survey also recorded the methods used for net–gross conversion. The first question regarding net–gross conversion was whether income is recorded in net form or gross form or in both forms. Table 16.1 presents the answers. In half the countries (11), income is recorded in net and gross forms; in two countries, the decision is dependent on the component. The remaining countries record income only in net form or only in gross form. The form in which income variables are recorded also predetermines the kind of net–gross conversion to a minor extent: only countries where net income and gross income are recorded can use an empirical factor based on EU-SILC data themselves.

([88]) Meaning that (some) deductions for taxes and/or SICs are ignored.

Table 16.1: Net or gross collection of income variables

Country	Only net	Net and gross	Only gross	Depending on component
Austria		x		
Belgium		x		
Bulgaria		x		
Croatia				x
Cyprus			x	
Czechia		x		
Estonia		x		
Finland			x	
France	x			
Germany		x		
Greece	x			
Ireland		x		
Italy	x			
Latvia				x
Netherlands			x	
Romania	x			
Slovakia			x	
Slovenia		x		
Spain		x		
Sweden	NA	NA	NA	NA
Switzerland		x		
United Kingdom		x		

NB: NA, not available.

Source: Net-SILC3 survey on weighting and imputation.

Five of the 22 countries (Latvia, the Netherlands, Slovakia, Sweden and the United Kingdom) specified that they do not apply any form of net–gross conversion during the data editing of income data (Table 16.2). For countries that do not use register information, the net–gross conversion is possibly not at the level of the components of the household income but at the level of the household income itself, or no conversions are necessary because no values are missing at all.

Eight countries (about one third of all responding countries) use an empirical factor (a fixed ratio between net income and gross income) to calculate the missing gross or net value. This empirical factor (or factors) is calculated either on the basis of the EU-SILC data themselves (the collected relations between gross and net values in the data set) or on the basis of other data sources (for example the tax data). If in the latter case the data source is a complete inventory (e.g. tax data), the empirical factor may not be impaired by biases connected with surveys (sampling, measurement errors, etc.).

Ideally, this empirical factor is not derived from the total distribution but is calculated for different classes and then applied accordingly. These classes can be defined by income or any other characteristics that are relevant for taxation or SICs (employment status, children, etc.). The implication here is that the empirical factor will differ between classes, thus allowing for specific properties of the taxation and SIC systems to an extent (minimum and maximum contributions, progressivity of taxes, etc.). Ideally, the classes for the calculation of the empirical factor should be informed by the system of taxes and social benefits. Unfortunately, there is no information about how these empirical factors are calculated in particular countries participating in EU-SILC.

Table 16.2: Methods used for net–gross conversion

Country	Empirical factor	Country-specific model	Siena model2
Austria	Other sources		
Belgium	EU-SILC		
Bulgaria		x	
Croatia		x	
Cyprus	Other sources		
Czechia		x	
Estonia	EU-SILC		
Finland	Other sources		
France	Other sources		
Germany		x	
Greece			x
Ireland		x	
Italy			x
Latvia		No net–gross conversion method applied	
Netherlands		No net–gross conversion method applied	
Romania		x	
Slovakia		No net–gross conversion method applied	
Slovenia	Other sources		
Spain	EU-SILC		
Sweden		No net–gross conversion method applied	
Switzerland		x	
United Kingdom		No net–gross conversion method applied	

NB: The Siena model2 is a microsimulation model designed particularly for net–gross conversion (Betti, Donatiello and Verma, 2011).
Source: Net-SILC3 survey on weighting and imputation.

Calculating the empirical factor on the basis of the surveyed data themselves is prone to bias when there is a substantial difference between missing data and non-missing data: the empirical factor can then be calculated on the basis of only a fraction of the income distributio and this may not include the fraction of the distribution where the missing cases are located. Seven countries use country-specific models for the net–gross conversion; two countries use the Siena microsimulation model (SM2; Betti, Donatiello and Verma, 2011) ([89]). Surprisingly, no country uses Euromod for the net–gross conversion, even though this microsimulation model seems to be specifically appropriate as it incorporates country-specific tax transfer structures.

([89]) The Siena microsimulation model will be briefly described below.

16.2.2. How many cases are subject to conversion?

All individuals and households for whom only gross values or only net values are available could be subject to conversion. No net values are available in the 2015 EU-SILC for the following countries: Denmark, Malta, the Netherlands, Norway, Slovakia and the United Kingdom. These countries use predominantly administrative information on income to convert from gross to net. Finland does not use

administrative information for all variables (but it does for most).

Tables 16.3 and 16.4 describe how many households and people, respectively, report an income for the respective income target variables in the 2015 EU-SILC. At the personal level, the main income source is unsurprisingly income from employment. The second most important income source is old-age benefits: between 20 % and 35 % of all individuals receive old-age benefits. All other benefits are of lesser importance; the share of recipients is between 1 % and 15 %.

At the household level, almost all households have a household income ([90]); for all other income target variables, the share of receiving households is significantly smaller. The highest share of income recipients is observable for interest, dividends, etc. On average, the lowest share of income recipients at the household level is for income received by individuals aged below 16 years.

Of particular importance are countries (for example Bulgaria, Cyprus and Portugal) in which for the majority of the income components the gross values are equal to the net values in the data set. From the analyst's perspective, it is not possible, without any further knowledge of the tax regime and system of SICs in the respective country, to understand why gross values equal net values: is it because there are no taxes and SICs to be deducted, meaning that the income components are not taxable or subjected to SICs, or is it because the deduction was omitted from the data production process? Apart from these countries, the country group in which only gross income is recorded (Denmark, Malta, the Netherlands, Norway, Slovakia and the United Kingdom) is also of interest. For this group, the conversion is set up at another level: not at the level of income subcomponents but at the level of the income target variables and the household income. Detailed information about the conversion procedures in these countries is sparse.

Table 16.5 and Table 16.6 present, for each income target variable, the percentage of cases in which the net income and the gross income differ after the net–gross conversion. For some income variables, the percentage of cases in which the gross value is different from the net value is very high. This is the case for most of the income from employment (mainly because low employment income is not taxable in some countries). In terms of social benefits, countries clearly differ. In some countries, for most of the variables for social benefits there is no difference between net and gross values. In other countries, there are differences between net and gross values in more or less all of these social benefit variables. This highlights obvious differences between tax regimes in Europe. However, in the case of countries where almost all gross variables equal the net variables, it should be investigated whether there are really no taxes and SICs on social benefits for all these variables. For data users, there is no possibility of knowing (from the data) whether all these benefit variables are actually tax exempt or whether net–gross conversion was not part of the data production process. Information can be found either in the EU-SILC quality reports or in the European system of integrated social protection statistics.

([90]) The derived variable, summing up all income components of all household members, is greater than zero for nearly all households.

Table 16.3: Share of households receiving no income by net income target variables, 2015 EU-SILC (%)

Variable	BE	BG	CZ	DK	EE	EL	ES	FR	HR	IT	CY	LV	LT	LU	HU	MT	NL	AT	PL	PT	RO	SI	SK	FI	SE	UK	NO
HY020	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
HY040	100	85	93	100	92	88	87	85	94	86	100	99	95	100	97	100	100	92	98	94	100	91	100	100	99	100	100
HY050	66	83	91	100	68	90	97	74	88	78	77	75	93	67	71	100	100	70	88	85	76	74	100	100	71	100	100
HY060	100	94	98	100	98	92	97	91	95	99	98	90	100	92	95	100	100	96	96	97	99	87	100	100	98	100	100
HY070	100	100	96	100	99	100	99	78	99	98	97	90	100	89	91	100	100	95	98	93	100	99	100	100	92	100	100
HY080	100	89	60	100	97	91	95	93	94	97	100	91	100	100	84	100	100	92	94	95	99	96	100	100	97	100	100
HY090	38	99	100	100	82	93	48	14	93	47	100	77	83	100	98	100	100	30	98	80	89	58	100	100	18	100	100
HY100	100	98	100	100	88	90	75	79	96	85	100	94	100	100	88	100	100	77	92	71	100	92	100	60	43	100	100
HY110	100	99	100	100	99	100	99	99	94	100	100	99	100	100	100	100	100	99	98	100	100	100	100	100	86	100	100
HY120	100	30	25	100	62	22	99	99	76	40	100	21	100	100	36	100	100	100	14	36	8	15	100	100	32	100	100
HY130	100	95	52	100	96	94	93	94	95	96	100	93	100	100	83	100	100	89	95	94	99	90	100	100	99	100	100

NB: In Croatia, 1 % of all households have no household income. For DK, MT, NL, SK, UK and NO, all income variables are recorded as gross variables.
Source: EU-SILC user database, 2015.

Table 16.4: Share of people receiving no income by net income target variables, 2015 EU-SILC (%)

Variable	BE	BG	CZ	DK	EE	EL	ES	FR	HR	IT	CY	LV	LT	LU	HU	MT	NL	AT	PL	PT	RO	SI	SK	FI	SE	UK	NO
PY010	54	50	52	NA	40	75	53	44	66	59	NA	43	50	43	52	NA	NA	44	59	58	67	43	NA	NA	30	NA	NA
PY020	78	92	NA	NA	85	97	94	84	96	95	NA	94	NA	94	100	NA	NA	100	88	97	100	92	NA	83	84	NA	NA
PY050	94	89	NA	NA	93	86	90	94	89	83	NA	95	90	97	74	NA	NA	88	90	94	87	84	NA	NA	87	NA	NA
PY080	100	100	99	NA	99	100	98	100	100	100	NA	100	NA	100	100	NA	NA	95	100	99	100	99	NA	NA	89	NA	NA
PY090	90	95	98	NA	95	98	85	90	98	90	94	94	95	96	96	NA	NA	90	98	95	100	95	NA	NA	96	NA	NA
PY100	78	63	64	NA	70	71	82	68	78	73	77	64	NA	80	68	NA	NA	72	72	72	67	77	NA	NA	70	NA	NA
PY110	99	99	90	NA	99	95	93	99	93	92	95	99	NA	96	99	NA	NA	94	97	93	94	97	NA	NA	100	NA	NA
PY120	98	87	95	NA	87	100	98	99	99	NA	97	89	88	99	96	NA	NA	96	99	98	100	87	NA	NA	83	NA	NA
PY130	95	94	94	NA	90	98	97	98	91	97	97	95	NA	96	95	NA	NA	97	95	97	97	93	NA	NA	97	NA	NA
PY140	98	100	100	NA	98	100	98	99	99	99	97	98	NA	98	99	NA	NA	98	99	99	100	94	NA	NA	87	NA	NA

NB: In Belgium, 54 % of all individuals aged at least 16 years receive no income from employment. For DK MT, NL, SK, UK and NO, all income variables are recorded as gross variables. NA, not available.
Source: EU-SILC user database, 2015.

Table 16.5: Share of households in which gross income is not equal to net income, 2015 EU-SILC (%)

Variable	BE	BG	CZ	DK	EE	EL	ES	FR	HR	IT	CY	LV	LT	LU	HU	MT	NL	AT	PL	PT	RO	SI	SK	FI	SE	UK	NO
HY020	95	87	94	100	82	98	93	100	67	98	88	94	70	100	96	93	100	98	99	93	96	97	98	99	99	90	97
HY040	0	3	100	NA	81	100	37	99	66	93	0	0	49	0	100	NA	NA	50	76	16	0	93	NA	0	100	NA	NA
HY050	6	0	0	NA	6	97	16	100	0	0	0	1	40	8	74	NA	NA	0	13	0	0	25	NA	0	46	NA	NA
HY060	0	0	0	NA	0	0	5	0	0	0	0	0	0	42	13	NA	NA	0	0	0	0	0	NA	0	0	NA	NA
HY070	0	0	0	NA	0	0	0	99	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	NA	0	0	NA	NA
HY080	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	11	NA	NA	0	0	0	0	0	NA	0	0	NA	NA
HY090	0	12	0	NA	11	100	100	89	0	99	0	90	13	0	100	NA	NA	100	100	100	9	40	NA	0	89	NA	NA
HY100	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0		0	NA	99	99	NA	NA
HY110	0	0	0	NA	10	0	3	0	0	0	0	80	100	50	100	NA	NA	46	18	15	100	10	NA	0	55	NA	NA
HY120	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	NA	0	0	NA	NA
HY130	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	NA	0	0	NA	NA

NB: In Belgium, for 95.1 % of all households the gross income differs from the net income. For DK, MT, NL, SK, UK and NO, all income variables are recorded as gross variables (except HY020). NA, not available.

Source: EU-SILC user database, 2015.

Table 16.6: Share of people for whom gross income is not equal to net income, 2015 EU-SILC (%)

Variable	BE	BG	CZ	DK	EE	EL	ES	FR	HR	IT	CY	LV	LT	LU	HU	MT	NL	AT	PL	PT	RO	SI	SK	FI	SE	UK	NO
PY010	97	92	100	NA	99	99	95	99	100	99	100	94	99	100	100	NA	NA	94	99	100	100	86	NA	0	94	NA	NA
PY020	25	0	0	NA	100	0	78	0	87	95	0	0	0	100	98	NA	NA	0	0	0	2	18	NA	91	98	NA	NA
PY050	91	61	0	NA	38	94	88	100	56	94	100	46	64	47	100	NA	NA	68	91	100	8	28	NA	0	67	NA	NA
PY080	0	0	0	NA	0	0	22	64	0	100	0	0	0	100	0	NA	NA	13	100	24	0	0	NA	0	100	NA	NA
PY090	39	0	0	NA	52	91	54	53	0	88	0	0	12	100	46	NA	NA	0	100	0	7	86	NA	0	97	NA	NA
PY100	76	0	0	NA	0	100	50	0	25	86	19	43	0	100	80	NA	NA	97	100	50	23	10	NA	0	99	NA	NA
PY110	66	0	0	NA	0	96	32	42	0	93	7	5	0	100	0	NA	NA	95	100	55	1	3	NA	0	88	NA	NA
PY120	13	0	0	NA	100	38	33	0	0	0	0	70	100	100	100	NA	NA	21	41	0	100	100	NA	0	100	NA	NA
PY130	8	0	0	NA	0	71	32	0	16	23	0	8	0	91	81	NA	NA	75	92	24	0	15	NA	0	95	NA	NA
PY140	0	0	0	NA	0	0	1	0	0	0	0	0	0	0	0	NA	NA	0	0	0	3	0	NA	0	0	NA	NA

NB: In Belgium, for 97.3 % of the recipients of income from employment the gross income and the net income differs. For DK, MT, NL, SK, UK and NO, all income variables are recorded as gross variables (except HY020). NA, not available.

Source: EU-SILC user database, 2015.

16.3. What are the main differences between the methods used?

As laid out in the introduction, differences between the methods used for NG/GN conversion are important if they lead to different outcomes. Based on the work of Leventi, Papini and Sutherland (2021), Guio, Marlier and Nolan (2021, p. 29) highlight that it matters whether the net or the gross value of transfers is taken into account: 'Depending on whether transfers are considered gross or net, the ranking of countries also changes substantially in terms of the anti-poverty effectiveness of their monetary social provision systems.' It therefore matters how income is recorded and whether the net–gross conversions are done properly.

The net–gross conversion changes or influences just a (more or less) small part of this relationship when only gross or net values are known. For a small fraction of countries, this is of particular importance, since for all (or nearly all) income target variables only one part of the net–gross relationship (mainly the gross variables) is known. When net and gross variables are collected, net–gross conversion is necessary only for a part of the sample: either when only one of the two values is given by the respondents or when the relationship between gross and net values collected from respondents is considered not plausible in the data editing ([91]).

The success of poverty reduction can mainly be measured on the basis of net values, because the effect of the tax system and the SIC system (which causes the difference between net and gross values) is an essential part of this policy. In cases in which there are no differences between net and gross values (gross equals net), it can be assumed that the 'poverty-reducing effect' is greater, since there are no deductions from the gross value. In some cases – as pointed out by Leventi, Papini and Sutherland (2021) – there is some uncertainty about whether it is actually true that the net value equals the gross value.

Net–gross conversion practices in EU-SILC can be distinguished between procedures that use an empirically determined factor to calculate the missing value and procedures that somehow aim to model the missing value. However, the result can be the same ([92]); these different approaches or families of approaches have different demands on the data producer and typically differ regarding complexity. Model-based approaches require extensive knowledge of the tax–benefit system and the characteristics of the individual case. A constraint is whether all the information necessary is available in the data set at hand. In addition, the question is whether the model is restricted to the information available in the data records (basically the extent of EU-SILC) or whether additional information is available (e.g. administrative data) for the modelling.

In contrast, an empirical conversion factor can in principle be determined without any knowledge of the characteristics of the individual case or the tax transfer system. It relies only on the observed ratio of net to gross among all cases with reported net and gross values. These conversion factors can be calculated either on the basis of the net and gross values available in the current data set or on the basis of external data (e.g. tax data). However, the calculation of this empirical factor can be made relatively complex by adapting the calculation of the factor to various circumstances relevant in the particular tax–benefit system. If all these complexities are fully taken into account for the calculation of these empirical conversion factors, the result would be the same as for a model-based approach.

For a model-based approach towards net–gross conversion, countries participating in EU-SILC use either SM2 or customised models (microsimulation models modelling the system of taxation adapted to a country's situation with more or less linkage to standardised models ([93])). The SM2 is a general tool designed to estimate or model income variables in the comparable and standardised form required

([91]) The extent of these latter cases is not known and not documented in the EU-SILC flag variables, or, at least on the basis of the flag variable stating that there has been net–gross conversion, it cannot be determined whether the missing corresponding value was not given by the respondent or is missing because of the data editing process.

([92]) Depending on the complexity of the tax–benefit system, the model and/or the calculation of the conversion factor.

([93]) As an example, see Farinha Rodrigues (2007) for the situation in Portugal.

for EU-SILC. 'Starting from data on household and personal income given in different forms (including some missing data), and country, the model estimates full information on income by component, with breakdown of gross, amounts into taxes, social insurance contributions, social transfers, and net and disposable income' (Betti, Donatiello and Verma, 2011, p. 37). The gross-to-net algorithm is designed to allow for adaptations to suit the country-specific situation as far as taxable units and level of aggregation, specific deductions by income component, tax credits, and so on, are concerned.

Another model-based approach in this context would be the use of Euromod (for a description of Euromod, see Sutherland, Immervoll and O'Donoghue (1999) and Immervoll and O'Donoghue (2001)). Euromod is a tax–benefit model based on the microdata sets from EU-SILC and is currently available for 28 countries (the EU Member States and the United Kingdom). Euromod is used for comparative analysis and simulations and policy swaps between countries as well as at the EU level. Euromod models the tax–benefit system of each individual country. It calculates the results for each household (micro level) and summary statistics (macro level) for the assessment of a policy (or policy change). Hence, the effects of policies can be analysed at the household level as well as at the level of income distribution. Users can change the setting of the tax–benefit system to analyse changes in household income, income composition, the distribution of household income and the costs of the policy change (Sutherland, 2018). Since Euromod reproduces the tax systems and SIC systems of countries, it can be used to calculate the corresponding net variables on the basis of gross variables. Therefore, it can be used for net–gross conversions. Euromod is not currently available for all countries participating in EU-SILC (non-EU countries, accession countries) but it is highly adapted to the country-specific tax system and SIC system. Apparently – and perhaps surprisingly – no NSI participating in the survey uses Euromod for net–gross conversion.

When comparing SM2 and Euromod, they seem to be equally useful for net–gross conversions in EU-SILC. The differences are the more 'model-based' character of Euromod and the more generic approach of SM2. Euromod requires full knowledge of the taxation and SIC systems in a country; in return, Euromod is less dependent on the input data (considering net–gross relations etc.). Euromod is fully adaptable to country-specific requirements. In contrast, SM2 seems to be more data driven and more generic, with a clear distinction between a common standardised part and country-specific parts. Therefore, SM2 seems to be easier to adapt to country-specific needs ([94]) but somewhat less specific to country details.

16.4. Conclusions

Given the different aspects covered in this chapter, the conclusions are introduced in several parts. The first part proposes modifications and extensions to the documentation of net–gross conversions in EU-SILC (and of imputation in general as well). The second part presents several questions directed at producers of EU-SILC data to provide a framework for considering improvements to net–gross conversion practices in a country. These questions should serve as guidance in reviewing and rethinking net–gross conversion procedures. The third part of the conclusion considers the question of whether harmonisation of net–gross conversion procedures in EU-SILC is desirable or not.

Conclusion I: Documentation

The analysis in this chapter illustrated gaps in the documentation of income target variables and in the quality reporting of EU-SILC. Here, the focus should not be primarily on a revision of the flag variables (which is currently in progress at the European level), but should be more on adaptation of the quality reporting.

Qualitative description. Besides the description of the imputation process and the extent of imputations in the qualitative report (and the documentation of the flag variables), a qualitative description

([94]) SM2 is also used by countries acceding to the EU, such as Albania, Montenegro and Turkey. In addition, it is in development for Bosnia and Herzegovina and Kosovo (*). (*) This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo declaration of independence.

of the net–gross conversion procedures would support an understanding of what is happening in the countries. The idea is not to increase the burden for the data producers (or people responsible for the documentation process). Instead, it is to allow for a better understanding of the net–gross conversion process. Countries using a customised model for the conversion could and should explain (briefly) the basics of the model. Countries using a fixed factor (or more than one factor) could explain how this factor was (or these factors were) determined. Countries using some kind of model could describe their adaptations of the model specification. These questions and answers are difficult to formalise; therefore, the suggestion would be a qualitative description.

Description per variable. Complementary to the above, the documentation of the net–gross conversion procedures should also be expanded at the level of individual variables rather than at the aggregate level. Currently, the flag variable may ([95]) give the information that a net–gross conversion procedure was applied but not how it was applied or which net–gross conversion was applied. This individualisation of the documentation of the net–gross conversion procedure is relevant, even though we can assume that the income target variables themselves are aggregated variables (a sum of income components). So far, there is no information in the documentation available to users on whether the implementation of net–gross conversion practices happened in practice and which tools were used. This information should be (ideally) available at the level of individual variables.

Taxation and SICs. In addition, the documentation should inform users about whether taxes and/or social insurance are included or considered. At present, the flag variable (ideally) provides information about how the variable (or all underlying income components) is recorded (gross, net of personal income tax, etc.). However, information about what should have happened to the data is missing, and it is not possible with the information at hand to identify whether taxes and/or SICs should have been deducted or not (and were not).

([95]) For some cases, it was not fully clear whether the flag variable was filled correctly; therefore, the interpretation was difficult.

Flag variables. Nonetheless, Member States and Eurostat are asked to consider whether the information in the flag variables concerning net–gross conversion procedures is sufficient. Naturally, there is a trade-off between the needs and demands of analysts and the desire of data producers and people committed to the documentation of EU-SILC user files for simplicity.

Conclusion II: Improving conversion procedures

It is recommended that data producers considering revisions to their net–gross conversion procedures carefully consider the following questions.

What are your resources for the revision of net–gross conversion procedures?

Any changes in the routines of the data production need resources: time and intellectual capacity. Both can be translated into money, more or less directly, by equating them with working hours and staffing. A change in the net–gross conversion procedures may necessitate the development of models, learning to adapt existing models and advanced software routines. The question is whether there are adequate resources to manage a change in the net–gross conversion procedures. If there are insufficient resources for a change in strategy or change in procedure, the question of how to manage such a change is obsolete. The key questions are whether a change in the routines is necessary and what resources are needed to manage such a change. Whether change is necessary may depend on how many cases are treated by the net–gross conversion procedures.

How many cases are subjected to net–gross conversion?

It may sound simple, but the need for rethinking net–gross conversion procedures will depend on the number of cases that are treated by these procedures. If there are no or hardly any cases subjected to net–gross conversions in one country, the need for rethinking these routines will be small. If

all the income of all subcomponents – or at least a great share of income – is subjected to net–gross conversions, the importance of these conversions and any consideration about changes in procedures are evident. Net–gross conversions are of particular importance in countries that survey (or take from a register) income or income components only in one form – either in gross form or in net form. The overview in this report of the number of income recipients and of net–gross relations of income components offers an insight into where a possible reworking of net–gross conversion procedures is relevant and where it is of minor importance.

What information about the net–gross ratio is available?

This question is also partly about a specific resource: the information available about the net–gross ratios within a country. From the perspective of data production, the question is whether the information about the net–gross relationship can be taken only from the data themselves or whether there are other external information sources that can be employed for the net–gross conversion procedures. In general, the more information is available about the net–gross relationship, the more opportunities are available to rework the net–gross conversion procedures. Thus, if only information from the survey is available, the possible reworking of the conversion procedures is limited. This is particularly of concern to countries that use (empirical) factors (simple or in classes) for the net–gross conversion. The employment of models for the net–gross conversion – for example Euromod or other tax–benefit microsimulation models – requires solid knowledge of the tax–benefit system and SIC system. Evidently, solid knowledge of the tax–benefit system of one's country is generally vital for the implementation of deterministic net–gross conversion procedures.

Complexity of the tax system and system of social insurance contributions

The complexity of the tax–benefit system may influence how smoothly and unproblematic changes in the net–gross procedures can be implemented. In addition, the complexity determines what information is needed to implement the net–gross conversion. If information needed to calculate taxation or SIC is not available in the EU-SILC data set, it is presumably difficult to build a customised model for the net–gross conversion.

Conclusion III: Harmonisation versus individualisation

The decision between harmonisation of methods and individualisation has accompanied comparative surveys in the social sciences possibly from the start. EU-SILC is no exception. In the context of this chapter, the specific question is whether a harmonised approach, meaning using one method for all countries, towards net–gross conversion would be desirable (and for whom). As for imputation practices in general, we observe a variety of approaches, in which we can distinguish between more model-based approaches and approaches based more on the empirical evidence of the data or of other sources. For some countries, it can be assumed that net–gross conversion is not undertaken. From the perspective of the data production, any improvement of the net–gross conversion practice and of the documentation is desirable. If there is no net–gross conversion in a country, any net–gross conversion procedure is an improvement. If there is a simple gross–net factor, any improvement (and individualisation) of this factor is an improvement. If a country uses specific group-wise factors, a (simple) model-based approach is desirable. The implicit ordering of approaches in terms of quality implies a guideline for improvement, which could be made explicit by Eurostat.

This guideline could represent a move towards harmonisation, given that resources and willingness for improvement are available. Finally, of course, a common approach – with regard to comparability – would be desirable; this approach cannot be based on 'empirical factors', since these are not available for countries where only gross variables are recorded. Ideally, it would be a model-based approach such as SM2 or Euromod. Which one of these two approaches is preferable should be openly discussed. In short, Euromod seems to

be more complex and more adaptable to country-specific standards, but the costs to adapt the model seem to be higher. SM2 is more generic and more 'data based', and therefore possibly easier to adapt and to integrate into the data production process. Another topic in this context concerns the level at which this harmonisation could and should be implemented: if the implementation of imputation flags is done correctly, the implementation of harmonised net–gross conversion could be done centrally. This centralisation would increase the comparability and the efficiency of the net–gross conversion.

References

Betti, G., Donatiello, G. and Verma, V. (2011), 'The Siena microsimulation model (SM2) for net gross conversion of EU-SILC income variables', *International Journal of Microsimulation*, Vol. 4, No 1, pp. 35–53.

Farinha Rodrigues, C. (2007), 'Income in EU-SILC – net/gross conversion techniques for building and using EU-SILC databases', in Eurostat, *Comparative EU Statistics on Income and Living Conditions: Issues and challenges*, Office for Official Publications of the European Communities, Luxembourg, pp. 157–172.

Guio, A.-C., Marlier, E. and Nolan, B. (2021), 'Improving the understanding of poverty and social exclusion in Europe', in Guio, A.-C., Marlier, E. and Nolan, B. (eds), *Improving the understanding of poverty and social exclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 25–38.

Immervoll, H. and O'Donoghue, C. (2001), 'Imputation of gross amounts from net incomes in household surveys – an application using Euromod', *Euromod Working Paper Series*, No EM1/01, University of Essex, Colchester (https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em1-01).

Leventi, C., Papini, A. and Sutherland, H. (2021), 'Assessing the anti-poverty effects of social transfers: net or gross? And does it really matter?', in Guio, A.-C., Marlier, E. and Nolan, B. (eds), *Improving the understanding of poverty and social exclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 123–138.

Sutherland, H. (2018), 'Quality assessment of microsimulation models: the case of Euromod', *Euromod Working Paper Series*, No EM19/18, University of Essex, Colchester (https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em19-18).

Sutherland, H., Immervoll, H. and O'Donoghue, C. (1999), 'An introduction to Euromod', *Euromod Working Paper Series*, No EMO/99, University of Essex, Colchester (https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em0-99).

17 Longitudinal imputation of EU-SILC income variables

Nadja Lendle and Matthias Till ([96])

17.1. Introduction

When information on certain variables is not obtained in data collection or is considered erroneous in the editing procedure, it is often replaced by artificially generated values that are considered plausible. This procedure is called imputation. Imputation can involve the estimation of values through regressions on related variables or donor methods that replace missing values with values from other units that are considered similar. Cross-sectional imputation refers to the exclusive use of information from the same data collection exercise (survey wave). By contrast, longitudinal imputation refers to procedures that use information successfully obtained in other waves of a panel survey. This chapter aims to provide guidance on a simple method to perform longitudinal imputation for European Union Statistics on Income and Living Conditions (EU-SILC) income components.

Almost every survey is affected by item non-response – that is, when certain information items requested from respondents remain unknown. The level of item non-response can be considerable. In EU-SILC, this is aggravated by the large number of potentially missing variables that are used to construct total income. Verma, Gagliardi and Ferretti (2010) argued that missing items in an income survey not only lead to inaccurate measurement but can imply a coverage problem if they lead to exclusion of certain units from the analysis. Although conventional opinion surveys often simply ignore item non-response by excluding a usually small number of incomplete cases, this strategy is very dangerous for EU-SILC. Even if individual income components are missing for only a small number of observations, the number of records that are complete on all variables may be considerably reduced. The United Nations Economic Commission for Europe guide on measurement of poverty (UNECE, 2020, p. 127) reports that, in the Canadian Income Survey, income components need to be imputed for roughly 12 % of respondents. In 21 out of 31 national quality reports ([97]), countries reported the percentage of households with partial information (before imputation) for total disposable household income (HY020). For 12 countries, the percentage of households with incomplete information on this important variable amounts to more than 30 %. This gives an idea of the potential extent of the problem. As the likelihood of completeness depends on the diversity of the income portfolio, excluding incomplete cases would introduce unacceptable bias. A comparative assessment of panel surveys in Australia, Germany and the United Kingdom highlighted especially that the analysis of incomplete cases would underestimate mobility patterns (Frick and Grabka, 2007). More information on missing data in general can be found in Little and Rubin (2002).

([96]) Nadja Lendle is with the Institute of Applied Statistics at the Johannes Kepler University Linz, Austria, and Matthias Till is with Statistics Austria. The authors would like to thank Eric Marlier, Peter Lynn and several colleagues for their very useful comments. All errors are the authors' responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Nadja Lendle (nalendle@gmail.com) and Matthias Till (matthias.till@statistik.gv.at).

([97]) The authors reviewed quality reports for the 2018 operation, except for nine countries for which only previous versions of the quality report were accessible (Croatia, Denmark, France, Germany, Iceland, Latvia, Norway, Poland and Portugal).

Since EU-SILC follows an integrated design, in principle it provides longitudinal information that could be used for imputation. Using longitudinal information for imputation has the evident advantage that it may be highly correlated with missing information for each individual or household. This may not only improve the accuracy of cross-sectional estimates but may also be particularly relevant for estimates that involve information on change over time. However, treating observations independently over time as in cross-sectional methods implies that only the correlation within waves is restored, whereas correlations over time are ignored. Estimates of longitudinal indicators such as the persistent at-risk-of-poverty (AROP) rate may thus be biased if longitudinal information is ignored.

It should be noted that all cross-sectional imputation methods have been found to considerably add variance (and potentially bias) to the AROP estimate (Psihoda et al., 2021), so imputation should be used only when all possible efforts have been made to avoid missing information in the first place. Strategies to prevent item non-response usually focus on the design of questions (see, for example, de Leeuw, Hox and Huisman, 2003). With computer-assisted data collection, for example, it is possible to issue an alert to the interviewer or the respondent that information for a certain item is missing. However, this has clear limits in the context of the detailed measurement of every single income component. For example, small or irregular sources of income may be unknown to the respondent, especially if he or she provides information as a proxy for another household member. Likewise, information that has been collected may have been found to be erroneous during the editing procedure, and data that are normally retrieved from registers may be missing for a particular income component for a particular household member. Therefore, in practice, imputation will be unavoidable in the context of income surveys. Making good use of longitudinal information can be considered a strategy to reduce uncertainty and improve comparability. Frick and Grabka (2007) found that the level of imputed values seems to depend critically on the imputation method. Policy indicators, such as the AROP rate, may therefore be sensitive to the way that missing data are handled. For comparative data collection such as EU-SILC, it is therefore desirable that methods used to impute missing values are standardised as much as possible. This chapter aims to provide some guidance on how longitudinal imputation may be best accomplished in the context of EU-SILC.

Section 17.2 reviews the current and potential use of information from prior waves of EU-SILC for longitudinal imputation of missing income values. It builds on information gathered by a survey of current country practices (see Chapter 24 of this book), which revealed that cross-sectional imputation methods are widely used. The current chapter puts an emphasis on longitudinal methods such as last value carried forward with or without uprating, which are used in several countries. Section 17.3 outlines the row-and-column method of Little and Su (1989), which is only slightly more sophisticated than the last value carried forward and is a viable alternative to more advanced approaches. The row-and-column method has been proposed by Eurostat in the context of business statistics (see Eurostat, 2014). Despite its theoretical merits and simplicity, it does not yet appear to be widely used for EU-SILC.

Section 17.4 presents some simulation results that illustrate the potential impact of different longitudinal imputation approaches on income poverty measures. The chapter concludes (Section 17.5) with recommendations to make more and better use of longitudinal information for imputing missing income data.

17.2. Current practices and potential for longitudinal imputation of income variables

Strategies for imputation can be broadly divided into regression-type methods and donor-type methods (Gelman and Hill, 2006). They can be applied to one variable or to impute a vector of variables at the same time. They may also involve an iteration of imputations, whereby imputed values are used to impute other missing values as implemented in IVEware software, which can be used with SAS, Stata, IBM SPSS Statistics and R packages

or as a standalone in Windows, Linux or Mac OS operating systems (Raghunathan et al., 2001). The sequential multivariate model ensures that correlation between variables is preserved. Multiple imputation is an approach that considers the uncertainty that is necessarily introduced by any of these methods. It produces several plausible values. The variance between the values can then be added to the sampling variance.

Imputation methods (and models) for EU-SILC are currently not standardised across countries. This may limit the comparability of indicators such as the AROP rate. Statistics Austria surveyed EU Member States on which imputation methods are used for EU-SILC (see Chapter 15 of this book). Most of the imputation methods used in practice use only cross-sectional information. The countries use a variety of different cross-sectional imputation methods, but generally simple approaches are dominant. They range from treating missing income values as zero to relatively sophisticated models and may vary between income components. More information on the results of the survey can be found in Chapter 15.

Although it may be justified to avoid complexity, it can be problematic to leave longitudinal information – perhaps even from the identical variable in a previous year – unused. To illustrate this, consider a simple cross-sectional imputation procedure: all missing values for households that belong to a certain category are replaced by the mean income of households in the same category (e.g. defined by household size). The procedure may, of course, be improved by increasing the number of characteristics used to determine the category of households (or by using a model to predict the missing value). In a cross-sectional approach, such characteristics would, however, always come from the same year. Thus, a household with a low income in the previous year will get the same imputed value as a household with a high income in the previous year provided it has the same characteristics in the current year. As long as the analysis of the data remains cross-sectional, this is only a problem of efficiency. However, once the analytical perspective shifts to the longitudinal dynamics of income, there will be obvious implications for the mobility observed. Restoring both longitudinal and cross-sectional multivariate structures is often seen as difficult. This is perhaps primarily a consequence of the large number of potential associations between variables. In practice, we will often lack a precise understanding of how different income components are related to each other. Every ad hoc theory will have its exceptions. This chapter proposes the use of longitudinal information as a strategy to make best use of the available information without needing to specify complex models.

Longitudinal imputation is particularly useful for complementing cross-sectional methods. Ideally, if used simultaneously in the form of multivariate models, the disadvantages of one method can be balanced by the other method. In practice, cross-sectional methods will be mostly used to impute cases in which longitudinal imputation is not possible. In the EU-SILC integrated rotational design, longitudinal information is potentially available for only approximately three quarters of the sample in any year. The initial wave, which cannot (yet) include data from previous waves, will typically be imputed with cross-sectional information only.

Theoretically, it would be possible to use longitudinal information from later waves to impute retrospectively. However, this option is ruled out in practice because of time constraints. Final data processing would effectively be delayed until completion of a rotational cycle, that is, up to 4 years after the year of data collection. Although any such delay in the production cycle of policy-relevant indicators appears unrealistic, it could be argued that retrospective imputation may be part of an *ex post* quality assessment. In such an exercise, actual estimates would be compared with results that take full advantage of all information, including subsequent longitudinal data, to assess the sensitivity with regard to the imputation method chosen.

As regards the strategic approach to longitudinal imputation, in principle longitudinal information may be utilised like any other covariates for hot-deck or regression-type imputations. Instead of using strata or predictors only from time point t, variables that were observed at time $t - 1$ (or earlier) may be used. Such an approach may potentially lead to rather heavy models with a large number of explanatory variables. This may drastically dimin-

ish the number of donors in imputation cells or increase the risk of model (mis)specifications.

The MetaSILC survey of country practices on weighting and imputation showed that, if longitudinal information is used, methods are often fairly trivial (see Chapter 14 of this book). They impute missing income information either by the last observed value or by adjusting prior values by some factor to accomplish uprating.

17.2.1. Last value carried forward

This method is by far the simplest way of using longitudinal information for imputation. If, for example, an income component is missing for an individual or household for which a valid value was observed in the previous wave of data collection, the missing value is simply replaced by that amount without any further adjustment. If there is no prior observation or the income component was also missing in the previous wave, the item remains missing or may be imputed using another method. Income fluctuations, inflation and growth/recession of the economy are all assumed to be non-existent. If item non-response increases, this method leads to serious bias. Accuracy suffers most when valid measurements are carried forward only from a long time ago or when income components are volatile. In the MetaSILC survey, the following countries report using this method: Belgium, Bulgaria, Czechia, Germany, Ireland, Croatia and Finland.

17.2.2. Uprating

Uprating is a method in which the last value is simply multiplied by a constant growth factor of the current period. This growth factor can be obtained using different approaches, including adjustment for inflation or by using an empirical factor that can be calculated, for example, from the change in mean income among complete cases. Again, if item non-response is considerable, and the calculation of the growth factor is not standardised, this leads to reduced comparability between countries.

However the uprating factor is derived, all methods suffer from the general drawback that only part of the total income dynamics is captured. In particular, all uniform uprating disregards the individual volatility of income. Using the same trend for every individual systematically underestimates (personal) income mobility. On the positive side, the method is again very simple, making the mishandling of data very unlikely and the method appealing. At the time of the 2018 MetaSILC survey, the method was reported to be in use in Austria, France, Italy, Spain and the United Kingdom.

17.2.3. Row-and-column imputation

This method (described in detail in Section 17.3) involves three elements: the overall trend over time, the average income level of the record to be imputed and the observed volatility from a similar case (nearest neighbour). Together, these elements ensure that mobility and correlation over time are restored, at least to some degree.

The row-and-column method is only slightly more complex than uprating or carrying forward information from previous waves. Only a few simple rules need to be followed, and it can be almost universally applied to all settings and income components without auxiliary information from outside the sample. The row-and-column method shares the same disadvantages as all single imputation approaches in that the variance of estimates will be underestimated if imputed values are treated as real. Since the method involves a stochastic element to account for mobility, it may however also be implemented as a method for multiple imputation.

Eurostat published a manual on business statistics (Eurostat, 2014) that recommends the row-and-column method for longitudinal imputation, notably in short-term statistics because of its balanced statistical properties and its simplicity, which renders it particularly suitable as a standard procedure in a harmonised statistical process. According to the MetaSILC survey on imputation and weighting, the row-and-column method is not applied in current EU-SILC practice.

The row-and-column approach is genuinely univariate, meaning that missing income variables are imputed independently from each other. Most current practices in longitudinal imputation are

univariate, and there is currently no evidence that cross-sectional imputation of EU-SILC variables would be applied multivariately in many countries. However, it has been demonstrated using wealth data that the simple row-and-column approach performs almost equally as well as more sophisticated algorithms (Grabka and Westermeier, 2014). The merit of being easily applicable for all countries without great risks of misspecification would need to be carefully considered against the drawbacks of a relatively simple univariate procedure.

The row-and-column method does not require any external data sources. It takes into account only the overall empirical trend within the sample and the usual (average) income level of an individual. This is theoretically superior to simply carrying forward information from the previous year, which assumes perfect stability without any change. When such imputations are applied on a massive scale, and there is overall income growth, imputed values will systematically lag behind, implying that imputed households will have a higher propensity to fall below a relative AROP threshold, which typically increases over time, at least nominally. The row-and-column method also does not simply assume that change against the previous year's income is uniformly distributed; instead, it usually assumes a modest rate of growth for everyone. This situation may be realistic only for certain recipients of transfers or for long-term employees whose salaries or pensions are uprated annually. This assumption is questionable, at least for people in precarious employment. Imputation from previous observations implies a stability that may not be true for these cases. Overall, uprating can thus be expected to underestimate the dynamics of income poverty.

17.3. A simple guide to 'row-and-column' imputation

The row-and-column method (Little and Su, 1989) uses information from all time points as well as the specific sample element for which imputation is necessary. The main advantage of the method is that it uses more information than simple cross-sectional imputation and is simpler than other methods of longitudinal imputation. In this section, some theoretical aspects of the method will be explained and its application will be demonstrated using fictitious income data.

17.3.1. Introduction

Little and Su (1989) provide two different approaches for their method. Depending on the imputation variable, one might use either the additive or the multiplicative formula (Eurostat, 2014, p. 3):

$$Imputation = (row\ effect) + (column\ effect) + (residual)$$

$$Imputation = (row\ effect) \times (column\ effect) \times (residual)$$

In the additive formula, the imputation might produce negative values (Eurostat, 2014, p. 3); therefore, only the multiplicative approach is applicable for most income variables in EU-SILC. Of course, one could impute the logarithm using an additive model.

In the following illustration, the observation of household i in period t_0 is to be imputed. For each observation, there are measurements y_{it} for different time points (or periods) t.

Both formulae consist of three elements: the row (individual) effect r_i, the column (period) effect c_t, and the residual e_{jt_0}. The row effect r_i is the mean of the equalised (observed) measurements of the household i – the one whose value will be imputed:

$$r_i = \frac{1}{m_i} \sum_{t=1}^{m_i} \frac{y_{it}}{c_t}.$$

The row effect is not affected by the period t_0 that will be imputed. It should be noted that, before taking the average, the income is divided by the column effect. This equalisation is especially helpful for data sets including extreme periods – both high and low extremes.

The column effect c_{t_0} of period t_0 is the ratio of the mean of period t_0 to the overall mean. Using this equalisation, the imputation will be adjusted to the current period.

$$c_{t_0} = \frac{\overline{y}_{t_0}}{\frac{1}{M} \sum_{t=1}^{M} \overline{y}_t}$$

The imputation also makes use of a so-called error term. The donor j of the error component is found by calculating the nearest neighbour of the row effects: $e_{jt_0} = \frac{y_{jt_0}}{r_j c_{t_0}}$. The calculation of the imputation value of course depends on the model used: $\widetilde{y_{it_0}} = r_i + c_{t_0} + e_{jt_0}$ for the additive model and $\widetilde{y_{it_0}} = r_i c_{t_0} e_{jt_0} = r_i c_{t_0} \frac{y_{it_0}}{r_j c_{t0}} = \frac{r_i}{r_j} y_{jt_0}$ for the multiplicative model.

17.3.2. Example

Because of its simplicity, Little and Su's (1989) method can be considered good practice for making best use of longitudinal information. Here, the method is explained using an example in the EU-SILC setting. Twelve households are in the sample, with one quarter (three observations) being in the sample in the first year, and another quarter being added in each of the the second, third and fourth years. Table 17.1 shows the observed income values. In this example, all prior item non-response is already imputed; there is only item non-response in year 4.

Table 17.1: Fictional raw data for EU-SILC setting

Household (i)	Income (y_{it})			
	$t=1$	$t=2$	$t=3$	$t=4$
1	220	250	265	270
2	570	500	510	
3	333	340	340	340
4		100	110	120
5		240	300	350
6		160	150	
7			400	
8			215	210
9			335	350
10				450
11				125
12				
Column means	374.33	265.00	291.67	276.88

NB: The table shows, for example, that household 1 participated in four waves of EU-SILC. All four incomes are observed. Household 2 also participated in four waves, but the last income is missing. Households 10 and 11 participated only in one wave.

To calculate row-and-column imputation for the four missing values (one per rotation group), the mean income for each survey year is calculated (last row) ([98]). In this example, it is only in the first year that we find an average income above the grand mean over the 4 years, indicating a decline in income over time. Hence, the first year is associated with a column effect greater than 1 ([99]). The column effects are displayed in the last row of Table 17.2.

In Table 17.2, the original income values from Table 17.1 are divided by the column effect to make the individual income values comparable. For a numerical example, let us consider the original value of household 9 in $t = 3$, which amounted to 335.00. Dividing this figure by the column effect of wave 3 (0.97) results in a slightly higher amount: 346.83. From these figures, we obtain the row effect as the mean of all equalised observations.

To impute missing values, a nearest neighbour is sought as a donor case. Little and Su (1989) suggested sorting the data to be sorted by their row effects (Table 17.3). The imputed value is then obtained as the product of the row effect r_i and the ratio of the empirical value ($y_j t_0$) and the row effect (r_j) of the next lowest observed value (household j).

For example, value for household number $i = 7$ is imputed (with $j = 9$ as donor for the error term) as: $(381.72/364.28) \times 414.13 = 433.96$. The observation of the nearest neighbour, namely household 9, is divided by its row effect and multiplied by the row effect of household number $i = 7$. In this case, the imputed value is slightly higher than the row effect, because the residual from the nearest neighbour is positive.

There are two special cases in which problems arise. Take, for example, household $i = 2$. The nearest neighbour is houssehold $j = 10$, which has only one observation. The row effect is therefore equal to the last equalised observation. In this case, the row effect of household $j = 2$ is used without any adjustment. However, in contrast to carrying the last value forward, the method will impute an equalised row effect, which is simply an average of all uprated previous income. The situation of the

([98]) The grand mean of the four wave-specific means is 301.97 ((374.33 + 265 + 291.67 + 276.88)/4).
([99]) In this example, 1.24 (374.33/301.97).

Table 17.2: Row-and-column effects for the fictional data

Household	Column-equivalised income (y_{it}/c_t)				Row effects (r_i)
	t =1	t = 2	t = 3	t =4	
1	177.47	284.88	274.36	294.47	257.79
2	459.81	569.75	528.01		519.19
3	268.63	387.43	352.01	370.81	344.72
4		113.89	113.89	130.88	119.57
5		273.48	310.60	381.72	321.93
6		182.32	155.30		168.81
7			414.13		414.13
8			222.59	229.03	225.81
9			346.83	381.72	364.28
10				490.78	490.78
11				136.33	136.33
12					
Column effects (c_t)	1.24	0.88	0.97	0.92	

NB: The table shows, for example, that the row effect of household 1 is 257.79. This household shows an average performance. The column effect of period 4 is 0.92. The mean of this period is below the overall mean. On average, people earn less money in this period than in general.

Table 17.3: Observed values in period 4 and imputed values using different longitudinal methods

Household	Income			
	t = 4 (observed values)	Last value	Uprating	Little and Su (1989)
1	270			
2		510	484.14	519.19
3	340			
4	120			
5	350			
6		150	142.39	168.81
7		400	379.72	433.96
8	210			
9	350			
10	450			
11	125			
12				
Column effects (c_t)				

NB: The table shows, for example, that the last value method gives an imputed value at t = 4 of EUR 510 for observation 2.

donor case used here holds generally for all observations of the second newest quarter with item non-response.

Household i = 12 shows the worst case. As there is no prior observation, no longitudinal imputation is possible here. Alternative cross-sectional methods have to be used in such a situation. The need to combine longitudinal imputation with other methods obviously complicates the process. In contrast to a lack of longitudinal information, prior observations of zero income are not a problem but will again result in an imputed value of zero.

17.4. Possible sensitivity of policy indicators to longitudinal imputation

The simulation is based on the 2013–2016 Austrian EU-SILC data with 500 randomly generated data sets. In this scenario, 15 % of cases for each of the 18 major income components are artificially set to be missing completely at random (MCAR) and to be independent of all other components. As a consequence, almost every household is affected by at least one missing component. Although an item non-response rate of 15 % is a fairly realistic assumption for many countries, it should not be assumed that, in reality, this rate is identical for each component of income or for every group in the population. The simulation can hence serve only as an illustration of the potential sensitivity; it cannot provide information on the true impact of the different imputation methods. Because countries have seen quite different growth trajectories, it is also plausible that effects may differ between countries. For example, carrying the last value forward could be a reasonable approximation when the income distribution is rather stable, whereas uprating would be a good fit when income is universally increasing. Only values in the last year of the observation period – 2016 – are set to missing. All other data are considered already imputed and therefore complete (see Chapter 15 for the detailed simulation set-up).

All missing values are imputed using the longitudinal method if possible. If the last value for the observation is zero, the information required for longitudinal imputation is assumed to be not available. Otherwise, the missing value is replaced by the median income of this component.

For the row-and-column method, records are accepted as donor values only if the last observation is different from zero, and there must be at least two values observed (otherwise the last observation and the row effect would cancel to 1).

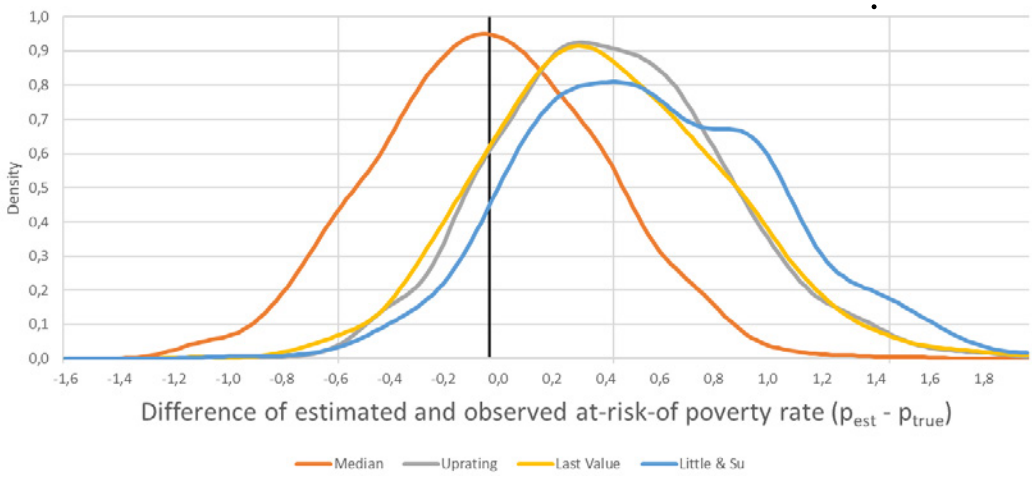
17.4.1. Potential imputation sensitivity of at-risk-of-poverty rate

After the imputation, equivalised income (in 2016) is calculated using all the imputed income components, and the AROP rate is calculated for each of the 500 (artificial) data sets. Thresholds are then recalculated for each of the 500 runs, using the original weights, to adhere as closely as possible to the reality of statistical production.

The focus here is on the aggregate impact on the level of indicators with a high level of policy relevance ([100]). Figure 17.1 gives a visual representation of the distribution of AROP rate estimates. Each curve represents the percentage point difference from the assumed true value ($p_{est} - p_{true}$) over 500 simulations.

([100]) An assessment of correlation coefficients revealed very little difference between methods. In principle, it would also be possible to inspect accuracy at the individual level in the form of some confusion matrix that compares predicted and the true values.

Figure 17.1: Comparison of estimated AROP rates (15 % item non-response rate and MCAR), Austria, 2016



NB: The true value of the official Austrian AROP rate in 2016 was 14.1 %. The maximum of the density curves represents the expected difference from this true value; the spread reflects the uncertainty of estimates when 15 % of each income component needs to be imputed.

Source: Authors' computation based on data from EU-SILC2010–2016 (retrieved 1 October 2018).

Most of the results are relatively close to the true value (14.1 %). The imputation of median income for each component does not alter the poverty threshold, and tends to overestimate the poverty rate in approximately half of the cases and underestimate it in the other half. The simulation of a completely random missing pattern would therefore suggest that median imputation leaves the AROP rate unbiased.

It is important, however, to acknowledge that a scenario with 15 % random item non-response on each component introduces considerable uncertainty to the estimates. Different results occur with each imputation. Variance across the 500 different AROP estimations varies only slightly between methods, from 0.17 for uprating and cross-sectional imputation to 0.19 for last value imputation and 0.21 for the Little and Su (1989) method. For comparison, the standard error published by Statistics Austria is 0.7 (variance of 0.49).

The results suggest that, whatever method is used, imputation errors add about one fifth to the value of the sampling variance, reducing the effective sample size accordingly. Therefore, as a priority, missing information needs to be avoided, considering the cost in precision attached to it.

With regard to an assessment of longitudinal methods for the full sample, it is noteworthy that only about 30 % of all missing values can be imputed by using longitudinal data. Here, the remaining cases were imputed using the median value. Median imputation was chosen as a neutral method that allows the inclusion of all observations for all methods in a computationally intensive simulation set-up. With 70 % of missing cases imputed using a cross-sectional method, it is clear that the potential impact of longitudinal imputations must be limited. However, the idea of including all cases does justice to the conditions in practice under which longitudinal imputation will inevitably have to be combined with other methods.

Different uses of longitudinal information affect the accuracy of estimates in only slightly different ways. The dispersion of last value carried forward and uprating is almost the same as and somewhat smaller than that of the row-and-column method. This is a plausible result, as the row-and-column method has more parameters, including some residual parameters. It can be noted that, in the Austrian simulation data, all the approaches tend to somewhat overestimate AROP rates. The discrep-

ancy appears to be largest for the row-and-column method.

All longitudinal imputation assumes a certain continuity over time, which can be problematic when imputation is performed at the component level and the principal source of income changes. Income insecurity is a typical characteristic of vulnerability, and it may be possible that a person who has a low income in one year due to job loss may be on social transfers in the next year. Hence, the assumed correlation over time or stability was apparently not completely accurate for the situation in the Austrian data – at least at the level of income components.

Because the potential impact of longitudinal imputations is expected to be mostly relevant for mobility patterns, we turn to the sensitivity of persistent poverty risk in the following section.

17.4.2. Potential sensitivity of persistent at-risk-of-poverty rates to imputation

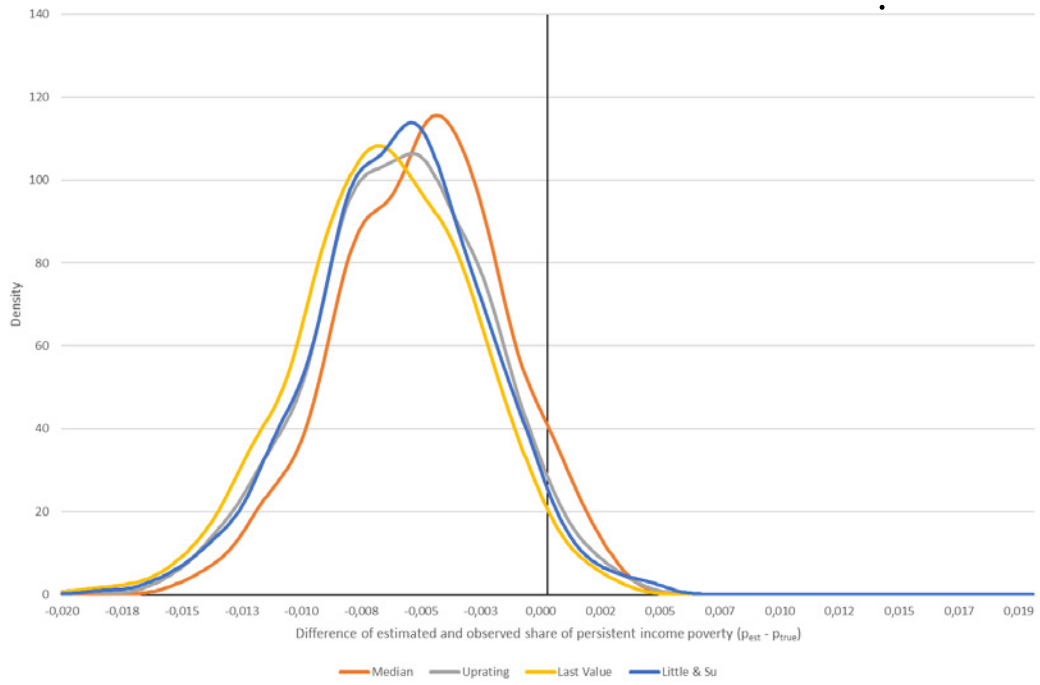
Figure 17.2 presents simulation results for the persistent AROP rates in 2016 – with AROP defined in the EU indicator framework ([101]) as people at risk of poverty – and also in at least two of the three previous years. The Austrian EU-SILC panel rotation that was in the sample between 2013 and 2016 is used for this analysis, comprising 2 421 people. To simplify the simulation, only values from 2016 have been imputed, whereas for the previous years the known 'true' poverty status has been used. The impact of imputations is therefore limited to the final observation. We refrain from any inspection of individual imputed values and base the assessment solely on the estimates for the final indicator.

For this indicator, all approaches tend to underestimate the true value very slightly. The uncertainty introduced is not large either. Although this simulation set-up did not consider any longitudinal patterns that may be observed in reality, median imputation turns out to produce some bias; however, it is a bit smaller than that of the longitudinal approaches. Carrying last values forward appears to be the least appropriate assumption in this simulation.

The limited variance in this simulation is likely to reflect the fact that previous years have been assumed to be measured by the true value. In essence, the approach aims to preserve the cross-sectional distribution only, which favours cross-sectional imputation. If the simulation had allowed missing values to occur in all waves and imputed them using all of the longitudinal information available regardless of whether it was collected prior to or after the occurrence of missing income, the result may have been different. Creating a more appropriate simulation set-up is complex, involving imputations of $t - 3$ feeding into imputations of $t - 2$ and those feeding into $t - 1$ and finally t.

([101]) The poverty status in each year is determined by comparing equivalised income of a household with the official cross-sectional poverty threshold (see https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology_-_monetary_poverty#Calculation_method).

Figure 17.2: Differences between estimated and observed persistent AROP rates (4 years, 15 % item non-response rate and MCAR), Austria, 2016



NB: The differences between estimated persistent income poverty rates and the true observed value vary only between – 0.018 and 0.006 (i.e. – 1.8 to + 0.6 percentage points).

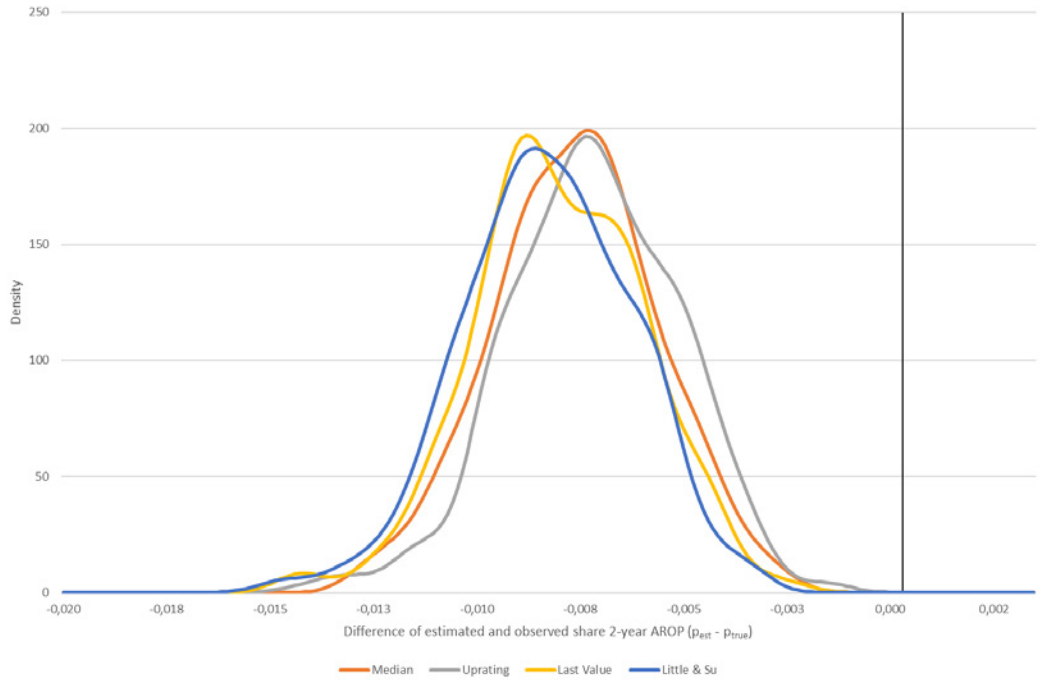
Source: Authors' computation based on data from the 2010–2016 EU-SILC (retrieved 1 October 2018).

17.4.3. Potential imputation sensitivity of 2-year at-risk-of-poverty rates

The effect of the method on the estimated persistence of income poverty over two consecutive years is shown in Figure 17.3. As before, only the income values in 2016 were imputed, whereas for 2015 the original income poverty status was used as the true value. The database slightly changed: all observations that entered EU-SILC in 2013–2015 ($n = 8\,524$) were used.

The method of using last values with or without uprating tends to underestimate a 2-year income poverty rate. On the other hand, a purely cross-sectional approach using median imputation also appears to underestimate longitudinal income poverty. However, in this simulation, the theoretical advantage of longitudinal methods cannot be confirmed.

Figure 17.3: Differences between estimated and observed persistent income poverty rates (2 years, 15 % item non-response rate and MCAR), Austria, 2016



NB: The differences between estimated persistent income poverty rates and the true observed value vary only between – 0.015 and 0.
Source: Authors' computation based on data from the 2010–2016 EU-SILC (retrieved 1 October 2018).

17.5. Recommendations on good practice

The avoidance of missing income information and investment in the convergence of timely imputation strategies are essential to make EU-SILC more accurate and comparable between countries. Precision loss due to missing data and inappropriate imputation can be put in the context of sample size requirements. The essential formula is simple: a smaller sample without missing values may yield the same precision as a larger sample with considerable item non-response. Hence, it is possible to translate variance into survey cost, attached to sample size. It pays to invest in avoiding item non-response in the first place. This is particularly relevant to statistical offices, which face constant resource constraints. Compared with the potential precision gain when more complete data are collected, the potential of imputation as an *ex post* remedy can only be limited.

In the governance mechanisms of the European Statistical System, compliance is achieved when output variables are harmonised and basic requirements regarding sample size and documentation are met. In particular, data may be collected by interviews or from registers. Consequently, there are no common fieldwork protocols such as in international survey programmes, such as the European Social Survey, the Survey of Health, Ageing and Retirement or the Organisation for Economic Co-operation and Development's Programme for the International Assessment of Adult Competencies (Glaser et al., 2015). Against this background, it appears unrealistic to recommend any strategy that will eliminate item non-response. Missing items may be considered a distinctive feature of income surveys, considering the sensitivity of this topic in some countries, but also the need to aggregate

income components over many detailed items as well as the need for editing procedures, which may lead to the deletion of erroneous information.

At the very least, it should be considered that the impacts of imputation – especially on the measurement of income poverty dynamics – may be considerable. Using purely cross-sectional information implicitly assumes that income is uncorrelated over time. Generally, it is therefore advisable to make use of longitudinal information whenever possible. Even if our own simulation results, which were adjusted specifically for the EU-SILC context, were not particularly conclusive on whether bias could be reduced when missing patterns are completely random, the effect of longitudinal imputation is well documented in an Australian simulation study. Evaluating eight approaches to longitudinal imputation on different income components collected in the Household, Income and Labour Dynamics in Australia survey, the authors found that the imputation method matters, especially for estimates of change (Watson and Starick, 2011).

One option is to use longitudinal variables in combination with cross-sectional variables. This, however, would require substantial efforts in model specification, as the number of potentially useful variables is considerable. Standardisation would be feasible only at the level of target variables, whereas missing patterns are likely to be specific to certain subcomponents that are collected at national level. There is little doubt that such standardisation would come at a substantial cost to country specificity. Generally, it does not appear advisable to centralise the processing of national income data, as that may negatively affect the usefulness of data in the national context. Consider, for example, the need to reflect the national specificities in the tax–benefit systems when executing imputations of missing income. At the same time, it is obvious that some level of standardisation is essential to ensure comparability.

The choice of a method for longitudinal imputation has potential consequences when comparing measures of stability or change, such as the persistent poverty rate. A particularly popular option is to simply carry values forward from previous years. As with all deterministic imputation approaches, this will artificially suppress variance. Even when uprating is applied, an overestimation of persistent income poverty rates is to be expected. This is simply reflecting the fact that mobility would occur only if the factor used for uprating income values substantially exceeded the increase in the AROP threshold over time. As the threshold will typically follow a similar (or identical if it is derived from the increase in median values) path to the uprating factor, this is rather unlikely. Depending on the amount of missing information and the imputation method, biased estimates of the persistent poverty rate could be expected. Cross-sectional imputation will tend to inflate mobility, whereas last value carried forward will tend to overestimate persistence.

The row-and-column method includes a stochastic element that does allow for a certain degree of mobility. This makes it the preferred method when a simple and effective strategy is sought. Its application will enable all countries to restore correlation patterns over time in a robust manner. The row-and-column method and certain variants were also found to perform particularly well in the Australian simulation study mentioned above (Watson and Starick, 2011).

In the present chapter, only univariate imputation was considered. Many income components will be correlated in some way; for example, if both employment income and salary income are received, the values of each are unlikely to be independent. In addition, income may affect means-tested social benefits. However, although multivariate imputation would be desirable on theoretical grounds, present practice appears to suggest an emphasis on incremental improvements in using longitudinal information for univariate imputation of missing income information. In any case, research into practicable imputation strategies appears worthwhile, provided that all efforts have already been made to avoid missing income information.

References

de Leeuw, E. D., Hox, J. J. and Huisman, M. (2003), 'Prevention and treatment of item nonresponse', *Journal of Official Statistics*, Vol. 19, No 2, pp. 153–176.

Eurostat (2014), 'Method: Little and Su method', in Eurostat, *Memobust Handbook on Methodolo-*

gy for Modern Business Statistics (https://ec.europa.eu/eurostat/cros/system/files/Imputation-06-M-Little%20and%20Su%20v1.0_7.pdf).

Frick, J. and Grabka, M. (2007), 'Item non-response and imputation of annual labor income in panel surveys from a cross-national perspective', *IZA Discussion Papers*, No 3043, IZA Institute of Labour Economics, Bonn.

Gelman, A. and Hill, J. (2006), 'Missing-data imputation', in Gelman, A. and Hill, J. (eds), *Data analysis using regression and multilevel/hierarchical models (analytical methods for social research)*, Cambridge University Press, Cambridge, pp. 529–544.

Glaser, T., Kafka, E., Lamei, N., Lyberg, L. and Till, M. (2015), 'European comparability and national best practices of EU-SILC: a review of data collection and coherence of the longitudinal component', *Net-SILC2 Working Papers*, No 5/2015, Statistics Austria, Vienna.

Grabka, M. and Westermeier, C. (2014), 'Persistently high wealth inequality in Germany', *DIW Economic Bulletin* Vol. 6, pp. 3–15.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd edition, Wiley, New York.

Little, R. J. A. and Su, H.-L. (1989), 'Item non-response in panel surveys', in Kasprzyk, D., Duncan, G. and Singh, M. P. (eds), *Panel Surveys*, Wiley, New York, pp. 401–425.

Raghunathan, T. E., Lepkowski, J. L., Van Hoewyk, J. and Solenberger, P. (2001), 'A multivariate technique for imputing the missing values using a sequence of regression models', *Survey Methodology*, Vol. 27, No 1, pp. 85–95.

UNECE (United Nations Economic Commission for Europe) (2020), *Poverty Measurement: Guide to data disaggregation*, United Nations, New York and Geneva (https://unece.org/sites/default/files/2021-01/ECECESSTAT20204.pdf).

Verma, V., Gagliardi, F. and Ferretti, C. (2010), 'Cumulation of poverty measures to meet new policy needs', in *Proceedings of the Italian Statistical Society 2010*, Italian Statistical Society, Padua

Watson, N. and Starick, R. (2011), 'Evaluation of alternative income imputation methods for a longitudinal survey', *Journal of Official Statistics*, Vol. 27, No 4, pp. 693–715.

Comparability and validity of measures

18 Lessons and recommendations regarding the comparability of the EU-SILC income variables

Tim Goedemé and Lorena Zardo Trindade ([102]) ([103])

18.1. Introduction

The European Union Statistics on Income and Living Conditions (EU-SILC) are currently the most important comparative microdata on household incomes in Europe. The survey is regularly used for studying income patterns, poverty and income inequality in the EU. It is also an important source of information for studying the impact of tax–benefit policies on income distribution and for carrying out *ex ante* policy evaluations of planned policy reforms (see, for instance, Atkinson, Guio and Marlier, 2017, and Chapter 2 of this book for an introduction to EU-SILC). In other words, cross-national policy learning and monitoring of poverty and inequality are important purposes of EU-SILC. As a result, the cross-country comparability of EU-SILC and, in particular, its measurement of incomes are key. Therefore, in this chapter, we evaluate several factors that could affect the comparability of the income variables included in EU-SILC. We look at both factors affecting the comparability of the aggregate income variables (i.e. those related to total household income) and factors affecting the comparability of more disaggregated income variables. While the total income variables are most often used in comparative studies of poverty and inequality in the EU and are the basis for the EU's social indicators related to income, the disaggregated variables are very important for policy evaluations and recommendations as well as cross-national learning. Many stages in the life cycle of a survey may affect comparability. In this chapter, we limit ourselves to reviewing the definition of the target variables, compliance with these definitions when constructing the target variables, and variations in how the underlying data are collected.

On the basis of a survey among national statistical institutes, an analysis of the national quality reports and the comparative quality reports, the national EU-SILC questionnaires and an analysis of the EU-SILC data, we compiled a database – MetaSILC 2015 – that documents the exact classification of income components into the EU-SILC target variables (Goedemé and Zardo Trindade, 2020a). The focus of the database is on the 2015 EU-SILC, covering 26 EU-SILC countries. The database contains information on the composition, source (survey versus register) and way of collecting (gross or net) the variables on total income before and after transfers; income from benefits, work and capital; and social contributions and taxes. Special attention is given to self-employment income, imputed rent and income from production for own consumption, as well as outlier detection and data error correction. The database was constructed in the context of the Net-SILC3 project and is freely available ([104]). In

([102]) Tim Goedemé and Lorena Zardo Trindade are at the University of Antwerp. The authors are grateful to Jeroen Horemans, Adeline Otto, Tess Penne and Wim Van Lancker for their contribution to the background report of the MetaSILC 2015 project; and to Anne-Catherine Guio, Cristina Lopez-Vilaplana, Eric Marlier, Lars Lyberg, Peter Lynn and Teresa Munzi for their most helpful comments and suggestions. In addition, they would like to thank all those involved in responding to the Net-SILC3 questionnaire they circulated to collect the information for MetaSILC 2015. All errors are the authors' responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Tim Goedemé (tim.goedeme@uantwerpen.be).

([103]) This chapter is a shortened and slightly revised version of Zardo Trindade and Goedemé (2020).

([104]) https://timgoedeme.com/tools/metasilc-2015/

addition, we compiled a detailed report that discusses for each income variable the results of the analysis of the database, and general limitations as regards comparability (Goedemé and Zardo Trindade, 2020b).

Rather than analysing the comparability of each variable separately (as is done in the detailed report), in this chapter we focus on some general conclusions with regard to the current state of procedural comparability in terms of collecting the income target variables, and formulate some recommendations about how comparability could be improved in the future. We stress that assessing the comparability of the income variables involves an element of subjectivity, especially when sufficient background information is lacking. In addition, many more factors than those discussed in this chapter may undermine cross-country comparability. While we focus on measurement and processing errors, other factors that may affect comparability include, for instance, variations in coverage errors of the sampling frame, non-response bias, and imputation and weighting strategies (see also Verma and Betti, 2010; Eurostat, 2016a; Di Meglio et al., 2017; and Chapters 3–17 of this book).

This chapter is organised as follows. First, we briefly describe the purposes, set-up and contents of the 2015 MetaSILC database. Subsequently, we report on the main factors that may undermine the comparability of the income target variables and the composite total income variables. We conclude with a brief summary of the main challenges, areas for improvement and some recommendations.

18.2. MetaSILC 2015

Many factors potentially undermine the comparability of income data in EU-SILC (Iacovou, Kaminska and Levy, 2012). From the perspective of total survey error, one could make a distinction between the definition of the target population, quality of the sampling frame, non-response bias, adjustment errors, construct validity, measurement errors, response errors and processing errors (e.g. Groves et al., 2009, and, with an extension to comparative research, Pennell et al., 2017). When comparing two populations, one hopes that only the target population differs (if defined in a 'comparable' way), and that otherwise all errors play out similarly in both populations. In this context, it is useful to define comparability somewhat more precisely. In this chapter we look only at 'procedural comparability': the extent to which the same procedures are used for constructing a variable in various countries or years (Goedemé et al., 2015). This should be distinguished from 'substantive comparability', which implies that the same phenomenon is captured in a similar way across time, subpopulations or countries. In contrast to procedural comparability, substantive comparability should be assessed with reference to the purpose of the analysis, and is more demanding. For instance, if a specific income component is not important in one country but it is in another (e.g. production for own consumption), leaving the component out of the survey in both countries complies with the principle of procedural comparability but not necessarily with substantive comparability, depending on the objective of the study. We limit ourselves to assessing to what extent procedures for measuring and processing the income variables vary across countries, under the assumption that for many research purposes cross-national variation in these factors potentially results in a limitation of substantive comparability too.

Several documents are available to assess procedural comparability of the EU-SILC income variables, including the methodological guidelines, national quality reports and the comparative quality reports. Nonetheless, information is not always entirely complete and is sometimes contradictory. For instance, it is often not clear how exactly each of the national income components is classified and aggregated into a target variable in practice, and how the data were collected. Therefore, we built MetaSILC 2015, an accessible database that allows both EU-SILC producers and EU-SILC users to easily find more information on the content, classification and comparability of the income variables.

The MetaSILC 2015 database was set up on the basis of two rounds of consultation with the national statistics institutes (NSIs), the data producers of EU-SILC. The purpose of the consultations was to gather detailed information on the collection, processing and aggregation of EU-SILC income

components. Eurostat invited the contact person for EU-SILC in each NSI to participate in an online survey, with each email invitation containing a customised link to an online questionnaire focusing on the 2015 EU-SILC cross-sectional wave. Data collection took place from July 2016 to January 2017. The first round focused on questions on the composite variables and income from benefits, while the second round focused on income from work and other sources. Complementary information to reported income components was collected over the course of 2017, 2018 and 2019. More detailed information on the data collection, as well as the questionnaire itself, are provided in Goedemé and Zardo Trindade (2020b).

The online questionnaire on income variables was divided into three sections: (i) variables on total income before and after transfers; (ii) variables on income from benefits; and (iii) variables on income from work and other sources, social contributions and taxes. Since 2014 most (not all) EU-SILC countries have collected more detailed target variables related to benefits, making a distinction between benefits that are (i) contributory and means-tested, (ii) contributory and non-means-tested, (iii) non-contributory and means-tested and (iv) non-contributory and non-means-tested. When applicable, we collected information for each of these more detailed benefit variables. In this chapter, we focus only on the income variables that are used to construct total disposable household income. Overall, the response rate for completed surveys was 76 % (26 of 34 countries). Twenty-four countries participated in both rounds of the survey (Table 18.1).

In addition to the information provided by NSIs in the questionnaire, the MetaSILC 2015 database was supplemented with information from the national quality reports, the comparative quality reports and the national EU-SILC questionnaires as well as various other sources, including the European system of integrated social protection statistics (ESSPROS) (Eurostat, 2016b), the Mutual information system on social protection (MISSOC, 2015) and the Euromod country reports ([105]).

MetaSILC 2015 maps the exact classification of income components onto the EU-SILC target variables. An income component should be understood as a specific source of income, which typically is much more disaggregated than an income target variable (e.g. child benefits for civil servants, family benefits for employees of small and medium-sized enterprises, and maternity benefits are three examples of income components that are part of the target variable 'family-/children-related allowances'). For each of the income components, the database contains the official name (national language) and code in the national EU-SILC survey, the equivalent name in English, the target variable code and name, the source of the income information used (register data, questionnaire, imputation), the level of aggregation when it was collected, information on gross–net conversion, whether there were important changes between the 2010 wave and the 2015 wave, and if there are important changes planned for future waves. The database allows researchers to identify which income components are covered in EU-SILC and how they were classified into the EU-SILC target variables. In addition to the database, we compiled a report

Table 18.1: Countries' participation in the MetaSILC 2015 survey

Participation status	Countries
Both rounds	Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, France, Germany, Greece, Hungary, Italy, Latvia, Luxembourg, Malta, the Netherlands, Poland, Serbia, Slovakia, Slovenia, Spain, Sweden and the United Kingdom
Only first round	Portugal
Only second round	Finland
No information	Iceland, Ireland, Lithuania, North Macedonia, Norway, Romania, Switzerland and Turkey

Source: Information collated by the authors on the basis of MetaSILC 2015.

([105]) https://euromod-web.jrc.ec.europa.eu/resources/country-reports

that discusses for each income variable the results of the analysis of the database, looking at compliance with Eurostat guidelines (Eurostat, 2016a), misclassifications and omitted income sources that could undermine cross-national comparability (Goedemé and Zardo Trindade, 2020b). In this chapter, we summarise the main findings with regard to the aggregated income variables and the target variables that are used for computing these aggregated variables.

18.3. Findings

In what follows, we first consider the individual target variables. Subsequently, we look in more detail at the composite income variables that refer to various concepts of total household income.

18.3.1. Challenges to comparability of the income target variables

We focus on the definition of the variables, non-compliance with Eurostat guidelines, the level of detail of data collection, the source of the data, and the collection of net versus gross incomes.

The definition of the variables

We identified three challenges to the definition of income variables: (i) the need for more precise definitions and to define target variables more clearly in a mutually exclusive way; (ii) the need to make sure that all countries provide the benefit variables with the same level of detail; and (iii) the potential for and usefulness of providing the benefit variables at an even more disaggregated level.

Even though there are many commonalities between European countries, tax systems and social benefits differ greatly. As a result, the definitions of the target variables should be sufficiently generic to cover all sources of income with a similar function. However, they should also be sufficiently specific that data producers can easily understand which income source should be aggregated into which target variable. Our conclusion is that often the Eurostat methodological guidelines (Eurostat, 2016a) and description of EU-SILC target variables leave too much room for interpretation, resulting in what could be called 'borderline cases': some income sources might be classified under the heading of at least two variables. This leads to comparability problems, as countries do not always make the same choice. These borderline cases are especially common with respect to the following variables.

- Support for bearing the costs of rent, gas, electricity, heating, water and utility bills, that is, all compensation for housing costs can be found in the variable on housing allowances (HY070), but Eurostat guidance is not clear about whether it is wrong to also include these in the variables on social benefits not elsewhere classified (HY060) and old-age benefits (PY100). Bulgaria and Greece include help with heating costs under HY060, while Denmark and the United Kingdom include it under PY100.
- HY145 contains payments and receipts for tax adjustments, but its definition leaves room for interpretation and, as a result, Estonia, Spain, Croatia, Austria, Poland and Slovenia opted for recording tax adjustments under both HY145 and HY040 (variable on income taxes and social contributions), while Belgium included adjustments only under HY145.
- The variable on social exclusion benefits not elsewhere classified (HY060), due to its very generic definition, often contains components that could be included in other benefit-type variables. There are relatively many borderline cases involving HY060 and other benefit variables (PY100, PY130 (disability benefits) and PY140 (education-related allowances)). The new disaggregation of target variables relating to benefits (making a distinction between means-tested and non-means-tested benefits) may be helpful in clarifying the definition of HY060 and the categorisation of benefits, by restricting HY060 only to those benefits that cannot be classified as means-tested versions of the other income target variables.

Apart from the lack of precision, another challenge of the benefit variables is that some countries provide them with more detail than other countries do. In particular, social benefit variables are now

disaggregated on the basis of eligibility criteria related to paying contributions and the existence or not of a means test. Since 2014, several countries have started to provide some of the benefit variables in disaggregated format. In 2015, with the exception of Latvia, Poland and Sweden, most of the countries under study applied the disaggregation. For comparative purposes, it would be very beneficial if all countries could apply the same level of disaggregation as soon as possible, and preferably going back to at least 2014.

Although this further disaggregation is a major improvement for EU-SILC users, allowing a much more refined analysis, the usefulness of EU-SILC could be further improved by providing information on benefits at an even more disaggregated level. Income from benefits is reported in EU-SILC following the eight functions of social protection defined by ESSPROS. From a policy point of view, the eight functions still group together very different types of benefits, with different functions, into one category, precluding a more detailed analysis. For example, old-age, disability and survivor benefits include benefits other than pensions (e.g. care allowances and other cash benefits as compensation for housing costs), precluding a clear-cut analysis of pensions using EU-SILC, the policy relevance of which is hard to overstate. In addition, one cannot disentangle the impact of regular cash support from individuals other than household members from remittances (regular cash support from households in other countries); or the impact of maternity benefits from child benefits; or the impact of educational private transfers, often merit based, from educational public social benefits (either merit based, income based or universal); or the impact of taxes from social contributions. As far as income from benefits is concerned, researchers wishing to carry out a more detailed analysis of these specific income sources are required to make use of (partially) simulated data, for instance data provided by Euromod (e.g. Sutherland and Figari, 2013). Although, strictly speaking, countries apply in these cases the same level of aggregation, for many types of analysis the different composition of these variables will undermine comparability in a substantive sense, and especially so when data users are not fully aware of the different kinds of income sources that are lumped together.

Misclassifications and omissions of income components

It is often difficult to judge whether countries are compliant or not with the Eurostat definitions, given the lack of clarity highlighted earlier. Therefore, in what follows we restrict ourselves to giving some examples of clear non-compliance, relating to 'misclassified' income sources, omitted income components (which may also affect the total income variables) and other potential 'errors'. Examples of income components that have been 'misclassified' by at least some countries include maternity benefits, spousal maintenance or support paid by the government, carer's allowance, death grants, funeral expenses, wages paid to oneself, income from own consumption, income in kind in the form of a company car, pension or annuities received in the form of interest or dividends, and income from individual private insurance plans. To give one example, payments for fostering children are understood differently among the countries included in MetaSILC 2015. Eurostat guidelines indicate that this type of income should be treated as employee cash or near-cash income (PY010), but most countries do not seem to treat it that way. Only Croatia, Greece and Serbia explicitly reported this type of income under PY010. In contrast, payments for fostering children are treated as family-/children-related benefits (HY050) in the case of Bulgaria, Germany, Cyprus, Latvia, Luxembourg Malta, Poland and Slovakia. Another important example is income in the form of a company car (PY021), which in France and Austria is recorded under employee cash or near-cash income (PY010) instead of PY021.

We have also observed some cases in which some income and tax components (e.g. maternity benefits, payments for fostering children, fringe benefits, land taxes and tax credits) do not seem to have been included in any of the EU-SILC target variables. This is the case for Denmark, which does not register the land value tax (a municipal tax on the land value of residential property), and Spain, which omits real estate tax and urban real estate value tax. In our understanding, both taxes could have been included under regular taxes on wealth (HY120). If the taxes were assessed on holdings of property, land or real estate, when these holdings are used as a basis for estimating the income of their owners, one could also argue that they could

have been included under tax on income and social insurance contributions (HY140G). However, this was not observed in any country. Malta and Slovakia do not include information on tax credits for taking parental leave in any target variable. As tax credits for taking parental leave can be considered as benefits received in addition to a salary for bringing up children, they should probably be included under family-/children-related allowances (HY050). Many countries (Austria, Belgium, Denmark, Estonia, Finland, France, Hungary, Luxembourg, Slovenia, Spain, Sweden and the United Kingdom) also did not report where payments for fostering children are allocated. This component, which should be treated as employee cash and near-cash income, is not included under PY010 or in any other variable. The countries for which this omission was confirmed include, for instance, Italy and the Netherlands. This is also the case for fringe benefits: Denmark clearly omitted them from PY010 and does not include them elsewhere.

Besides the misallocation and omission of income components, other potential inconsistencies were observed in the computation of income target variables. Eurostat guidelines (Eurostat, 2016a) are sometimes very specific regarding the components that should be included or excluded from a variable, or if a variable should be filled or not. With regard to the variable for income from rental of a property or land (HY040), the guidelines are clear when defining that costs such as mortgage interest repayments, minor repairs, maintenance, insurance and other charges should be deducted from income from rental of a property or land. However, some countries (e.g. Italy and Luxembourg) do not deduct such costs from the values reported, while for quite a few other countries it is not entirely clear whether these costs are deducted or not.

The level of detail of the data collection

Even when countries comply with the Eurostat guidelines for defining a target variable, comparability may be undermined because of the different level of detail with which data are collected. In principle, one could expect that measurement errors are minimised when respondents are asked about their income at the most detailed level (i.e. by source of income). However, sometimes respondents may more easily recall their total income than the exact level of each individual income component. In any case, it is likely that variations in the level of detail with which income data are collected may affect comparability. Although the 'optimal' level of detail may vary across countries, it is very unlikely that the variations identified below are such that they maximise the quality of data collection in each country and optimise comparability.

Variables that are collected with strongly varying levels of detail include income from interest, dividends and profits from capital investment in an unincorporated business (HY090), pensions from individual private plans (PY080), housing allowances (HY070), survivor benefits (PY110), sickness benefits (PY120), disability benefits (PY130), education-related allowances (PY140), income from rental of a property or land (HY040), income received by people aged under 16 (HY110), cash benefits or losses from self-employment (PY050), regular inter-household cash transfers received (HY080) and tax on income and social contributions (HY140).

Typically, when data are collected from registers, information can be collected for many income components separately, and the problem of under-/over-reporting is equal to the problem of mistakes in administrative data and problems related to measuring income from the grey/black economy (non-taxable incomes are usually collected through the questionnaire). The general problem is then an issue of survey versus register data (see below). However, for countries that collect the information on the basis of a survey, EU-SILC questionnaires vary greatly in the extent to which they ask for each income source separately, or in whether or not they list extensively the different types of income sources to be added up for an aggregate question. Furthermore, some countries collect more income components at the individual level before aggregating them to the household level, whereas others collect them through the questionnaire from the main respondent only. Finally, not every country provides the opportunity to report the income amount on the basis of prespecified income intervals, even though this is often recommended in the Eurostat guidelines. This may lead to different patterns of non-response and non-re-

sponse bias. There is probably some room here for further input harmonisation. Having a more refined template that countries should use by default, unless they have good reasons for doing it in a different way, would probably be an important step forward in terms of comparability.

A good example to illustrate this point is the variable on income from interest, dividends and profits from capital investment in an unincorporated business (HY090). Questionnaires differ with regard to (i) whether or not the question is asked only of the main respondent (e.g. Belgium) or of each adult separately (e.g. Greece); (ii) whether or not examples of financial products are listed, and in the types and number of examples given; and (iii) whether there is one aggregate question or separate questions by type of financial product. When questions are asked separately by type of financial product, the selected types of products vary by country. It should be noted that the Eurostat guidelines mention explicitly that respondents should be given the opportunity to choose their answer from a range of values rather than giving the exact amount. It is not entirely clear to us whether this option was offered everywhere.

Survey versus register data

As also highlighted in the previous section, EU-SILC combines information from survey and register data. While most countries use only survey data, an increasing number combine register data with survey data or use only register data. It is important to stress that variations in the extent to which countries rely on register data exist not only between countries in terms of the scope of data collected, but also within countries over time. Even within target variables the use of survey and register data is sometimes combined. Of the countries included in our study, only Denmark and Sweden collect all their income variables from registers. Estonia, Spain, France, Italy, Cyprus, Latvia, Malta, the Netherlands, Austria, Slovenia and Finland use registers and/or a combination of register and survey data to construct some income variables (i.e. in the 2015 EU-SILC). The number of countries that use register data is increasing; for instance in Belgium the use of registers is maximised since the 2019 EU-SILC. In addition, the use of imputation techniques to compute income target variables is also observed among various countries for seven different target variables: social exclusion benefits (HY060), housing allowances (HY070), old-age benefits (PY100), tax on income and social contributions (HY140), family-/children-related allowances (HY050), employee cash or near-cash income (PY010) and income in the form of a company car (PY021).

The main issues in using register data compared with survey data have been discussed by Jäntti, Törmälehto and Marlier (2013), Törmälehto, Jäntti and Marlier (2017) and a number of other studies that analyse how the use of register data compared with survey data may bias conclusions from comparative research (e.g. Rendtel et al., 2004; Lehmann, 2011). Certainly, more research is necessary to evaluate the impact of the variety of data collection approaches in EU-SILC on different research questions. However, it is clear that sound conclusions from comparative analyses require an awareness that some results are probably driven by the data collection approaches used. One way to make data users more aware of the potential impact of the difference between survey data and register data is to create a repository in which all countries that switch at some point from survey to register data deposit their preparatory validation studies. Given the increased and planned use of register data in many countries, Eurostat should continue to encourage countries to evaluate the change over the course of several years (by collecting data from both registers and surveys) before survey data are replaced by register data, and to publish and disseminate the results of these assessments (e.g. Statbel, 2018).

Collecting net versus gross amounts

EU-SILC guidelines suggest that countries report 'total gross income' and 'gross income at component level'. However, as some countries still do not collect gross income values, data on income components are often collected net of taxes and/or of social contributions, and gross values are imputed (Eurostat, 2016a). Because of this difference in the types of values collected, the quality of information is not uniform across countries and issues with comparability may arise. In addition to the lack of uniformity in the types of data collected, it is important to mention that the lack of uniformity in

how to treat tax credits and the different approaches for converting net values into gross are also potential challenges for cross-country comparability. Belgium, Bulgaria, Czechia, Estonia, Greece, France, Croatia, Italy, Latvia, Luxembourg, Hungary, Austria, Poland and Sweden make use of gross–net or net–gross conversion techniques, as they do not collect both gross and net values for most income target variables.

MetaSILC 2015 did not collect information on the conversion approaches used. However, EU-SILC national quality reports include descriptions of the approaches used to convert net values into gross values. Some of the approaches used are (i) microsimulation models taking into account either withholding or final taxes, (ii) statistical methods and (iii) matching survey data with register data (see also Chapter 16 of this book). Therefore, more consistency in the methods used for converting net incomes to gross incomes may reduce problems with cross-country comparability. In addition, including tax credits in a separate variable and providing income taxes separately from social security contributions may also contribute to better conversion outputs.

18.3.2. Comparability issues with regard to total household income

Many studies and EU social indicators do not rely on the disaggregated income target variables but make use of the composite variables that relate to total household income instead. This is, for instance, the case for the at-risk-of-poverty indicator (with and without social transfers) and the indicators on income inequality produced by Eurostat. Gaining more insight into the comparability of these variables is therefore of paramount importance, even though it is hard to assess their overall comparability. In what follows we report on what we think can be learned from the information collected in MetaSILC 2015 and our multiple exchanges with NSIs.

Of the 26 countries that responded to the survey, all except Finland completed the questions on the composite income variables. According to the information reported, 18 countries reported compliance with the Eurostat definition for all aggregated income variables. However, this is not to say that these variables can be considered fully comparable. Other factors that are not covered in this section may affect comparability, such as differences in the source of data collection (register data versus survey data), mode of data collection, and degree and method of imputation. As far as the other countries are concerned, the main reasons for non-compliance include deviations from the standard definition suggested in the Eurostat guidelines (Eurostat, 2016a), omission of income components and misallocation of income components. This is shown in more detail in Table 18.2.

Table 18.2: Reported issues for cross-country comparability with regard to the composite income variables

Composite income variables	Potential issues for comparability		
	Deviations from the standard definition (equation to compute the variable)	Omission of other income target variables	Misallocation or omission of income components
HY010 (total household gross income)		Serbia (*)	
HY020 (total disposable household income)	France and Slovenia	Serbia (*)	
HY022 (total disposable household income before social transfers other than old-age and survivor benefits)	Spain (*), France, the Netherlands (*) and Slovenia	Belgium and Serbia (*)	Denmark (*)
HY023 (total disposable household income before social transfers including old-age and survivor benefits)	Spain (*), France, the Netherlands (*) and Slovenia	Belgium and Serbia (*)	Denmark (*)

(*) Cross-country comparability is compromised.

Source: Information collated by the authors on the basis of MetaSILC 2015.

Deviations from the standard definitions suggested by the Eurostat guidelines were observed for Spain, France, the Netherlands and Slovenia. In France and Slovenia, the variables are calculated on the basis of net income components, rather than gross income components. In theory, this does not undermine comparability as long as the tax–benefit systems in both countries allow for a 'clean' collection of net income amounts. Similarly, Spain calculates HY022 and HY023 on the basis of net income components, rather than gross income components. However, while France and Slovenia reported adjustments that allow for cross-country comparability, Spain seemingly did not.

Although the Netherlands reported using the recommended equations (Eurostat, 2016a) to compute HY022 and HY023, modifications compared with the standard definition of taxes on income and social contributions (HY140G) were observed. The variable was calculated without taking account of social transfers (PY090G, PY120G, PY130G, PY140G, HY050G, HY060G, HY070G); consequently, HY140G refers to the fictitious amounts that should have been paid if such social transfers were not received. Even though this is arguably a better way of computing these variables ([106]), it is different from how other countries calculate them, compromising cross-country comparability.

Limits to cross-country comparability reported in the more disaggregated income target variables may also affect cross-country comparability of HY010, HY020, HY022 and HY023. We limit ourselves to describing some misallocations of specific income components. Obviously, in contrast to the disaggregated variables, the composite income variables are affected only insofar as misallocations lead to inclusion or exclusion of specific income components in the composite variables, and not just the underlying variables. The main challenges we identified are listed in Table 18.3.

([106]) Leventi, Papini and Sutherland (2021) study more systematically alternative ways of computing the 'before taxes and transfers' variables.

Table 18.3: Limits to cross-country comparability of the composite variables: overview by disaggregated variable

Variables with limitations in cross-country comparability	Potential threats to cross-country comparability	Countries	Affected variables
Income from rental of a property or land (HY040)	*Omission* of (1) costs such as mortgage interest payments, minor repairs, maintenance, insurance and other charges, which must be deducted from the values reported and (2) smaller values of income from rental of a property – these might not be reported at all or are reported under different variables.	(1) Hungary and Serbia; (2) Denmark	HY010, HY020, HY022 and HY023
Family-/children-related allowances (HY050)	*Omission* of (1) family- and children-related tax credits. Tax credits for taking parental leave should be considered as benefits received in addition to a salary for bringing up children. *Misclassification* of (2) payments for fostering children, which must be included under PY010. *Differentiation* of (3) payments for fostering children from transfers paid by the government as a form of social benefit (included in HY050), even though Eurostat guidelines are not clear about this differentiation.	(1) Malta and Slovakia; (2) Cyprus, Latvia and Malta; (3) Bulgaria, Germany and Poland	(1) HY010 and HY020; (2) and (3) HY022 and HY023
Old-age benefits (PY100)	*Misclassification* of income from private pensions, which should be considered under PY080 (pension from individual private plans).	Denmark	HY023
Regular taxes on wealth (HY120)	*Omission* of (1) municipal taxes on the land value of residential property and (2) real estate tax and urban real estate value tax.	(1) Denmark; (2) Spain	HY020, HY022 and HY023
Employee cash or near-cash income (PY010)	*Omission* of (1) fringe benefits, (2) payments for fostering children and (3) allowances paid for working in remote locations. *Inclusion* of (4) allowances for purely work-related expenses.	(1) Denmark and Hungary; (2) Austria, Belgium, Denmark, Estonia, Finland, France, Hungary, Italy, Luxembourg, Netherlands, Slovenia, Spain, Sweden and United Kingdom; (3) Estonia; (4) Croatia and Finland	(1) and (2) HY010 and HY020; (1), (2), (3) and (4) HY022 and HY023
Cash benefits or losses from self- employment (PY050)	*Misclassification* of own consumption, which should be included under HY170 (income from production for own consumption) and not considered when calculating household income.	Croatia	HY010, HY020, HY022 and HY023
Pension from individual private plans (PY080)	*Misclassification* of pensions from mandatory employer-based schemes, which should be accounted under old-age benefits (PY100).	Malta and the United Kingdom	HY022 and HY023

Source: Information collated by the authors on the basis of MetaSILC 2015.

18.4. Conclusion and recommendations

It is very hard to make an overall assessment of the comparability of the income data brought together in EU-SILC. Comparability depends on many factors, and substantive comparability depends on the purpose of the study for which the data are used. It is clear that data collection and data processing in EU-SILC vary greatly across countries. Nonetheless, EU-SILC is probably the best resource to study the distribution of income from a comparative perspective in Europe. Therefore, those using the data and the indicators that are based on EU-SILC should be made more aware of the limitations to comparability and, whenever possible, the potential impact this may have on their (policy) conclusions. Hence, there is a need for better documentation of the 'contents' and comparability of the EU-SILC income variables. With the MetaSILC 2015 database, we set out to provide a useful addition to the toolbox for EU-SILC users and producers. This exercise has taught us that there are quite a few challenges for EU-SILC users and policymakers who rely on EU-SILC for monitoring poverty and inequality in the EU and for analysing the distributive impact of particular policies or policy reforms. Some of these challenges are unavoidable, given the diversity in policy systems and social realities in Europe. However, there is also considerable room for improving both the documentation and the comparability of the income target variables in EU-SILC.

It is hard to find detailed information on the exact implementation of EU-SILC in each country, and to document the comparability of the income variables. A survey among NSIs (the data producers) proved very helpful for collecting detailed metadata, especially with regard to how specific income components are aggregated into the income target variables. Therefore, we believe that it would be a useful strategy to repeat and expand MetaSILC in the future, preferably led by Eurostat, as it clearly provides essential information that is not available anywhere else. Ideally, this should be done during the data production process, as at that point all required information is available to NSIs, and problems of misallocation or inconsistent allocation can then be usefully addressed before the microdata are transferred to Eurostat and other EU-SILC users. We are confident that for future updates it should be possible to limit the required time investment on the part of NSIs. Obviously, the results of this exercise should be made easily accessible to both data producers and data users. One could also think about ways in which data users could contribute to sharing more information on the comparability of specific variables for specific research purposes, for instance by creating a repository with notes by variable. A more ambitious expansion that could be considered, and that would be very useful to many users, is to complete the MetaSILC data set with an institutional description of each income component (especially benefits) covered by each variable.

In this chapter, we highlight some of the main challenges for comparative research on the distribution of income using EU-SILC. There is quite a lot of room for improvement, in various directions. First, we think that there could be more rigid guidelines, but also more guidance for NSIs when they design their data collection strategies. More guidance is essential in terms of (i) the interpretation of the definitions of quite a few target variables; (ii) the aggregation of specific income components in the target variables and, in particular, the consistent classification of income components across countries; and (iii) the implementation of the data collection. Some target variables are not defined in a sufficiently clear and mutually exclusive way. To some extent, this is unavoidable, given the broad range of tax–benefit systems in the EU and differences in economic and financial realities. However, the current situation results in difficulties in judging whether or not all countries comply with the Eurostat guidelines, as there can be confusion regarding how each income component should be classified in terms of the EU-SILC target variables. This implies that countries, but also data users, should receive more guidance about which sources of income should be collected for which target variable, and in which target variables 'borderline cases' should be included. To this purpose, we strongly recommend that Eurostat establishes an expert panel consisting of members of public administrations, NSIs and academic institutions to ensure the consistent interpretation and implementation of the definitions of target variables in a way that

is most relevant for policymakers and researchers, and to support NSIs to classify income components in a consistent way across countries. We are strongly convinced that many NSIs would favour such an approach, given the many positive reactions received and multiple corrections that NSIs announced in response to our report (for details on announced changes to target variables, see Zardo Trindade and Goedemé, 2020). In addition, it would be advisable for this expert panel to monitor how the data collection is implemented in practice.

When incomes are collected through a survey, practices regarding how the income questions are asked, the examples that are given and the level of detail (and number of questions) with which the information for target variables is collected vary greatly. Even though complete input harmonisation is not feasible and not desirable, currently, survey questions seem to vary in ways that are difficult to justify. We are strongly convinced that it is possible to strengthen substantially the 'guided output harmonisation' process that EU-SILC currently follows, while fully respecting national interests and specificities.

Concrete recommendations on collecting data through a survey or registers are beyond the scope of this chapter. However, practices vary greatly, and the move from survey to register data is not always clearly documented, even though it is obvious that this may have a major impact on the measurement of income, especially at the tails of the distribution. We would expect such analyses to be carried out by NSIs in preparation for this change. Therefore, it would be useful if Eurostat could ask for research notes on the change from survey to register data to be made publicly accessible in an online repository (such as the Communication and Information Resource Centre for Administrations, Businesses and Citizens).

Another major variation between countries concerns the collection of income in gross or net terms. Data producers should strive for more consistency in the methods used for converting net to gross incomes. If this higher level of consistency is not feasible, then at least the net-to-gross or gross-to-net procedures used for each country should be documented more transparently and in far more detail than is currently the case.

Finally, more research is necessary to evaluate the impact of the variety of data collection approaches in EU-SILC on the conclusions of (comparative) distributive analyses. The few studies that have addressed the issue (in relation to EU-SILC) indicate that reaching sound conclusions from comparative analyses requires an awareness that some results are probably driven by the distinctive data collection approaches used. EU-SILC users should be made more aware of this and be trained to be sensitive to the potential impact of cross-national variations in data collection on their (policy) conclusions.

References

Atkinson, A. B., Guio, A.-C. and Marlier, E. (eds) (2017), *Monitoring Social Inclusion in Europe*, Publications Office of the European Union, Luxembourg.

Di Meglio, E., Dupré, D., Montaigne, F. and Wolff, P. (2017), 'Investing in statistics: EU-SILC', in Atkinson, A. B., Guio, A.-C. and Marlier, E. (eds), *Monitoring Social Inclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 51–61.

Eurostat (2016a), *Methodological Guidelines and Description of EU-SILC Target Variables – DocSILC065 (2015 operation)*, European Commission (https://circabc.europa.eu/sd/a/afb4601b-4e5c-4f40-86bb-0c3d0d94aa12/DOCSILC065operation-2015VERSION08-08-2016.pdf).

Eurostat (2016b), *ESSPROS Manual and User Guidelines: European system of integrated social protection statistics (ESSPROS)*, Publications Office of the European Union, Luxembourg.

Goedemé, T. and Zardo Trindade, L. (2020a), *MetaSILC 2015: A database on the contents and comparability of the EU-SILC income variables*, Herman Deleeck Centre for Social Policy, University of Antwerp, Antwerp, and Institute for New Economic Thinking, University of Oxford, Oxford, doi:10.7910/DVN/TLSZ4S.

Goedemé, T. and Zardo Trindade, L. (2020b), *The Comparability of the EU-SILC Income Variables: Review and recommendations*, Eurostat Statistical Working Paper, Publications Office of the Europe-

an Union, Luxembourg (https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-tc-20-001).

Goedemé, T., Storms, B., Stockman, S., Penne, T. and Van den Bosch, K. (2015), 'Towards cross- country comparable reference budgets in Europe: first results of a concerted effort', *European Journal of Social Security*, Vol. 17, No 1, pp. 3–31.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009), *Survey Methodology*, 2nd edition, Wiley, Hoboken, NJ.

Iacovou, M., Kaminska, O. and Levy, H. (2012), 'Using EU-SILC data for cross-national analysis: strengths, problems and recommendations', *ISER Working Paper Series*, No 2012-03 (https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2012-03.pdf).

Jäntti, M., Törmälehto, V.-M. and Marlier, E. (2013), *The use of registers in the context of EU-SILC: Challenges and opportunities*, Publications Office of the European Union, Luxembourg.

Leventi, C., Papini, A. and Sutherland, H. (2021), 'Assessing the anti-poverty effects of social transfers: net or gross? And does it really matter?', in Guio, A.-C., Marlier, E. and Nolan, B. (eds), *Improving the understanding of poverty and social exclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 123–138.

Lohmann, H. (2011), 'Comparability of EU-SILC survey and register data: the relationship among employment, earnings and poverty', *Journal of European Social Policy*, Vol. 21, No 1, pp. 37–54.

MISSOC (2015), 'Organisation', European Commission, Brussels (https://www.missoc.org/missoc-database/organisation/).

Pennell, B.-E., Cibelli Hibben, K., Lyberg, L. E., Mohler, P. P. and Worku, G. (2017), 'A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts', in Biemer, P. P., de Leeuw, E. D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, C. and West, B. T. (eds), *Total Survey Error in Practice*, Wiley, Hoboken, NJ, pp. 179–202.

Rendtel, U., Nordberg, L., Jäntti, M., Hanisch, J. and Basic, E. (2004), 'Report on quality of income data', *CHINTEX Working Papers*, No 21, Statistiches Bundesamt, Wiesbaden.

Statbel (2018), *Using Registers in BE-SILC to Construct Income Variables – Eurostat grant: Action plan for EU-SILC improvements*, Statbel, Brussels (https://statbel.fgov.be/en/themes/households/poverty-and-living-conditions/risk-poverty-or-social-exclusion#documents).

Sutherland, H. and Figari, F. (2013), 'Euromod: the European Union tax–benefit microsimulation model', *International Journal of Microsimulation*, Vol. 6, No 1, pp. 4–26.

Törmälehto, V.-M., Jäntti, M. and Marlier, E. (2017), 'The use of registers in the context of EU-SILC', in Atkinson, A. B., Guio, A.-C. and Marlier, E. (eds), *Monitoring Social Inclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 499–508.

Verma, V. and Betti, G. (2010), 'Data accuracy in EU-SILC', in Atkinson, A. B. and Marlier, E. (eds), *Income and Living Conditions in Europe*, Publications Office of the European Union, Luxembourg, pp. 57–77.

Zardo Trindade, L. and Goedemé, T. (2020), 'The comparability of the EU-SILC income variables: review and recommendations', *Eurostat Statistical Working Papers*, 2020 edition, Publications Office of the European Union, Luxembourg, doi:10.2785/047001.

19 The validity and comparability of EU-SILC health variables

Stefaan Demarest and Rana Charafeddine ([107])

19.1. Introduction

This chapter analyses the comparability of the European Union Statistics on Income and Living Conditions (EU-SILC) rolling health module across EU Member States. Where possible, questions that are part of this module will be compared with similar questions in the European Health Interview Survey (EHIS; Eurostat, 2013a). This section describes the development and content of the new EU-SILC rolling health module. Section 19.2 then introduces the analytical approach used to assess the comparability of the instrument and the analysis is presented in Section 19.3. Section 19.4 presents recommendations for enhancing the validity and comparability of EU-SILC data across Member States.

19.1.1. Development of the EU-SILC health module

Since its first wave in 2004, EU-SILC has included a set of three general health variables (self-perceived health, chronic morbidity and activity limitations), commonly referred to as the Minimum European Health Module, and four variables related to unmet healthcare. As for all other modules of EU-SILC, the Member States have only to provide the necessary data to enable Eurostat to calculate the requested health variables (post harmonisation). Nevertheless, Eurostat has proposed a model questionnaire that Member States can use to collect the data (Iacovou, Kaminska and Levy, 2012; Iacovou and Lynn, 2013).

Further integration and streamlining of the European official social surveys were initiated in 2011, and in the context of this reform it was stated that EU-SILC had to better cover the multidimensional aspects of living conditions, poverty and social exclusion by addressing additional themes, including health, access to services and quality of life, themes that cannot be accommodated by the flexible mechanism of the ad hoc modules (Duprez and Di Meglio, 2014).

To enable this, it was decided to reduce the number of EU-SILC variables that were required every year and to use the extra space for fixed rotating modules of about 20 variables with a periodicity of 3 years for variables dealing with labour, health, children and housing and a periodicity of 6 years for other topics such as social participation, quality of life, access to services, wealth and debt (Eurostat, 2013b).

In parallel, the directors of social statistics of the EU Member States decided to reschedule future EHIS waves; instead of every 5 years, EHIS would be organised every 6 years and this would start from EHIS wave 4 (scheduled for 2025) under Regulation (EU) 2019/1700 on EU-SILC ([108]). The timing of the surveys should allow EHIS and EU-SILC to be carried out simultaneously, including a rolling health module every 6 years. During the task force meet-

([107]) Stefaan Demarest and Rana Charafeddine are both with Sciensano, Brussels, Belgium. The authors would like to thank Lucian Agafitei and Didier Dupré from Eurostat. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Stefaan Demarest (stefaan.demarest@sciensano.be).

([108]) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.LI.2019.261.01.0001.01.ENG

ings on the revision of the EU-SILC legal basis, it was stressed that, provided that this module in EU-SILC includes variables that are harmonised with other European surveys focused on health (in particular EHIS), some advanced techniques of data pooling could potentially be used for the calibration of these proxy measurements obtained from EU-SILC. This would be in line with one objective of the modernisation of social statistics, which anticipates the harmonisation of social surveys and a better use of their complementarity.

Before implementing the re-design of EU-SILC into a new legal act, some testing of the variables was found to be necessary. It was agreed that the space offered by the ad hoc modules would be at least partly used for testing variables that would be new to EU-SILC. Regarding health, it was decided to use the 2017 EU-SILC ad hoc module for testing.

This ad hoc module of the EU-SILC is divided into two parts.

- **Part 1: Module on health and children's health.** This includes variables proposed for a future 3-yearly module on health and variables on the health of children intended for a future 3-yearly module on children. This part is implemented in accordance with Commission Regulation (EU) 2016/114, which means that all variables are collected in all Member States in accordance with EU-SILC legislation ([109]).
- **Part 2: Supplementary variables on health (and on labour, over-indebtedness, consumption and wealth).** These cover various topics considered for future EU-SILC modules. This part is implemented through a special non-legal instrument called the European Statistical System agreement, which consists of a commitment by Member States to implement variables for at least one topic.

19.1.2. Content of the EU-SILC rolling health module

The list of health variables to be included in the EU-SILC rolling health module is the result of several rounds of consultations with other European Commission services and Member States. The initial proposal from Eurostat consisted of 43 variables that were ranked in priority order. These were subject to further discussion in the successive task force meetings on the revision of the EU-SILC legal basis (Eurostat, 2013c). A reduced list of 27 variables was discussed during the eighth meeting of the task force (4–5 March 2014), at which the importance and relevance of each variable were assessed using criteria such as policy needs, appropriateness in EU-SILC and robustness.

After applying these criteria, the list of variables was again reduced and presented during the ninth meeting of the task force (17–18 September 2014); at the meeting 17 variables were approved and a decision was pending for four variables (Eurostat, 2014). The modules proposed covered the following topics: health status, healthcare (use of formal care, financial burden of healthcare) and health determinants (body mass index (BMI), physical activity, consumption of fruit and vegetables and smoking, and possibly alcohol consumption). Ultimately, 21 variables were approved for inclusion in the rolling health module.

These variables were divided into two groups: 'first priority' and 'second priority' variables.

- The first priority variables are implemented in all Member States and are listed in Commission Regulation (EU) 2015/2256 ([110]) (10 variables).
- The second priority variables are implemented in a limited number of Member States that volunteered to add the corresponding questions (Bulgaria, Estonia, Greece, France, Romania, Slovenia and Slovakia) and are listed in Commission Regulation (EU) 2016/114 ([111]) (11 variables).

19.2. Analytical approach

To ensure the validity of the EU-SILC results, given the multicountry focus, it is essential to assess the comparability of the data among Member States. Data can be considered comparable when:

([109]) https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=uriserv:OJ.L_.2016.023.01.0040.01.ENG

([110]) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32015R2256

([111]) https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=uriserv:OJ.L_.2016.023.01.0040.01.ENG

> *data (estimates) for different populations (whether countries or different groups within the same country) can be legitimately (i.e. in a statistically valid way) put together (aggregated), compared (differenced), and interpreted (given meaning) in relation to each other and against some common standards.* (Verma, 2006, p. 6)

To assess comparability, 'it is essential to examine both the input side (an analysis of the methodology and implementation of the process of production of the data) and output side (a comparison of the substantive results actually obtained)' (Verma, 2006, p. 12).

On the input side, a number of factors need to be assessed, including target population, sampling scheme, mode of data collection, questionnaire design and weighting. However, these factors are not unique to the health module and have already been extensively studied elsewhere for EU-SILC (Clémenceau and Museux 2007; see also Chapters 3, 9 and 24 of this book) and more generally on health and other surveys (Jäckle, Roberts and Lynn, 2010; Hoebel et al., 2014; Garbarski, Schaeffer and Dykema, 2015; Berger et al., 2016; Croezen, Burdorf and van Lenthe, 2016). In the context of this study, the focus is on the content of the EU-SILC rolling health module, specifically on whether the questions included by the Member States are in line with the EU-SILC model questionnaire.

As described above, a model questionnaire has been developed by Eurostat but, as flexibility is an essential feature of the EU-SILC, Member States are not required to strictly implement this questionnaire. Therefore, a number of Member States follow the questionnaire as provided while other Member States introduce changes to fit with their own national questionnaire and national data needs. In this chapter, an analysis is undertaken to assess to what extent the Member States have followed the model questionnaire in relation to the order of the questions, the level of measurement (individual or household level), the wording of the questions, the wording of the answer categories, the framing of the questions and the reference periods applied. However, no evaluation was carried out concerning cultural specificity or the translation process. For questions that appear in both EU-SILC and EHIS, the wording of the questions listed in the EU-SILC model questionnaire was compared with that of the relevant EHIS questions from the wave 2 (2015) model questionnaire. To undertake this analysis, the statistical offices of all Member States responsible for conducting the 2017 EU-SILC were contacted by post and asked to provide the national version of the rolling health module. If available, the offices could also provide an English version of this module. All Member States provided their version of the module; some provided it only in the national language, some provided it in both the national language and English, and some provided it exclusively in English. For the versions in the national languages, rough translations into English were carried out using Google Translate and other online translation programs. Each national version was compared with the model questionnaire, which is described in the EU-SILC methodological guidelines (Eurostat, 2016).

On the output side, some weighted prevalence rates in the population of both EU-SILC and EHIS variables (wave 2) are presented without assessing whether differences in prevalence are (partially) due to differences in order or wording of the questions or answer categories.

19.3. Analysis of comparability of the EU-SILC rolling health module

This section analyses in a systematic manner all the variables of the EU-SILC revolving health module. For each variable, possible deviations of the Member State questionnaires from the model questionnaire are described. Where applicable, a comparison with similar variables in EHIS (wave 2) is performed.

19.3.1. Use of healthcare goods and services

Number of visits to a healthcare provider

The 2017 EU-SILC rolling health module included three closed questions on the number of visits to

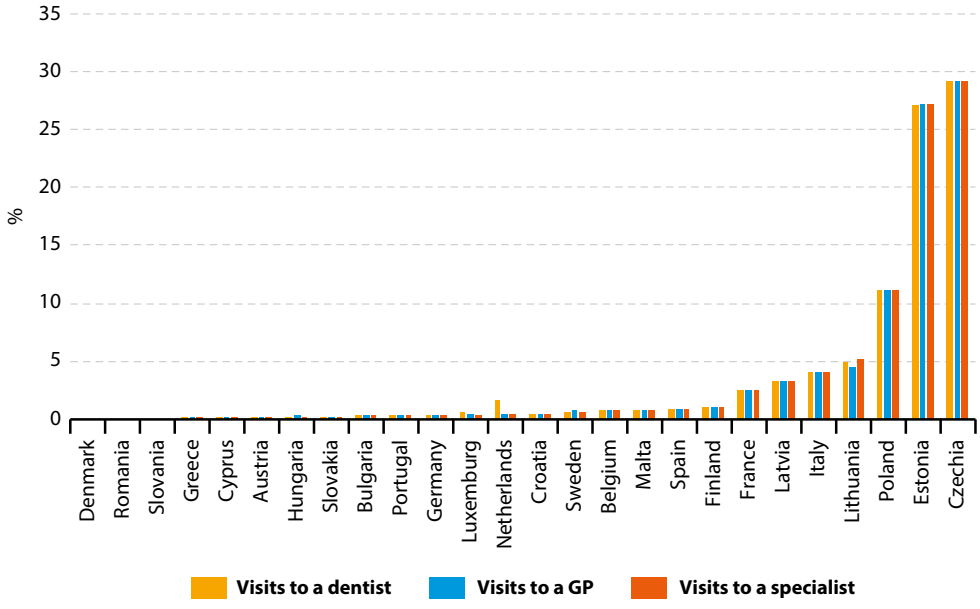
a healthcare provider in the past 12 months: visits to a dentist (PH080), a general practitioner (GP) or family doctor (PH090) and a medical or surgical specialist (PH100) (proposed order as listed in the model questionnaire). It is important to note that the model questionnaire states that these questions, addressed to all, should be included in the questionnaire after the questions on unmet needs for medical and dental examinations (PH040, PH050, PH060 and PH070) and with no inclusion of any other variables between them to ensure higher comparability of results (Lee and Schwarz, 2014). Data derived from EU-SILC on visits to a healthcare provider are not comparable with similar data derived from EHIS, as the latter measures the time of the last visit to each type of healthcare provider, rather than the number of visits.

Denmark did not include questions on the number of visits to a healthcare provider, as it used register information. In Estonia, France, the Netherlands and Finland the questions on the number of visits were preceded by a filter question ('Have you consulted a dentist/GP/specialist in the past 12 months?'). In Luxembourg and Poland, the questions on the number of visits were combined with the questions on unmet need, but were asked of all respondents regardless of their responses to the unmet needs questions. In Greece, the questions on the number of visits were asked only of respondents indicating a need for care. In Czechia, open-ended questions were used to assess the number of visits to a healthcare provider.

The framing of the question on the number of visits to a dentist was identical across the Member States. For the consultation with a GP, some Member States excluded consultations by phone (Malta and Portugal), others did not mention such consultations (Germany, Spain, Italy, Romania, Slovakia and Sweden) and one Member State mentioned paediatricians in this question (Latvia). For the consultation with a specialist, some Member States did not mention that emergency department visits should be included and/or that hospitalisations should be excluded (Germany, Spain, Italy, Cyprus, Latvia, Malta, Portugal, Romania, Slovenia, Slovakia and Sweden).

Deviating from the model questions on visits to healthcare providers impacts the level of missing data (and consequently the distribution of visits), as shown in Figure 19.1. Czechia, which

Figure 19.1: Missing values for the number of visits to a healthcare provider, by Member State



Source: EU-SILC database, 2017.

used open-ended questions to assess these visits, shows the highest levels of missingness. In addition, in Estonia and Poland, where the questions on visits to healthcare providers were conditional questions, preceded by questions on whether the respondents had visited healthcare providers, high levels of missingness could be observed (missingness in these cases probably refers to 'no/zero visits').

Number of nights spent as a patient in a hospital

In the 2017 EU-SILC, the number of nights spent as a patient in a hospital (PH140T1) was measured (second priority variable). To assess this, two questions were proposed in the model questionnaire. The first question asks respondents whether they have been hospitalised in the past 12 months as an inpatient, that is, overnight or longer (PH140_Q1). For those who have been hospitalised, a follow-up question asks about the number of nights in total spent in hospital (PH140_Q2). The variables on hospitalisation used in the 2017 EU-SILC are identical to those used in EHIS.

Five of the seven Member States specified that hospital stays excluded any stay concerning childbirth, which is in line with the model questions and is specified in the technical guidelines. Only Bulgaria was silent on this matter. In addition, Bulgaria added a response category for those who were currently hospitalised.

Use of any home-care services for personal needs

The use of any home-care services for personal needs (PH150T1) was assessed (second priority variables) using two questions. The first question asked whether respondents had used or received any home-care services in the past 12 months (PH150_Q1). For those who had not used these services, a follow-up question asked about the reasons for non-use (PH150_Q2). While in EHIS the use of home-care services is assessed, no data are collected on the reasons for non-use.

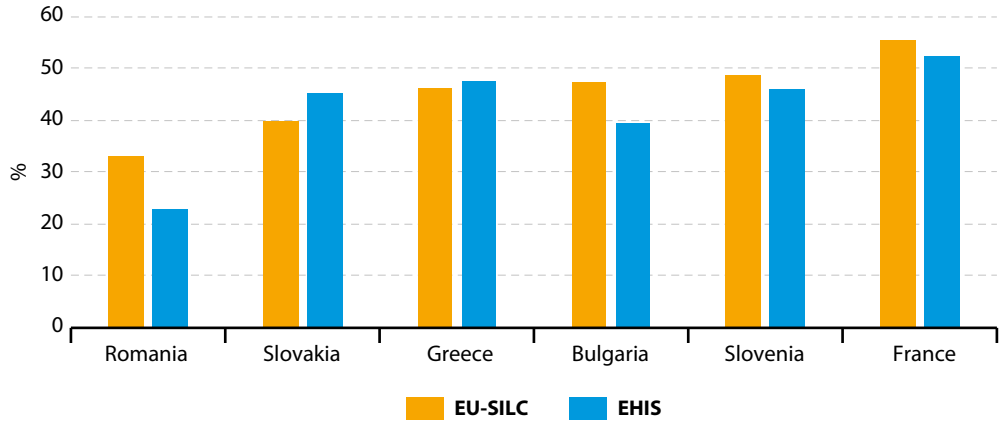
Only two Member States deviated significantly from the model questions. Slovenia preceded these questions with a question on the need for such services, and those who said that they did not need home-care services omitted these questions. As a consequence, the second question had only four response categories, as the first category concerning the need for these services was dropped. Estonia included the first question and dropped the question on the reasons for not using these services, replacing it with a question on whether home-care services were needed in the past 12 months. The other Member States deviated slightly in the framing of the first question: Bulgaria, Greece and Romania did not include a definition of home-care services, while Greece and France explicitly mentioned that the question referred to home-care services related to health problems.

Use of any medicines prescribed by a doctor

One variable was included on the use of medicines (PH160T1) (second priority variable). This assessed whether, in the past 2 weeks, respondents had used any medicines prescribed by a doctor. This variable is identical to that measured in EHIS. This question, as applied by the seven Member States, did not deviate significantly from the model question. Two Member States deviated slightly in the framing of the question. France expanded the concept of interest to include prescriptions from a dentist (in addition to those from medical doctors) and explicitly mentioned the inclusion of prescribed homeopathy and dietary supplements. Greece mentioned that the medicines prescribed before the reference period (past 2 weeks) but used during the reference period should also be reported. In addition, Greece included prescribed medicinal herbs or vitamins in the question in addition to medicines. All of the Member States excluded contraceptive pills from the medicines to be reported.

A comparison of the results obtained in the 2017 EU-SILC and those obtained in EHIS (wave 2) does not show consistent differences: while in Bulgaria, France, Romania and Slovenia the percentages obtained were higher for the 2017 EU-SILC than for EHIS (wave 2), the inverse is true for Greece and Slovakia (Figure 19.2).

Figure 19.2: Use of any medicines prescribed by a doctor, by Member State: comparison between the 2017 EU-SILC and EHIS (wave 2)



Source: Authors' own analysis based on EU SILC database, 2017, and EHIS wave 2 data.

19.3.2. Health status

The 2017 EU-SILC asked the respondents whether they have difficulty performing basic universal activities in four core functional domains: seeing, even when wearing glasses or contact lenses (PH100T1), hearing, even when using a hearing aid (PH110T1), walking or climbing steps without the use of any assistance or device (PH120T1) and remembering or concentrating (PH130T1) (second priority variables). In the model questionnaire two questions were proposed to address the domain of seeing and hearing: an introductory question on the use of glasses / contact lenses (PH100_Q1) and a hearing aid (PH110_Q2) followed by a question on difficulties in seeing (PH100_Q2) and hearing (PH110_Q2). Variables on seeing, hearing and walking are identical to those measured in EHIS, while the variable on remembering is not part of the data collected by EHIS.

Both Bulgaria and France did not apply the introductory questions on the use of glasses / contact lenses and a hearing aid; they directly asked the questions on difficulties in seeing ('even wearing glasses or contact lenses') and hearing ('even using a hearing aid'). In Estonia, it was stressed that the glasses had to be prescribed by a doctor for daily use, while in Slovenia it was mentioned that glasses used only for reading, watching television or driving should be included. In France, it was mentioned that difficulties in seeing could concern only one eye and difficulties in hearing could concern only one ear. Four out of the seven Member States (Bulgaria, Estonia, Slovenia and Slovakia) did not explicitly mention that possible difficulties in walking or climbing steps should be assessed without the use of any assistance or device.

19.3.3. Health determinants

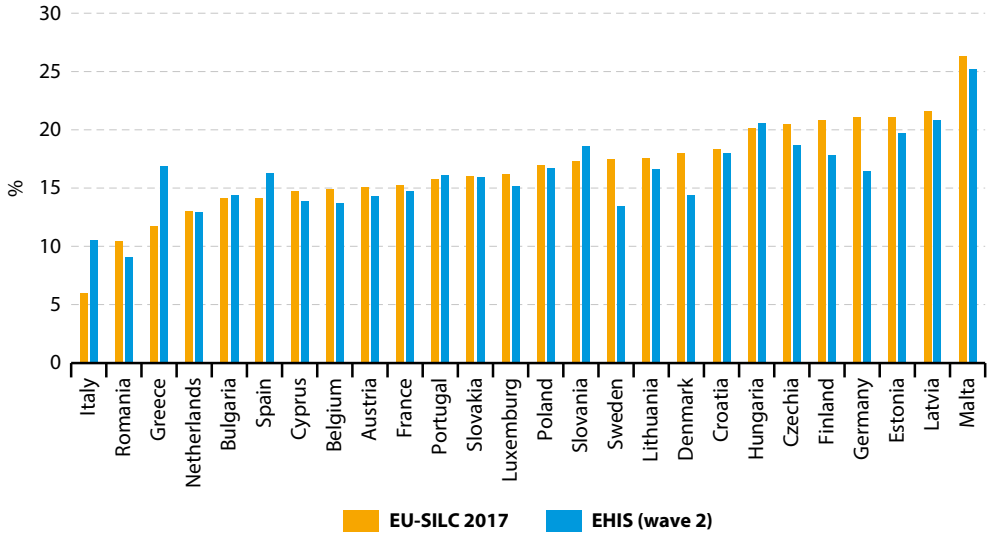
Body mass index

Information on BMI (PH110) was collected. In the model questionnaire, two alternative methods of collecting the data are described: by asking the respondents about their height (in centimetres) (Q1) and their weight (in kilogrames) (Q2), which would provide the necessary information to calculate BMI during the analysis of the data, or by providing a showcard so that respondents could provide their BMI directly based on the combination of their height and weight. The variable BMI was measured in EHIS (wave 2) using the same approach as in the 2017 EU-SILC (either by asking respondents for their height and weight or by using a showcard).

The overwhelming majority of Member States opted to ask respondents their height and weight; only in Ireland and Luxembourg was a showcard used, and this was only as an alternative in case the respondents preferred this way of responding. In all Member States it was specified that height without shoes should be provided and that weight without clothes and shoes should be provided. For pregnant respondents, weight before pregnancy was specified.

In general, the results obtained in the 2017 EU-SILC and EHIS (wave 2) coincided regarding both the percentage of the population who are overweight and the percentage of the population who are obese (Figure 19.3). It can be presumed that the results obtained from the questions used to assess BMI are not impacted by the context or the content of the survey. An assessment of a potential mode effect was outside the scope of this review, but reporting of height and weight is well known to be subject to social desirability bias, which tends to result in mode effects (Spencer et al., 2002; Dekkers et al., 2008; Uhrig, 2012).

Figure 19.3: Percentage of the population who are obese, by Member State: comparison between the 2017 EU-SILC and EHIS (wave 2)



NB: Member States are ordered in ascending order of the percentage of the population who are obese in the 2017 EU-SILC database.
Source: EU-SILC database, 2017, and EHIS (wave 2) database.

Physical activity

Two measures of physical activity were collected. The first assessed the type of physical activity undertaken while working, using a broad definition of working (see below) (PH120). The second variable was the total number of hours and minutes spent on physical activities (excluding working) per week (PH130). In the model questionnaire, the introduction specifies that only physical activities carried out 'for a continuous period of at least 10 minutes and that cause at least a small increase in breathing or heart rate' should be accounted for. The variable on physical activity while working is quasi-identical to the one used in EHIS. The variable on the time spent on physical activities is part of a specific battery of variables in EHIS: the EHIS Physical Activity Questionnaire (EHISPAQ).

The introduction to the model question on the type of physical activity undertaken when working mentions a broad definition of 'working': 'think of work as the things that you have to do such as paid and unpaid work, work around your home, taking care of family, studying or vocational training'. The notion of 'work' is thus wider than a professional activity, or activities performed in the context of paid work. In the national versions of the question, the boundaries of what should be understood as 'working' are very diverse. Czechia and Spain, for example, do not refer to the notion of work, but use the notion 'what you do' and 'your activity in the workplace, school or at home', respectively. In Bulgaria the strict notion of 'work' is used ('when you are working'), while Croatia opted for 'physical activity at your workplace'. Therefore, these versions do substantially deviate from the model question. In the Danish and Finnish versions the question on the type of physical activity undertaken when working is preceded by a (national) question that assesses whether respondents work or not. Depending on the response to this question, respondents have to complete one of two separate (but similar) questions on the type of physical activity undertaken. In Sweden the description provided to respondents of what should be understood by 'work' depends on their age category (15–22 years, 23–64 years, 65+ years).

The second variable measured time spent on physical activities, excluding those activities undertaken 'when working'. A wide range of examples was used by the Member States to define which activities should and should not be taken into consideration, such as 'sport, fitness, yoga, dance and other' (Bulgaria), 'exercise in your free time (or on your way to work and back)' (Cyprus) and 'sport, fitness and physical activity during leisure time' (Austria). However, more important than the differences in the examples were the differences in relation to the minimal time and the intensity of the physical activities mentioned in the model question ('for a continuous period of at least 10 minutes and that cause at least a small increase in breathing or heart rate'). Only in a minority of Member States (France, Italy, Cyprus, Latvia and Austria) was reference made to both requirements. In Greece, Spain and Portugal only the notion of 'continuous activity of at least 10 minutes' was used, while in Germany only the intensity of the physical activity (increase in breathing or heart rate) was referred to. All other Member States did not refer to the minimal time or to the intensity of the physical activity, and should consequently be categorised as deviating substantially from the model questionnaire.

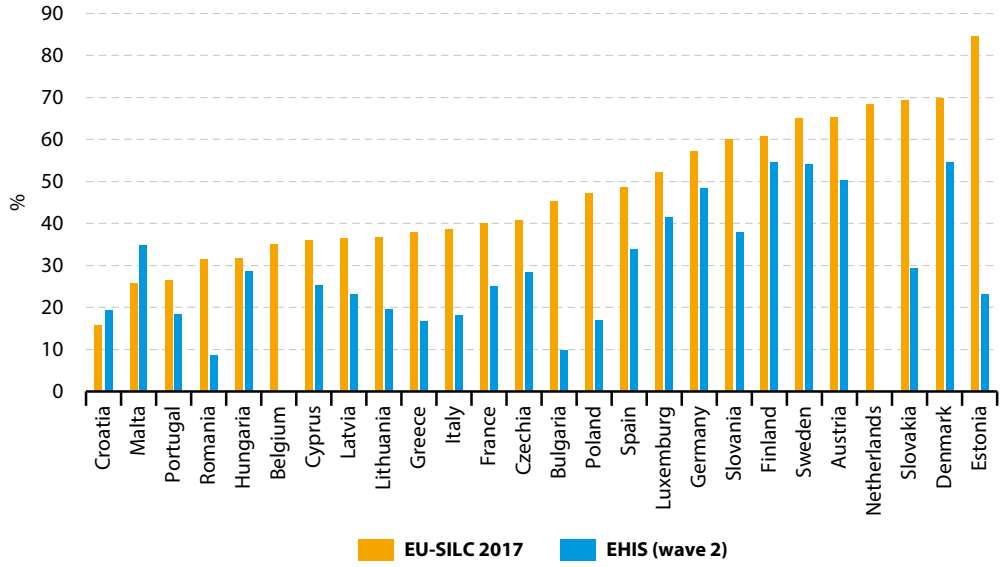
The format of the response – asking about the total number of hours and minutes spent on physical activities (excluding working) per week – was identical for all Member States.

As the 2017 EU-SILC used only one overall question on physical activities (excluding working), the results are not comparable with those obtained using the more detailed question on physical activities (excluding working) in EHIS (wave 2). For example, in comparing the percentages of the population fulfilling the World Health Organization (WHO) recommendation on health-enhancing physical activity (at least 150 minutes per week), the percentages derived from the 2017 EU-SILC are almost universally higher than the percentages derived from EHIS (wave 2); the exceptions are found in Croatia and Malta (Figure 19.4).

Frequency of eating fruit/ vegetables and salad

One variable was included on the frequency of consuming fruit (excluding juice made from concentrate) (PH140) and one was included on the

Figure 19.4: Percentage of the population aged 16 years (15 years in EHIS) and older fulfilling WHO recommendations on health-enhancing physical activity, by Member State: comparison between the 2017 EU-SILC and EHIS (wave 2)



NB: Member States are ordered in ascending order of the percentage of the population fulfilling WHO recommendations on health-enhancing physical activity in the 2017 EU-SILC database.

Source: EU-SILC database, 2017, and EHIS (wave 2) database.

frequency of consuming vegetables and salad (excluding potatoes and juice made from concentrate) (PH150). Both variables were identical to the those used in EHIS. The latter also assessed the portion sizes consumed.

The (two) questions on the consumption of fruit and vegetables, as applied in the different Member States, showed limited deviations from the model questionnaire. Estonia applied a two-step approach in which respondents were asked, first, if they consumed fruit/vegetables in a typical week ('yes'/'no') and, then, the number of times that they consumed fruit/vegetables (without the response category 'never'). In the Netherlands, 'daily' was used as a first response category (instead of 'two or more times a day', as proposed in the model questionnaire). An additional follow-up question then asked if respondents ate fruit/vegetables once a day or more than once a day.

However, more heterogeneity was observed regarding the definitions of fruit and vegetables used by the Member States. The model questionnaire states that fruit juice made from concentrate is to be excluded, but each Member State defined differently the juices to be excluded. In Austria, it was stipulated that 'fruit juices from concentrate or with added sugar' were to be excluded, while in Greece and Cyprus 'juice made from concentrate or artificially sweetened' were to be excluded, in France 'concentrated or powdered fruit juices' were to be excluded, in Italy 'industrial fruit juices' were to be excluded, in Malta 'juice not made from fresh fruit' was to be excluded, in Portugal 'soft drinks, nectars and concentrated juices' were to be excluded, and in Slovakia 'juices prepared from concentrate or processed fruits, or juices artificially sweetened' were to be excluded. In the Danish version, all juices were to be excluded, while in Germany and Croatia no reference at all was made to juices.

Some heterogeneity in the definitions of vegetables could also be found. The model questionnaire stipulates that 'potatoes and juice made from concentrate' should not be taken into consideration,

but each Member State defined differently the items to be excluded. In Greece it was stipulated that 'potatoes and juice prepared from concentrated or processed vegetables' should be excluded, in Malta 'juice and soups not made from fresh vegetables' were to be excluded, in Portugal 'potatoes, yams, manioc and concentrated or processed vegetable juices' were to be excluded, and in Slovakia 'juices prepared from concentrate or processed vegetables, or artificially sweetened' were to be excluded. In Estonia, France, Croatia, Italy, Romania and Finland, (vegetable) juices were not mentioned in the question wording.

Type of smoking behaviour / average number of cigarettes a day

A variable was included on the type of smoking behaviour (daily smoking, occasional smoking, non-smoker). For cigarette smokers, an additional variable was included on the number of cigarettes smoked. The variables on smoking in the 2017 EU-SILC were similar to those measured in EHIS wave 2.

Deviations were identified in three Member States. Estonia used the same question for the type of smoking behaviour ('Do you smoke?') but the response categories included only 'yes'/'no' and therefore no information was available on the pattern of smoking. However, the questions on the number of cigarette smoked in Estonia were in line with the model questions. The only minor variation here was that the number of cigarettes smoked could be expressed in terms of packs of cigarettes, but it was specified that one pack contains on average 20 cigarettes. Slovenia used a multistep approach: respondents were first asked whether they smoked ('yes'/'no'); smokers were then asked whether they smoked daily or occasionally, and daily smokers were asked about the type of tobacco product used and finally the average number of cigarettes smoked. France used slightly different response categories for the question on smoking behaviour ('yes, every day'; 'yes, but not every day'; 'no or rarely'). However the major difference was that the questions on the number of cigarettes smoked were asked not only of daily smokers but also of occasional smokers. A minor variation was the explicit exclusion of e-cigarettes by France, Romania and Slovenia.

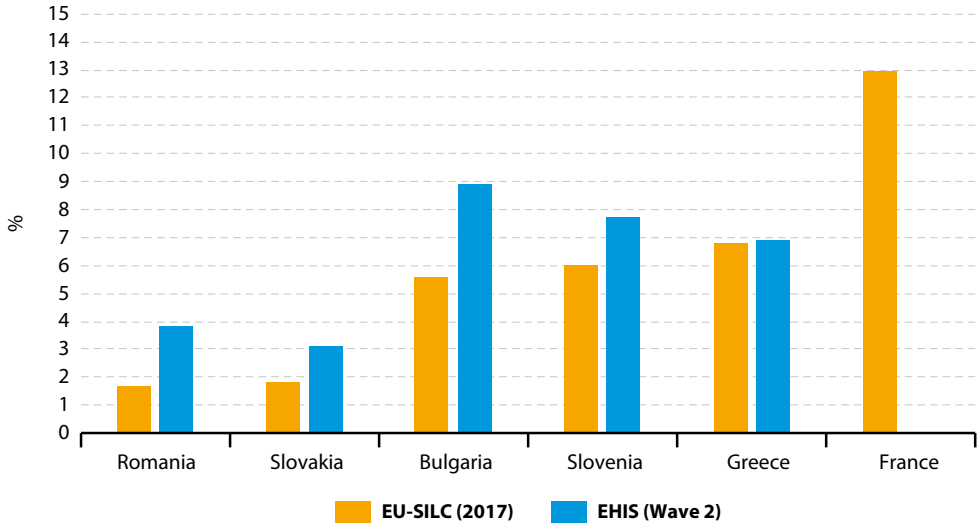
Frequency of consumption of an alcoholic drink of any kind

A variable was included on the frequency of alcohol consumption in the past year. This was identical to the one used in EHIS, although in EHIS the quantity of alcohol consumed was also measured.

No significant deviations among the seven Member States could be identified. In Slovenia, the eighth response category, 'Not in the past 12 months, as I no longer drink alcohol', was split into two categories: 'Never in the last 12 months' and 'I do not drink alcohol anymore'. In Estonia, the question on the frequency of consumption of alcohol was supplemented with two questions: one on the kind of alcohol consumed (in terms of alcohol percentage) and one on binge drinking (the consumption of at least six alcoholic drinks on one occasion).

The question on the frequency of consumption of an alcoholic drink of any kind was identical in both the 2017 EU-SILC and EHIS (wave 2), yet the results of EHIS show, in general, higher frequencies of daily drinking than the results of EU-SILC (for France it was not possible to calculate this indicator for EHIS) (Figure 19.5).

Figure 19.5: Percentage of daily drinkers aged 16 years (15 years in EHIS) and older, by Member State: comparison between the 2017 EU-SILC and EHIS (wave 2)



NB: Member States are ordered by ascending order of the percentage of daily drinkers in the 2017 EU-SILC database.
Source: EU-SILC database, 2017, and EHIS (wave 2) database.

19.3.4. Financial burden of healthcare

The 2017 EU-SILC included three newly developed variables on the financial burden of healthcare, asking to what extent the costs of medical care, dental care and medicines were a financial burden to the household during the past 12 months (the response categories were 'Heavy burden', 'Somewhat a burden', 'Not a burden at all' and 'No one in the household needed medical care / dental care / medicines'). These variables were to be measured at the household level.

In all but one of the 28 national versions, a reference period of 12 months preceding the interview was mentioned in the questions on the financial burden of healthcare. Only in Denmark was reference made to the calendar year 2016 ('To what extent was the cost of (medical treatment) in 2016 a financial burden to the household?').

In Italy, the Netherlands, Austria and Slovenia the burden questions began with the burden of dental care, followed by the burden of medical care and the burden of medicines (while the model questionnaire starts with the burden of medical care). In Denmark, the order was completely modified: the first question was on the burden of medicines, and this was followed by questions on the burden of dental care and the burden of medical care. In Czechia, a question on the burden of hospitalisation was added, while the question on the burden of medicines was split into the burden of prescribed medicines and the burden of non-prescribed medicines. In Luxembourg, households were also asked to consider costs related to 'complementary medicine (for example an acupuncturist, an Ayurvedic practitioner, a Chinese herbal medicine practitioner, a homeopath or a naturopath)' when completing the question on the financial burden of medical care.

In Bulgaria, Czechia, Estonia, France, Italy, Luxembourg and Hungary, the questions on the financial burden of healthcare were preceded by a question on whether the household had expenses for dental care, medical care or medicines ('yes'/'no'). The questions on the burden of care were asked only of households that indicated that they had expenses. As a consequence, the fourth response category proposed in the model questionnaire

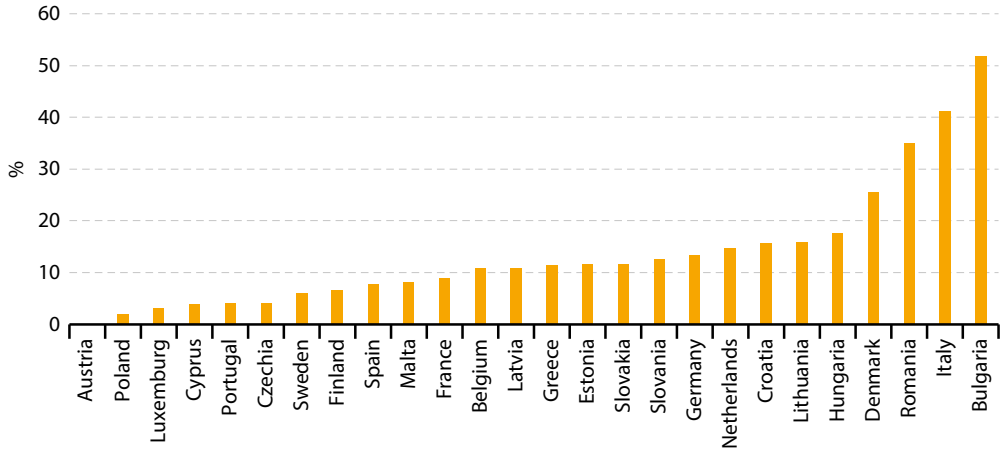
('No one in the household needed (medical examination or treatment, dental care, medicines)') was redundant.

In the model questions on the financial burden of healthcare, the response category 'No one in the household needed (medical, dental) care or medicines' was added. It was expected that the frequencies for this category would show only limited differences across Member States, presupposing that the need for care is equally distributed among the Member States. However, Romania, Italy and Bulgaria showed very high levels (> 30 %) of households indicating no need for (e.g.) medical care (Figure 19.6).

19.4. Conclusions and recommendations

As is the case with the other modules of EU-SILC, the rolling health module is output harmonised rather than input harmonised. The specific approach of this module could perhaps be better described as *ex ante* post harmonisation, given that Eurostat proposes guidelines and a model questionnaire that can be applied by the different Member States. In other words, some precautions are taken before the start of the survey in order to enable harmonisation. Our methodology of roughly translating the rolling health module questions into English and comparing them with the English version of the model questionnaire does not support an in-depth comparison between the different versions, but it allows – in broad terms – an assessment of the extent to which the national versions of the questions and response categories deviate from the model rolling health module questionnaire. Some subjectivity is inherent in inferring whether deviations from the model questionnaire are minor or major, since the impact on responses, and hence on comparisons between countries, has not been tested experimentally. In overall terms, it can be concluded that all Member States at least used the model questionnaire as inspiration for developing their own versions. As the comparison between the national versions of the rolling health module and the children's health module has shown, some deviations from the model questionnaire could be observed. These deviations range from minor differences to more substantial differences. In some cases it is hard to predict whether the deviation will affect the data; in other cases the direction of a likely effect seems clear on a priori grounds. Furthermore, any effects of differences in the wording of questions or response categories may also depend on the characteristics of the sample, non-response,

Figure 19.6: Percentage of households without the need for medical care, by Member State



NB: Member States are shown in ascending order of the percentage of households without the need for medical care.
Source: EU-SILC database, 2017.

the order of the questions, the mode of data collection and so on. Disentangling all these aspects promises to be a challenging task for the future if achieving comparability is to be taken seriously.

The development of a 3-yearly rolling EU-SILC health module has meant confronting a number of constraints, not least the imposed length of the module (20 variables). Choices had to be made with respect to both the number of topics that could be addressed and the level of detail for each topic.

In reviewing the different proposals (from the initial 43 items proposed to the final list of 21 items), it is clear that the 'health status' part is very limited, since this module has been developed to complement the three health status variables of the Minimum European Health Module, part of the EU-SILC nucleus. By adding four of the six questions derived from the Washington Group on Disability Statistics questionnaire on disabilities (Altman, 2016), a more complete picture of limitations and restrictions due to health problems will be obtained. Unfortunately, given that for two questions of the disability questionnaire no information will be collected, international comparison of the results will be hampered. It is therefore recommended that these two questions be added to the EU-SILC rolling health module, which would have a very limited impact on respondent burden.

Similarly to the EHIS, the EU-SILC rolling health module includes questions on the number of times a respondent contacted health providers during a given reference period. While this reference period in EHIS is mostly 'the past 4 weeks', in EU-SILC it is 'the past 12 months', a reference period that is also used in other EU-SILC modules. Maintaining the same reference period throughout the questionnaire has some advantages since it limits distractions and the confusion that could be generated if respondents are asked to use different reference periods. However, given the fact that the reference periods in the peer survey – EHIS – are different, a comparison between the two data sources is not possible. This, of course, is a major setback since it jeopardises one of the main goals of adding a rolling health module to EU-SILC: to serve as a bridge between two waves of EHIS. It is therefore recommended that identical reference periods be used in both surveys to enable a comparison of their findings. For the number of nights spent as an inpatient in hospital, the use of home-care services and the use of prescribed medicines, the reference period in both surveys is identical.

The bulk of the EU-SILC rolling health module is dedicated to questions on health determinants: BMI, consumption of fruit and vegetables, the amount of physical activity undertaken, smoking habits and alcohol consumption. To address so many topics in a (sub)module of such a restricted size jeopardises the prerequisite of a 'minimal critical mass', both for the respondents (the domain switches after almost every question) and for the interpretation and comparison of the data. For instance, the two questions on physical activity in the EU-SILC rolling health module are derived from the extended EHISPAQ used in EHIS, in which these two questions are embedded in a very detailed instrument on physical activity. It is clear that the preceding questions in EHISPAQ will impact the responses given to the two questions and will jeopardise comparability with the EU-SILC questions. In addition, the EU-SILC rolling health module includes questions on the frequency of consuming fruit and vegetables but, contrary to EHIS, no questions on portion size are included. This lack of information on portion size precludes estimation of the quantity of fruit and vegetables consumed. This is also the case for the consumption of alcohol: in EU-SILC information is gathered on the frequency of consumption, but not on the quantity of alcohol consumed. Extending the EU-SILC rolling health module would have a considerable – probably unacceptable – impact on the size of the module. It is therefore recommended that the number of topics is decreased but the number of questions per topic is increased in order to optimise comparability with the EHIS results.

In the 2017 EU-SILC rolling health module a new domain – the financial burden of healthcare – is addressed, based on three questions for which data are collected at household level. The importance or relevance of this new submodule cannot be assessed without making reference to the unmet needs questions, part of the EU-SILC nucleus health module. The unmet needs questions assess situations in which individual respondents had to postpone healthcare (generalist care, dental care)

for specific reasons (possibly financial reasons). As such, the unmet needs questions are more specific than the burden questions, since the latter assess the perception of the financial burden of medical care in general terms. However, it is unclear how the analyst should deal with situations in which the two sets of questions result in apparently conflicting answers. For example, a household member may report postponement of care due to financial reasons while at household level the burden is not described as a 'heavy burden'. Or individual members of a household (15+ years) may indicate not having needed care, while the household burden of medical is declared as a 'heavy burden'. A thorough analysis of the relationship between responses to the burden questions and responses to the unmet needs questions and an assessment of the implications for derived measures are recommended.

References

Altman, B. M. (ed.) (2016), 'International measurement of disability: purpose, method and application', *Social Indicators Research Series*, Vol. 61, Springer, Berlin/Heidelberg.

Berger, N., Robine, J. M., Ojima, T., Madans, J. and Van Oyen, H. (2016), 'Harmonising summary measures of population health using global survey instruments', *Journal of Epidemiology and Community Health*, Vol. 70, No 10, p. 1039.

Clémenceau, A. and Museux, J. (2007), 'Comparative EU statistics on income and living conditions: issues and challenges', *Proceedings of the EU-SILC Conference*, 6–8 November 2006, Helsinki, Office for Official Publications of the European Communities, Luxembourg.

Croezen, S., Burdorf, A. and van Lenthe, F. J. (2016), 'Self-perceived health in older Europeans: does the choice of survey matter?', *European Journal of Public Health*, Vol. 26, No 4, pp. 686–692.

Dekkers, J. C., van Wier, M. F., Hendriksen, I. J. M., Twisk, J. W. R. and Mechelen, W. (2008), 'Accuracy of self-reported body weight, height and waist circumference in a Dutch overweight working population', *BMC Medical Research Methodology*, Vol. 8, No 69.

Duprez, D. and Di Meglio, E. (2014), 'Planned future developments of EU-SILC', EU-SILC Conference, Lisbon, 15–17 October 2014.

Eurostat (2013a), *European Health Interview Survey (EHIS Wave 2) – Methodological guidelines*, Publications Office of the European Union, Luxembourg.

Eurostat (2013b), *Issues paper for the DSS board and DSS discussions*, Eurostat, Luxembourg.

Eurostat (2013c), *Draft summary minutes of the 7th meeting of the task-force on the revision of the EU-SILC legal basis*, Eurostat, Luxembourg.

Eurostat (2014), *9th meeting of the task-force on the revision of the EU-SILC legal basis*, Eurostat, Luxembourg.

Eurostat (2016), 'Working group meeting "Statistics on Living Conditions / HBS" – 2017 operation: Guidelines and other issues', Eurostat, Luxembourg.

Garbarski, D., Schaeffer, N. C. and Dykema, J. (2015), 'The effects of response option order and question order on self-rated health', *Quality of Life Research*, Vol. 24, No 6, pp. 1443–1453.

Hoebel, J., von der Lippe, E., Lange, C. and Ziese, T. (2014), 'Mode differences in a mixed-mode health interview survey among adults', *Archives of Public Health*, Vol. 72, No 1, p. 46.

Iacovou, M. and Lynn, P. (2013), 'Implications of the EU-SILC following rules, and their implementation, for longitudinal analysis', *Institute for Social and Economic Research Working Papers*, No 2013-17, University of Essex, Colchester.

Iacovou, M., Kaminska, O. and Levy, H. (2012), 'Using EU-SILC data for cross-national analysis: strengths, problems and recommendations', *Institute for Social and Economic Research Working Papers*, No 2012-03, University of Essex, Colchester.

Jäckle, A., Roberts, C. and Lynn, P. (2010), 'Assessing the effect of data collection mode on measurement', *International Statistical Review*, Vol. 78, No 1, pp. 3–20.

Lee, S. and Schwarz, N. (2014), 'Question context and priming meaning of health: effect on differences in self-rated health between Hispanics and non-Hispanic Whites', *American Journal of Public Health*, Vol. 104, No 1, pp. 179–185.

Spencer, E. A., Appleby, P. N., Davey. G. and Key, T. J. (2002), 'Validity of self-reported height and weight in 4808 EPIC-Oxford participants', *Public Health Nutrition*, Vol. 5, No 4, pp. 561–565.

Uhrig, S. C. N. (2012), 'Understanding panel conditioning: an examination of social desirability bias in self-reported height and weight in panel surveys using experimental data', *Longitudinal and Life Course Studies*, Vol. 3, No 1, pp. 120–136.

Verma, V. (2006), 'Issues in data quality and comparability in EU-SILC', Eurostat and Statistics Finland Conference 'Comparative EU Statistics on Income and Living Conditions: Issues and Challenges', 6–7 November 2006, Helsinki.

20 Recommendations on the validity and comparability of EU-SILC housing variables

Ross Bowen and Callum Clark ([112])

20.1. Introduction

There is a continuing drive to harmonise methodologies and measurements across countries. Harmonisation at European level is a difficult task, as even basic concepts such as those relating to dwellings, rooms, living spaces, occupancies and housing costs are influenced by cultural, institutional and socioeconomic norms.

The purpose of this chapter is to review the capture of EU-SILC housing variables by countries. The chapter has two primary aims:

1. to provide recommendations on how to improve validity – that is, ensure that the data that countries gather are fit for purpose;
2. to provide recommendations on how to improve comparability – that is, ensure that data are gathered consistently to enable valid comparisons to be made (Iacovou, Kaminska and Levy, 2012).

Focusing on the housing variables in EU-SILC, the processes that countries use to inform these variables are reviewed and compared across national questionnaires to assess comparability, consistency and compliance with the methodological guidelines. The scope of this topic is wide and we have therefore focused on selected priority recommendations for the purposes of this chapter. Greater detail can be found in Clark and Bowen (2018).

20.2. Methodology

We use information obtained through a consultation with countries that was conducted jointly with the University of Antwerp. We consider practices carried out in alternative housing surveys and make comparisons between the impacts of different practices where data are available. We also identify areas with the potential for future data collection.

No countries participating in the Antwerp consultation indicated that any comparative assessments between EU-SILC and their independent housing data had been made. Briant et al. (2010) provide an in-depth review of the sources of housing data available in France ([113]) and emphasise their complementary nature. They discuss the merits of the various sources and highlight the benefits of a more comprehensive approach to the treatment of housing data. In discussing EU-SILC, they conclude that its usefulness is in the breadth of topics covered and the level of harmonisation achieved across Europe, rather than in the depth of detail collected about housing; this idea is echoed in other literature (e.g. Bonnefoy, 2007; Sunega and Lux, 2016).

Our recommendations do not seek to replicate questions that occur commonly in national surveys. Rather, they are based on the idea that more

([112]) Ross Bowen and Callum Clark conducted this analysis while at the UK Office for National Statistics (ONS). All errors are the authors' responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. As both authors have left ONS, correspondence should be addressed to Dominic Webber (dominic.webber@ons.gov.uk).

([113]) These include the French housing survey *Enquête Logement*, the French national census, the household budget survey, a survey focusing on homelessness and, of course, EU-SILC.

detailed data could be collected in a way that maintains the validity of the European Commission's indicators, while better facilitating external users of the data in making their own inferences.

20.3. Data sets and guidelines

Through the Antwerp consultation, most countries indicated that EU-SILC was their main source of data on housing alongside national censuses. On the latter, the United Nations Economic Commission for Europe (UNECE) guidelines (UNECE, 2015) heavily influence their formation and so, as an example of guidelines with similar requirements to those used in EU-SILC, comparisons are made throughout.

Throughout we compare national EU-SILC questionnaires with the EU-SILC methodological guidelines (Eurostat, 2017). Where appropriate we suggest changes to the variables to bring them more in line with those used in other housing surveys, or to allow for more detailed or comprehensive information to be collected. This was possible thanks to the provision of (sometimes partial) translations or the full original English versions of questionnaires by the national statistical institutions of 19 countries ([114]).

20.4. Selected findings

In this section we outline our key points of discussion, findings and recommendations. The comprehensive recommendations from our review are prioritised and listed in the conclusion section. For convenience, the selected variables are grouped by whether they relate to dwelling size, housing affordability, housing conditions or housing characteristics.

([114]) These are Austria, Bulgaria, Cyprus, Denmark, Estonia, Finland, Germany, Greece, Hungary, Ireland, Italy, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Slovenia, Spain and the United Kingdom.

20.4.1. Dwelling size

EU-SILC includes a core variable relating to the size of the dwelling: number of rooms (HH030) ([115]). Two ad hoc variables, size of dwelling in square metres (HC020) and shortage of space (HC010), were included in the 2012 ad hoc module and provide some additional context. These variables were also reviewed; however, for the purposes of this section we focus our attention on the number of rooms variable, since this is part of the core EU-SILC module and is used in the construction of the official indicator of overcrowding, used in the social inclusion portfolio.

Regarding the number of rooms variable (HH030), we noted that countries faced difficulties collecting this in a consistent manner. There appear to be three main reasons for this.

1. The first relates to difficulties conveying what does and does not constitute a room for the purposes of EU-SILC in an easy to understand way. The definition is not intuitive, with bathrooms excluded and kitchens excluded only if used solely for cooking. Utility rooms and lobbies, which might otherwise be considered rooms, are not considered as such. Understandably, many countries had difficulties conveying this definition, and many questionnaires do not advise on which room types should be ignored. In the case of those that do give some instruction, the description of rooms to be excluded is generally incomplete. Countries also frequently do not establish whether a kitchen is used solely for cooking, and many exclude kitchens from the room count entirely.

([115]) For the purposes of EU-SILC, a room is defined as a space in a housing unit of at least 4 m^2, such as a bedroom, dining room, living room, habitable cellar, attic, kitchen or other separated space used or intended for dwelling purposes with a height over 2 m and that is accessible from inside the unit. Kitchens are excluded only if the space is used only for cooking. A single room used as a kitchen-cum-dining room is included as one room in the count of rooms. The following space in a housing unit does not count as a room: bathroom, toilet, corridor, utility room, lobby and veranda. A room used solely for business use is excluded; however, it is included if shared between private use and business use. If a dwelling is shared by more than one household and some rooms are shared with other households (within the same dwelling), the number of shared rooms should be divided by the number of households and an equal share should be added to each household.

2. The second relates to difficulties associated with the treatment of shared spaces, for example shared kitchens in multiple occupancy households. EU-SILC advises that rooms shared between households should be divided accordingly. For example, if a living room is shared between two households, each household should have 0.5 added to its corresponding number of rooms variable. Although this is a valid approach, it does not appear to be well adopted in practice. In 2014, only 8 out of 28 countries ([116]) recorded any households with a non-integer number of rooms.
3. The third relates to difficulties associated with the correct treatment of open-plan living and kitchen areas, which are becoming increasingly common. EU-SILC does not currently provide explicit guidance on this point. This means that an open-plan kitchen, dining and living room may in some cases be considered a single room for the purposes of EU-SILC. This could skew the EU-SILC data negatively towards houses with more modern, open-plan, designs.

Many of these problems could be addressed by collecting data on the number of bedrooms, which could replace or supplement the existing number of rooms variable (HH030). This would have the following benefits.

1. Respondents more intuitively understand what a bedroom represents. They instinctively know to exclude bathrooms, utility rooms, lobbies, and kitchens used only for cooking.
2. Sharing bedrooms between households is uncommon, but it is important to represent this to understand overcrowding. We should focus efforts on this dimension. By asking about the number of bedrooms, respondents will not need to spend time and cognitive effort on issues related to the accurate division of kitchens and living spaces between households, which is a more common arrangement.
3. As bedrooms are not generally open plan, the methodological difficulties noted in the case of the number of rooms variable (HH030) would be largely negated.

([116]) EU-SILC user database, 2014 – version 1 of January 2016 (excludes Germany and Ireland).

The number of bedrooms approach to inform overcrowding is used in Canada and New Zealand. For example, the Canadian National Occupancy Standard defines bedrooms as:

> *Rooms in a private dwelling that are designed mainly for sleeping purposes even if they are now used for other purposes, such as guest rooms and television rooms. Also included are rooms used as bedrooms now, even if they were not originally built as bedrooms, such as bedrooms in a finished basement. Bedrooms exclude rooms designed for another use during the day such as dining rooms and living rooms even if they may be used for sleeping purposes at night. By definition, one-room private dwellings such as bachelor or studio apartments have zero bedrooms.*

In summary, we conclude that collection of this variable could be made more valid and comparable through the following means.

- Consideration should be given to replacing or supplementing the number of rooms variable (HH030) with a number of bedrooms variable, which is likely to be more robust.
- Alternatively, consideration should be given to revising the guidelines on rooms to include kitchens where they meet the minimum volume requirements, counting only rooms that are for the sole use of the household and specifying that open-plan spaces should be counted as if there are walls between the different areas.

20.4.2. Housing affordability

Four variables are gathered in relation to housing affordability. These are current rent related to occupied dwelling (HH060), total housing costs (HH070), financial burden of total housing costs (HS140) and mortgage principal repayment (HH071). These are used to derive the official EU-SILC cost overburden rate, describing the proportion of individuals living in households where total housing costs represent more than 40 % of the household's disposable income (both net of housing allowances), and in comparative analysis (Pittini, 2012).

The financial burden of total housing costs variable (HS140) is in the core yearly survey. Eurostat

has proposed that in the future it should form part of the rolling 3-year module. This variable aims to assess respondents' feelings about the extent to which housing costs are a financial burden to the household. Examination of the questionnaires revealed inconsistencies in this variable. For instance, the Belgian questionnaire asks respondents for their total mortgage costs (interest plus capital repayment), whereas the EU-SILC definition of total housing costs includes only the mortgage interest. For this reason, we recommend that countries explicitly state in their surveys what is included in total housing costs.

The mortgage principal repayment variable (HH071) currently forms part of the core yearly survey and the proposal is that it will be collected annually in the future. EU-SILC enables users to distinguish between mortgage principal and interest payments, which is very useful. This is integral to the study of housing affordability, as mortgage principal repayment is a form of saving rather than a housing cost.

Analysis of countries' questionnaires shows that the variable is informed in two ways. Some countries ask respondent households to separate their principal and interest payments by estimation or by consultation of documentation, while others ask a sequence of questions to gather specific details about the mortgage. These typically relate to the year the loan was taken out, the principal amount borrowed, the term of the loan, the total amount paid up to that point in time, the amount of interest paid on the mortgage during regular payments (if known) and the interest rate. In addition, some countries include questions relating to country-specific elements.

In isolation we are limited by how much information can be gathered from the variable as it is currently collected. In housing-specific surveys, a wealth of detailed information is available on mortgages, including the start date of the mortgage and the mortgage duration. It would be more informative for users of the data to have additional information of this type to enable a more in-depth understanding of the sorts of commitments that homeowners make when entering into a mortgage, the commitments that remain over time and how these differ across society. In addition, with access to information on the start date and duration of mortgages, researchers could model the interest rates that homeowners face (as these are less likely to be known by respondents).

20.4.3. Housing conditions and measures of housing deprivation

Housing deprivation is assessed through the monitoring of four key variables, meaning that indicators are available for each item and the overall severity of deprivation (e.g. the proportion of the population experiencing one, two, three or four housing deprivation items).

The current housing deprivation items considered are:

1. leaking roof, damp walls/floors/foundations, or rot in window frames or floors (HH040),
2. no bath or shower in the dwelling (HH081),
3. no indoor flushing toilet for the sole use of the household (HH091),
4. dwelling too dark, not enough light (HS160).

In relation to the leaking roof, damp walls/floors/foundations, or rot in window frames or floors variable (HH040), we found that most questionnaires collect the three components of this variable in one question; however, some countries (including France and Slovenia) separate them out into two or three questions. Both France and Slovenia reported that they obtained different results when the components were separated from when the variable was established using one question.

Data from Slovenia appear to support this. Prior to 2008, Slovenia used a single question to inform the variable, but in 2008 this was split into three separate questions. The change was justified by experience showing that Slovenian respondents (who were surveyed using computer-assisted telephone interviewing) found a long, single question difficult to understand. The results of introducing this change are shown in Table 20.1, which reveals a sharp increase in 2008 in the percentage of people reporting one or more of these problems in comparison with the percentage who reported these problems when asked using a single question.

Table 20.1: Percentage of population reporting problems related to housing conditions, Slovenia, 2007–2015

	2007	2008	…	2013	2014	2015
HH040	17.2	30.2		27.1	30.3	28.0
Leaking roof		10.6		8.7	9.4	8.7
Damp walls/floors/foundations		17.3		18.1	21.2	20.4
Rot in window frames or floors		15.9		11.3	12.3	9.6

NB: The numbers are unweighted and are therefore not representative of the Slovenian population as a whole. For 2007, all three components were collected in a single variable (HH040), while from 2008 onwards the breakdown is shown for each component as well as the percentage who report at least one of these problems (HH040).

Source: Statistical Office of the Republic of Slovenia.

Regarding HS160, EU-SILC includes a single variable intended to assess the amount of natural light available within the respondents' dwellings. Households are generally asked a question such as 'Is your dwelling too dark?'. Some questionnaires make it clear that they are referring to natural light. For example, in Sweden the question asks, 'Do you have problems in the dwelling with it being too dark in any rooms, even on a sunny day?'. Given the subjective nature of the variable, all questionnaires make reasonable attempts to establish how respondents feel about the level of light in their dwellings, with a focus on external light.

We therefore recommend that the guidelines be revised to make clear that the question refers to the amount of natural light. For example, questionnaires could ask, 'Is there a problem with the amount of natural light (daylight) accessible from within the dwelling?'.

20.4.5. Housing characteristics

The EU-SILC survey includes two variables that capture additional housing characteristics. These are:

1. dwelling type (HH010),
2. tenure status (HH021).

The EU-SILC guidance on dwelling type (HH010) lists four possible responses. These are 'detached', 'semi-detached or terraced', 'apartment or flat' and 'other kind of accommodation'.

Consultation of countries' questionnaires shows that, generally, implementation of the dwelling type variable in EU-SILC occurs either by asking respondents which of the four categories they fall into or by asking a sequence of questions designed to establish the broad category of dwelling (house versus flat/apartment versus other), before asking more questions to obtain additional details. There does not appear to be any clear benefit to using one format over the other and both are consistent with methodological guidelines.

In some cases, questionnaires collect household composition alongside dwelling type, for example including responses such as 'single-family detached'. Countries should be careful when including the term 'single family' alongside the categories of dwellings: a dwelling being detached does not imply that it is a single-family dwelling.

The responses for this variable collected in EU-SILC (Table 20.2) are a simplification of those included in UNECE's more detailed recommendations for the 2010 censuses of population and housing (UNECE, 2006).

Comparison of EU-SILC data with English Housing Survey (EHS) data shows a similar breakdown of the population by dwelling type (Table 20.3). This evidence supports the validity of the EU-SILC data, because the dwelling type variable in the EHS is provided by qualified housing surveyors.

While the combination of semi-detached and terraced houses appears to have a minimal impact on the quality of the data collected, at least in the English case the analysis of results from the EHS reveals that terraced houses are more susceptible to damp and hazards and are less likely to meet the

Table 20.2: Comparison of dwelling type categories in EU-SILC, the UNECE recommendations and the EHS

Current EU-SILC categories	UNECE recommended categories	EHS categories
1. Detached house	(1.1) Detached house (1.1.1) Detached houses with one dwelling (1.1.2) Detached houses with two dwellings (with one above the other)	Detached house or bungalow
2. Semi-detached or terraced house	(1.2) Semi-detached house	Semi-detached house or bungalow
	(1.3) Row (or terraced) house	Small terraced house: a house with a total floor area of less than 70 m^2 forming part of a block where at least one house is attached to two or more other houses
		Medium/large terraced house: a house with a total floor area of 70 m^2 or more forming part of a block where at least one house is attached to two or more other houses.
		Mid-terraced house: a house attached to two other houses in a block
		End-terraced house
3. Apartment or flat in a building with fewer than 10 dwellings	(1.4) Apartment buildings (1.4.1) Apartment buildings with three to nine dwellings	Converted flat: a flat resulting from the conversion of a house or former non-residential building; includes buildings converted into a flat plus commercial premises
		Purpose-built flat, low rise: a flat in a purpose-built block less than six storeys high; includes cases where there is only one flat with independent access in a building that is also used for non-domestic purposes
4. Apartment or flat in a building with 10 or more dwellings	(1.4.2) Apartment buildings with 10 or more dwellings	Purpose-built flat, high rise: a flat in a purpose-built block that is at least six storeys high
5. Some other kind of accommodation	(1.5) Other residential buildings	A room or rooms Other: caravan, mobile home or houseboat; some other kind of accommodation

Source: UNECE (2015), Eurostat (2017) and MHCLG (2021).

decent homes standard ([117]) than semi-detached houses (Table 20.4). This implies that distinguishing between terraced and semi-detached housing in EU-SILC would provide useful insight.

([117]) The decent homes standard is based on four criteria: it meets the minimum standard for housing; is in a reasonable state of repair; has reasonably modern facilities; and provides a reasonable degree of thermal comfort. See Department for Communities and Local Government (2006) for further information.

The other housing characteristic variable is tenure status (HH021). The response options for this variable are 'outright owner', 'owner paying mortgage', 'tenant or subtenant paying rent at prevailing or market rate' and 'accommodation is rented at a reduced rate (i.e. lower than market price)'.

Table 20.3: Comparison of dwelling type categories in EU-SILC and the EHS, England, 2011

Dwelling type	EU-SILC (%)	EHS (%)
Detached	24	22
Semi-detached	—	31
Terraced	—	28
Semi-detached/terraced total	56	59
Apartment/flat < 10 dwellings	12	—
Apartment/flat 10 or more dwellings	7	—
Apartments/flats total	20	20

NB: For comparability, the EU-SILC estimates in this table are based on households in England; households in other parts of the United Kingdom have been excluded.

Source: EU-SILC user database, 2011, and EHS, 2011.

Table 20.4: Percentage of households with problems with damp or hazards or failing the decent homes standard by type of dwelling, England, 2011

Dwelling type	Damp	Hazards	Fails decent homes standard
Terraced houses	6.7	18.7	28.0
Semi-detached houses	2.6	15.6	23.4

Source: Authors' own calculations and EHS, 2011.

Commentary from countries indicates that respondents find it difficult to establish if they are renting at or below the market rate. Most countries use self-assessment as the prevailing method of differentiating between the two responses. Some countries, however, make assumptions, for example they categorise respondents as renting at below the market rate if they are renting in the social housing sector.

Further to capturing information concerning tenure status, the *Enquête Logement*, EHS and UNECE each requests information concerning type of ownership. In the case of UNECE, the response options are 'owner-occupied', 'cooperative ownership', 'rented dwelling (private ownership)', 'rented dwelling (owned by the local or central government and/or by non-profit organisations)', 'rented dwelling (mixed ownership)' and 'other types of ownership'. Were this information to be collected in EU-SILC, powerful insights could be made concerning the effect of the ownership model on the cost and quality of housing.

Interestingly, the *Enquête Logement* establishes the tenure status of each separate individual within the household. This allows more complex living arrangements to be distinguished in the data.

20.5. Conclusion

Although the variables gathered in EU-SILC are mostly consistent across countries, more work needs to be carried out to ensure that variables are interpreted in the same way regardless of the language of implementation.

Issues have also arisen in more complex cases, where multiple questions feed into a single variable. Future research should seek to establish a common framework for gathering the data that feed into these variables. This would need to be sufficiently flexible to accommodate different housing support systems.

In terms of external validity, overall the EU-SILC variables aligned well with the 2011 censuses, although within countries large variations were occasionally observed. Discussions on why these occurred for individual countries go beyond the scope of this

chapter, but these comparisons could provide a useful tool for further validity testing.

In Table 20.5 we provide a summary of recommendations. We have sought to provide recommendations where the rationale is clear. These have already been adopted by many countries, with priority given to changes to the guidance for existing variables.

However, where a particularly strong case can be made and for newer priority areas, such as adaption to needs and fuel poverty, we recommend some new variables that are not widely adopted by countries. These proposals, in particular, represent a greater collection burden and will need to be carefully considered within the wider process of prioritisation alongside our other recommendations.

Table 20.5: Summary of recommendations

Variable	Recommendation	Priority/frequency
Number of rooms (HH030)	Countries should review collection to ensure that implementation reflects the guidelines.	Medium
	Revise guidelines to include kitchens where they meet the minimum volume requirements, count only rooms that are for the sole use of the household and specify that open-plan spaces should be counted as if there are walls between the different areas.	Medium
	Introduce a new variable capturing the number of bedrooms available to households. Over the long term, this could replace the number of rooms variable, as it is much easier for respondents to estimate and is less susceptible to error, especially for shared or open-plan dwellings.	Medium/yearly
Size of dwelling in square metres (HC020)	Redesign the variable to refer to the 'useful floor space' of the dwelling.	Medium/3-yearly
Shortage of space (HC010)	Include the variable in a 3-yearly rolling module.	High/3-yearly
	Provide further guidance on phrasing to ensure that the variable is collected consistently across countries.	Low
Current rent related to occupied dwelling (HH060)	Improve clarity with regard to the inclusion of any housing allowances, whether paid by third parties or the respondent household directly.	Low
	Countries should make explicit that any payments for the use of a garage to provide parking in connection with the dwelling should be included.	Low
	Consider including and testing a new variable capturing whether a respondent household's rented dwelling is furnished or unfurnished.	Low/6-yearly
Total housing costs (HH070)	Collect utility costs separately from total housing costs. This could generate greater insight into fuel poverty and housing deprivation.	High/yearly
	Countries should consider the reference period specified to respondents so that estimates are less affected by seasonality or unfamiliar periods of reference.	Medium
	Revise guidance so that it explicitly states which types of costs are included within the definition of total housing costs.	Medium
Mortgage principal repayment (HH071)	Revise guidance on this variable to provide more consistency and generate insight from existing data collections regarding mortgage repayments. For instance, explicitly request the collection of the amount paid in interest, the start year of the mortgage, the mortgage duration and the principal amount borrowed.	Medium
Dampness and related structural damage (HH040)	Consider including separate variables for each component part of the leaking roof, damp walls/floors/foundations, or rot in window frames or floors variable to encourage consistency.	High
Housing facilities (HH081 and HH091)	Countries should ensure that their questionnaires refer to the presence of a bath or shower in the dwelling, rather than a bathroom (which is open to misinterpretation).	Medium

Variable	Recommendation	Priority/frequency
Problems with the dwelling: too dark, not enough light (HS160)	Revise the guidelines and format of the question to make it clear that the question refers to the amount of natural light in the dwelling.	Medium
Noise (HS170)	Countries should ensure that their questionnaires refer to noise from outside that can be heard within the dwelling (e.g. nearby train lines or overhead planes) rather than just noise from the street.	Medium
	Replace response categories with a non-binary scale describing the extent to which noise is felt to be a problem for the household.	Low
Pollution, grime and environmental problems (HS180)	Countries should ensure that specific reference to causes of pollution (e.g. traffic, industry) are not made, so that respondents are aware that their responses should be based on the existence of pollution, grime or other environmental problems independently of the cause of these problems.	Medium
	Replace response categories with a non-binary scale describing the extent to which respondents feel that pollution, grime or other environmental issues are a problem for the household.	Medium
Crime, violence and vandalism (HS190)	Replace response categories with a non-binary scale describing the extent to which respondents feel that crime, violence or vandalism is a problem for the household.	Medium
Dwelling type (HH010)	Replace the existing dwelling type variable with a new six-response variable that separates semi-detached and terraced dwellings into individual categories.	Medium
	Countries should review questionnaires to ensure that two-family houses are appropriately categorised as houses rather than apartments.	Low
Tenure status (HH021)	Consider amending response guidance to allow employer- and family-/acquaintance-owned dwellings to be captured separately.	Low
	Consider whether tenure status should and could be captured at personal level, given that tenure status often varies within households.	Medium
Fuel poverty	Collect utility costs separately from total housing costs. This could generate greater insight into fuel poverty and housing deprivation.	High/yearly
	Consider including a subjective variable asking whether any household members are particularly vulnerable to extremes of temperature for age or health reasons.	High/3-yearly
	The variable on ability to keep home adequately warm should account for seasonality. We suggest wording as either 'Ability to keep home adequately warm in the winter' or 'Ability to keep home adequately warm in the winter and cool in the summer'.	Medium/3-yearly
Adaptation to needs	Add a variable capturing whether the dwelling has been appropriately adapted to users' requirements (e.g. ramps, wide corridors/hallways, handrails on steps).	Medium/6-yearly

References

Bonnefoy, X. (2007), 'Inadequate housing and health: an overview', *International Journal of Environment and Pollution*, Vol. 30, No 3/4, pp. 411–429.

Briant, P., Donzeau, N., Marpsat, M., Pirus, C. and Rougerie, C. (2010), *Le dispositif statistique de l'Insee dans le domaine du logement – État des lieux et évaluation comparée des sources*, National Institute of Statistics and Economic Studies (https://www.insee.fr/fr/statistiques/1380822).

Clark, C. and Bowen, R. (2018), 'Suggestions for the improvement of housing indicators in EU_SILC and capturing new dimensions of housing disadvantage', available on request from the authors.

Department for Communities and Local Government (2006), *A Decent Home: Definition and guidance for implementation, June 2006 – Update*, Department for Communities and Local Government, London.

Eurostat (2017), *Methodological Guidelines and Description of EU_SILC Target Variables – DocSILC065 – 2017 operation (version September 2017)* (https://ec.europa.eu/eurostat/documents/203647/203704/Guidelines+SILC+2018/).

Iacovou, M., Kaminska, O. and Levy, H. (2012), 'Using EU-SILC data for cross-national analysis', *Institute for Social and Economic Research Working Papers*, No 2012-03, University of Essex, Colchester.

MHCLG (2021), English Housing Survey, 2019 to 2020: Home ownership, Ministry of Housing, Communities and Local Government, London. https://www.gov.uk/government/statistics/english-housing-survey-2019-to-2020-home-ownership

Pittini, A. (2012), 'Housing affordability in the EU: current situation and recent trends', *Cecodhas Housing Europe Observatory Research Briefs*, Vol. 5, No 1.

Sunega, P. and Lux, M. (2016), 'Subjective perception versus objective indicators of overcrowding and housing affordability', *Journal of Housing and the Built Environment*, Vol. 31, pp. 695–717.

UNECE (2006), *Recommendations for the 2010 Censuses of Population and Housing*, United Nations, New York and Geneva (https://unece.org/fileadmin/DAM/stats/publications/CES_2010_Census_Recommendations_English.pdf)

UNECE (2015), *Recommendations for the 2020 censuses of population and housing*, United Nations, New York (https://unece.org/DAM/stats/publications/2015/ECECES41_EN.pdf)

21 Methods for collecting data on production for own consumption

Tijana Čomić ([118])

21.1. Introduction

This chapter summarises the different approaches to data collection on the value of goods produced for households' own consumption (own consumption products (OCPs)) in European Union Statistics on Income and Living Conditions (EU-SILC). Although not part of the total disposable household income (TDHI) concept used for the computation of EU social indicators, data on the value of goods produced for own consumption are collected by most countries. However, there is no standard methodology for data collection, which causes cross-national data comparability issues. This chapter describes the national practices (in EU Member States and non-EU countries) used to collect own consumption data from households/individuals and how own consumption is valued/monetised, with the aim of suggesting the most efficient approach that might be accepted by all countries.

As a part of their Net-SILC3 research on the comparability of EU-SILC income variables, Goedemé and Zardo Trindade (2020) collected detailed information from EU-SILC countries on the methods used for data collection and compiled the MetaSILC 2015 database). Further details of this data collection and the countries that participated are presented in Section 21.4. Based on the information collected, in this chapter all Member States and some non-EU EU-SILC countries are divided into two categories:

1. countries that do not collect data on own consumption through EU-SILC; and
2. countries that do collect data on own consumption through EU-SILC.

For countries that do not collect data on own consumption, the reasons for this are discussed. In this part of the analysis, when possible, both EU-SILC quality reports and information available in the MetaSILC 2015 database are used.

For countries that do collect own consumption data using EU-SILC, the methods used for data collection are described, based on the MetaSILC 2015 database.

21.2. Why are own consumption products data needed?

Data on production for own consumption are collected in several surveys for different reasons. At the macro level, one of the basic economic distinctions is defined in the system of national accounts, in which a distinction is made between:

- establishments that are market producers,
- producers for own final use,
- non-market producers (Eurostat, 2014).

([118]) Tijana Ćomić is a researcher at the Institute of Economic Sciences, Belgrade, Serbia. Thanks go to Sofija Suvočarev for assistance with the research presented in this chapter and also Tim Goedemé, Anne-Catherine Guio, Eric Marlier and Peter Lynn for their coordination and valuable comments and suggestions. Special thanks also go to Eurostat colleagues for kindly providing extractions from the Household Budget Survey. These individuals are not responsible in any way for the contents of this chapter and all errors are the author's responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Tijana Ćomić (tijana.comic@gmail.com).

While in market establishments goods and services are mostly produced for sale at prices that are economically significant, in non-market establishments most of the goods and services are produced without charge or at prices that are not economically significant.

The third part of the economy is in fact production for own final use, in which goods and services are mostly produced for final consumption by the owners of the assets. For the purpose of this chapter, production for own final use can be divided into household sector production and business sector production. According to Ironmonger (2001), household production is the production of goods and services by the members of a household for their own consumption and using their own capital and their own unpaid labour. However, households can also use their own capital and their own labour to produce goods that are mainly market oriented but withdraw/keep part of them for their own consumption.

Before industrialisation, household production was more widespread. With the development of the market economy, production is transforming to market production and households are deciding to purchase goods and services rather than producing them themselves. From an economic point of view, household production requires the engagement of own resources – mainly labour but also capital, as it requires certain investments as well. Therefore, the difference between the resources engaged and the value of the produced goods is becoming smaller and smaller. Market economics has enforced economies of scale, which results in lower (fixed) costs of production. The household sector is not able to produce the goods (mainly food) at such low costs and it is therefore no longer economical for households to produce these goods themselves.

On the other hand, production for own consumption engages unpaid labour of household members. How do members of the household contribute to the household well-being if they are not paid in money but contribute to the household through other types of engagement? The main purpose of this chapter is to consider how production for own consumption could be included in TDHI. To do this, we need not only to quantify the production but also to monetise its value.

There is an increasing interest in monetising unpaid work, especially when it comes to gender equality – making women's work visible as, even with the increased involvement of women in the labour market, the majority of the labour needed within the household is provided by women (Ironmonger, 2001).

As its name states, a main purpose of EU-SILC is to collect data on the income of the household. Income is defined comprehensively, to include much more than just the salaries of employees, which are more or less easily collectable, especially in countries where register data are used. The biggest challenge is how to collect (i.e. monetise) non-monetary income, as it might be a significant part of overall income for some households and therefore can significantly improve the well-being of households.

Agenda 2030 calls for the eradication of extreme poverty for all people everywhere (UN General Assembly, 2015). Methodology for the calculation of poverty indicators, regardless of whether it concerns national or international poverty definitions, suggests that:

> *consumption or income data are gathered from nationally representative household surveys, which contain detailed responses to questions regarding spending habits and sources of income. Consumption, including consumption from own production, or income is calculated for the entire household.* (UNSD, n.d.).

Therefore, the importance of this component of household income is recognised worldwide and requires the commitment of national statistical institutes and the adjustment of national and international surveys in order to respond to the measurement requirements.

The following section describes the methods commonly used for collecting data on OCPs.

21.3. Methods of data collection

The two main sources of data on production for own consumption at EU level are the Household Budget Survey (HBS) and EU-SILC.

For national accounts, the HBS is considered the most reliable data source for data on OCPs. Here, data on income in kind from non-salaried activities include withdrawals from one's own garden, farm or enterprise for the household's private consumption.

> *HBS data collection involves, most of the time, a combination of one or more interviews, and diaries or logs maintained by households and/or individuals, generally on a daily basis. The period for which a diary is maintained is called the recording period. The duration and distribution of this time period is the most important determinant of the structure of the survey. The other time period that characterises the HBS is the survey period: it is the period of time for which the household consumption expenditure [collected in the interview] is recorded. The survey periods may vary from one year to multiple years (two or three years) depending of the nature of survey.* (Eurostat, 2016)

In other words, the survey period is longer than the recording period, as it is impossible to collect the data from all sampled households at the same time. Within the survey period different households have their own recording period, depending on when they start maintaining the diary. Therefore, the recording period differs between households, whereas the survey period is a characteristic of the survey as a whole.

In the HBS, data on income in kind from non-salaried activities are collected for each good withdrawn for consumption by the household. The HBS therefore contains detailed data on quantities of products used. Moreover, data are collected for a short period of time (the recording time is usually 2 weeks) and data collection does not require recall of the household members; rather, products are recorded 'as consumed'.

On the other hand, EU-SILC methodology requires recall for a reference year (usually the previous calendar year) and the data are collected through interviews, together with responses to all other questions. While the HBS records goods actually consumed, EU-SILC records all goods produced by households for their own needs and not for the market. An additional difference between the two surveys is in the treatment of withdrawals from an enterprise. In EU-SILC these are included under gross income benefits or losses from self-employment (including royalties) and are therefore excluded from the target variable on the value of goods produced for own consumption (HY170) (Eurostat, 2016). Table 21.1 summarises the main differences between the HBS and EU-SILC approaches.

The next section reviews national practices (in Member States and non-EU countries) used in EU-SILC for the collection of data on own consumption from households/individuals and how own consumption is valued/monetised.

21.4. National practices in data collection in EU-SILC

Data on national data collection practices to populate the MetaSILC 2015 database were collected by Goedemé and Zardo Trindade (2020) by means of a questionnaire. The questionnaire was distributed to 35 EU-SILC countries and was completed by 26 countries (25 Member States and one candidate country) (Figure 21.1).

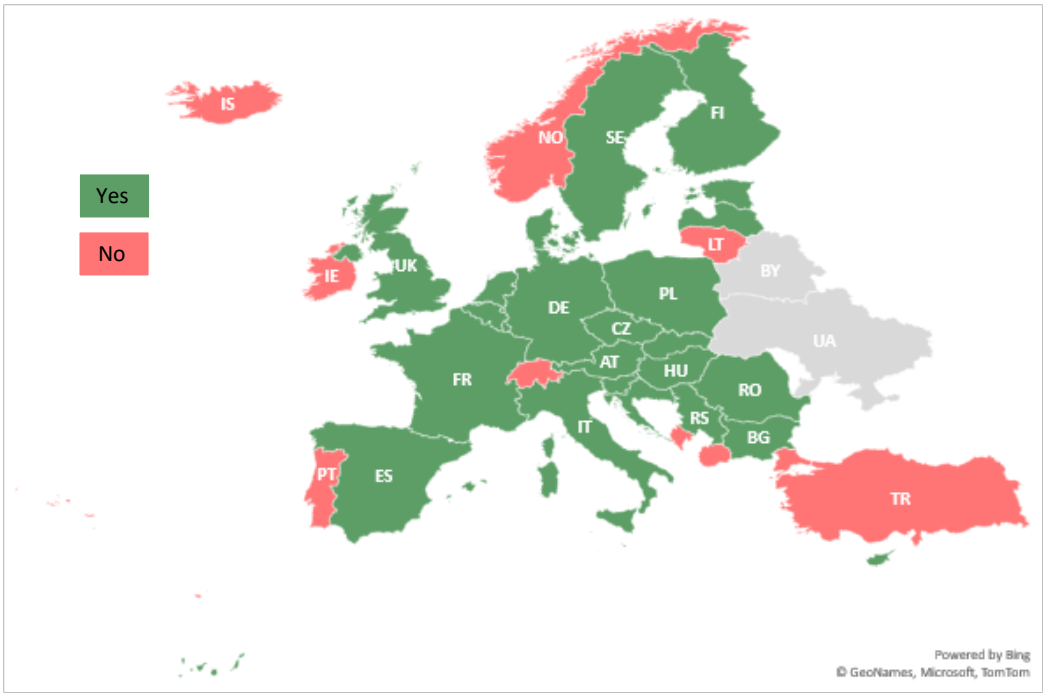
Of the 26 countries that responded to the MetaSILC 2015 survey, 17 include OCPs in HY170, four do not collect OCP data in EU-SILC but they collect them

Table 21.1: Main differences in the HBS and EU-SILC approaches to OCPs

Definition	HBS	EU-SILC
	Includes withdrawals from enterprise	Excludes withdrawals from enterprise
Method of data collection	Diary	Recall
Reference/recording period	Usually 2 weeks	Previous calendar year
Data collection period	Continuous	Up to 8 months after the reference period
Level	As consumed	Produced (harvested)

Source: Eurostat (2016, 2021).

Figure 21.1: Participation status for each country invited to participate in the MetaSILC 2015 survey



NB: Of the 35 countries that were invited to participate in the survey, 26 responded. For Iceland, Ireland, Lithuania, Norway, Portugal and Switzerland, data are available in the 2015 EU-SILC user database, but no information was provided for the MetaSILC 2015 database. In the case of Portugal, only partial information was provided. For Montenegro, North Macedonia and Turkey, data were not available in the 2015 EU-SILC user database.

Source: MetaSILC 2015 database.

in other surveys, and in five the national statistical institutes do not collect OCP data at all (Figure 21.2).

21.4.1. Countries that do not collect own consumption product data in EU-SILC

Data on own consumption are not collected in EU-SILC, but are collected in other surveys, in Belgium, Finland, Romania and the United Kingdom. In Austria, Denmark, Malta, the Netherlands, Sweden and Switzerland data on own consumption are not collected at all. The main reasons for not collecting data are the low prevalence rate of OCPs and the perception that this variable is not significant for the overall income of households. In addition, the difficulties associated with the evaluation of the value of own consumption are a factor (Table 21.2).

Figure 21.2: OCP collection status for each country that participated in the MetaSILC 2015 survey

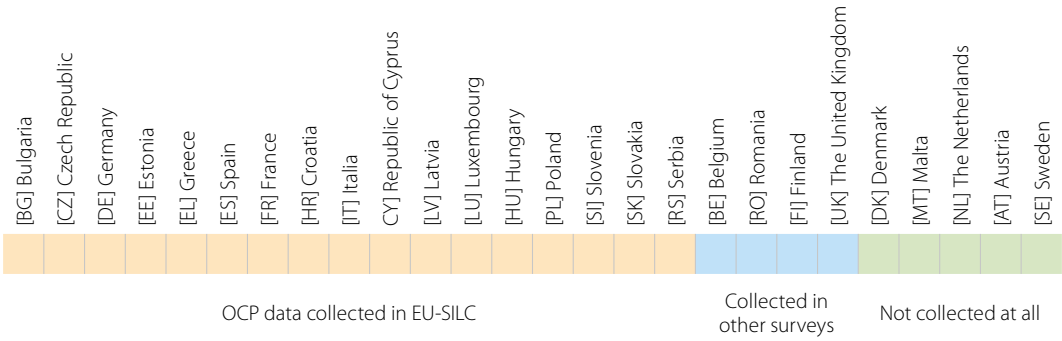


Table 21.2: Main reasons for not collecting own consumption product data in EU-SILC

Country	Main reason
Austria	The value of goods produced for own consumption is not very important for Austrian households and data are difficult to collect/evaluate.
Belgium	Based on HBS results, the value of goods produced for own consumption does not constitute a significant component of income (less than 5 %).
Denmark	Data are not collected because of the trade-off between response burden and the value of information: in this case, the response burden is higher than the value of the information. For the vast majority of people it is assessed that the value of goods produced for own consumption is very negligible compared with income. This, combined with the difficulties in accurately assessing the value of such goods, led to this question not being included in the questionnaire.
Finland	The value of goods produced for own consumption is relatively small (not significant).
Malta	The value of goods produced for own consumption is always very low.
Netherlands	It does not constitute a significant component of income.
Romania	No information available.
Sweden	No information available.
Switzerland	This variable is not collected as the value of goods produced for own consumption is not a material income component in Switzerland. According to the Federal Statistical Office HBS, in 2017 this variable represented an average of less than 0.1 % of gross income.
United Kingdom	The value of goods produced for own consumption is not asked in the United Kingdom and the variable is set to zero in the microdata. Home-grown fruit and vegetables are assumed to have a negligible benefit when calculating household income, in many cases being grown for pleasure rather than to save money. Monetary benefits may even be negative when production costs are taken into account. Data from the Living Costs and Food Survey (*) show that less than 3 % of households record this type of income and, among those that do, it accounts for less than 0.5 % of their disposable income.

(*) The Living Costs and Food Survey is the United Kingdom's name for the HBS.

Source: MetaSILC 2015 database (2018 national quality report for Switzerland).

21.4.2. Countries that do collect own consumption product data in EU-SILC

The method of data collection, in terms of formulation of the questions on own production and length and specificity of food lists, is not strictly defined in the EU-SILC methodology. The document including the methodological guidelines and description of EU-SILC target variables (Eurostat, 2017) states that the output variable HY170 has to be delivered to Eurostat but that each country can decide on the set of questions to be used to collect those data. The analysis shows that there are two main types of information that are collected: a monetary assessment by households about total income from own consumption and a self-assessment of quantities of goods produced for own consumption. Table 21.3 shows the method of data collection as reported by countries.

The majority of countries use monetary assessment. In other words, respondents are asked to estimate the value of the goods produced rather than to state the quantities produced. Furthermore, in most cases respondents are asked only to estimate the total value and not to provide separate values by food groups. Details on income by food groups are collected only by Czechia, Spain, Cyprus and Slovakia. Box 21.1 shows examples of different question formulations, including the food groups specified in Czechia.

Table 21.3: Methods used for data collection on production for own consumption

Country	Monetary value of OCPs				Quantities of goods produced	
	Point estimate (open question)		Range (closed question)		Short lists of goods (food groups)	More detailed lists (specific foods)
	By food group	Total only	By food group	Total only		
Bulgaria		✓				
Croatia		✓				
Cyprus	✓					
Czechia	✓		✓		✓	
Estonia						✓
France		✓		✓		
Germany		✓				
Greece		✓				
Hungary		✓		✓		
Italy		✓				
Latvia						
Luxembourg		✓				
Poland		✓		✓		
Portugal		✓		✓		
Slovakia	✓		✓		✓	
Slovenia						
Serbia		✓				
Spain	✓					

NB: Most countries use monetary assessment. Countries were asked first to specify whether they use the monetary value of OCPs or quantities of goods produced or both for data collection. Cells are shaded grey if a method is not used. Greece uses both methods but did not specify the approach used when collecting quantities of goods produced. Slovenia uses monetary value only but did not specify the approach used.

Source: MetaSILC 2015 database (national questionnaire for Portugal).

Box 21.1: Examples of different question formulations

1. When monetary assessment by households about **total income from own consumption** is used.

Bulgaria. 'Approximately what amount of agricultural production in total is consumed by your household?' and 'Approximately what amount of the above indicated (in terms of eggs, milk, meat, honey or other animal products) consumption by your household is own production?'

Greece. 'During the previous calendar year, did you save any income from your own/home production, such as from the production of food or drink?' and 'If yes, how much approximately did you save?'

2. When self-assessment of both **quantities and value of goods produced** is used.

Czechia. 'Can you please estimate the quantity and the value of the products that you consume from your own farm or you own business? Please indicate the amount and the value for the whole calendar year'.

Quantities of the following goods are included:

(i) meat (kilograms),
(ii) eggs (number),
(iii) potatoes (kilograms),
(iv) fruit (kilograms),
(v) vegetables (kilograms).

The question asks about the value in Czech koruna of wood grown on households' land, other food and drink, including meals, and industrial products and services.

Source: National questionnaires.

On the other hand, Estonia collects data on income from own consumption only through self-assessment of quantities of goods produced, whereas Czechia and Slovakia collect data through both monetary assessment of income and self-assessment of quantities. Only Estonia collects data on quantities using a more detailed list of goods.

For monetising the value of goods produced in Estonia, most quantities were imputed from answers provided by respondents and unit costs were taken from the HBS. Production costs were deducted from the total income obtained for own consumption goods, and the profits were transferred to the personal level.

Although Czechia and Slovakia also include questions on monetary assessment, when collecting data on quantities they use a similar method as Estonia. Namely, in Slovakia the value of food produced for own consumption is calculated on the basis of the market price of those goods after deducting the direct costs incurred in their production. The list of prices for selected food groups produced for own consumption is provided in the 'Guidelines for interviewers'. This list is created based on the HBS and is used for calculation of the value of goods produced and consumed by households. Czechia estimates the monetary value of goods based on the average price of products.

Latvia collects data on whether households consume their own food products using several questions about consumption of self-grown fruit and/or vegetables, products obtained from raising livestock for personal consumption (e.g. meat and dairy products, eggs and honey), collecting mushrooms and berries, and fishing and hunting animals for households' own needs ('yes' or 'no'). For assessing income from own consumption, HBS data are used.

21.5. Impact of type of questionnaire on estimated value of own consumption products

The challenges in collecting data on amounts (and, to a lesser extent, the value) of production for own consumption are similar to the challenges in collecting data on consumption more generally, and especially food consumption, bearing in mind that food is consumed several times every day in different forms, from different sources, etc. Smith, Dupriez and Troubat (2014) analysed the reliability and relevance of the food data collected in national household consumption and expenditure surveys and pointed out areas where surveys could be improved. Of these areas, the focus is on those already identified in the introduction when differences between EU-SILC and HBS were discussed.

21.5.1. Diary versus recall data

The recall period in EU-SILC is very long – data collection can take place up to 8 months after the end of the reference period, and the reference period is spread over a whole calendar year. The limitations of long-term memory may be a cause of significant errors in responses, which could therefore endanger the reliability of the data. When data are collected on the consumption of items that are consumed regularly, shorter recall periods are likely to provide more accurate data (e.g. 24-hour recall). However, expenditure surveys 'commonly use recall of one week or more in order to be better able to capture "usual" behavioural patterns' (Zezza et al., 2017, p. 2). There is, therefore, a trade-off between accurate measurement of potentially atypical consumption and less accurate measurement of more typical consumption. The former may be preferable if the objective is to estimate sample totals, while the latter may be preferable if the objective is to understand household/individual correlates of consumption.

According to Conforti, Grünberger and Troubat (2017, p. 49):

> *the data collection method affects food consumption measurement, as recall interviews report higher quantities compared to diaries. However, the data collection method interacts with the length of the reference period proposed to the respondents, which may lead to memory loss when long reference periods are proposed.*

Although errors due to faulty recall when collecting diary data are minimal, the major disadvantage of this method is that it is highly influenced by seasonality, which is very important with respect to consumption from own production, for obvious reasons. This brings us to another important characteristic of the data collection, namely the reference period, which is the length of time over which respondents are requested to report.

21.5.2. Reference period

While reference periods for survey data collected through diaries typically range from 24 hours to 2 weeks, the reference period for recall data can vary from 1 day to 1 year, with the latter being the case in EU-SILC. When it is necessary to capture the seasonality of consumption, or to capture consumption that is rare and therefore memorable, such as the consumption of white goods, furniture or holidays, 12-month reference periods are used. Seasonality can be 'overcome' if the sample is evenly distributed over a whole year (i.e. as in continuous surveys, such as the Labour Force Survey and the HBS) (see Conforti, Grünberger and Troubat, 2017). However, this overcomes seasonality only at the sample level – for example for estimating total consumption. It does not help with the estimation of correlates of consumption, especially for subcategories of consumption (small sample sizes in each harvest period).

The EU-SILC guidelines allow continuous data collection, but this is rare among Member States. According to the 2018 comparative quality report (Eurostat, 2020), only Ireland and the United Kingdom implement the survey throughout the whole year (12 months). Therefore, 12-month recall remains the most appropriate method to collect own consumption production in EU-SILC.

21.5.3. Consumed versus produced (harvested)

Consumption surveys that use diaries for data collection focus on goods that were consumed during the reference period. Consumption from own production tends to be subject to large variability between periods, depending on the season. On the other hand, households are more likely to know the annual amount of goods produced (harvested) than the amount of goods actually consumed. However, depending on the purpose of the survey, approaches vary. According to Zezza et al. (2017), for consumption structure and poverty analysts the focus is on the amount of money spent to acquire food, whereas for food security analysts the focus is on the amount of food available for consumption. This should be kept in mind when designing questions on own production. The question here is whether, for an income survey such as EU-SILC, it is more about the value of all goods produced for own consumption or about the value of goods actually consumed. When other income variables in the EU are observed, they relate not to the amount actually spent but to the overall amount received. A similar approach should be applied when OCPs are considered, bearing in mind that the production of these products requires the engagement of household resources (the investment of time, labour and money in production). EU-SILC should strive to measure the total value of goods produced in the reference year, even if not all those goods were also consumed. This is analogous to the treatment of other sources of income, as unspent income increases the 'wealth' of the household in savings and can be spent in the future.

21.5.4. Value or quantities of goods produced

As already mentioned, in consumption surveys the value of goods acquired is commonly captured. When using a diary, it is relatively simple to capture the price of purchases. However, considering that food consumption is an everyday activity and varies, recalling the value of each good acquired is more difficult than recalling the quantities of goods, as households are often familiar with their regular consumption habits. If the quantities of goods consumed (or produced) are known, it is relatively easy to impute the value of those goods at the data processing stage based on some simplifying assumptions and some reference data. For example, in Slovakia, a list of prices for selected food groups produced for own consumption is provided in the 'Guidelines for interviewers'. This list was created based on the HBS and is used for calculating the value of goods produced and consumed by households. However, this may be an additional source of error and, if the intention is that the same unit prices should be applied to all respondents, it would be better to ask only for quantities and apply the look-up table in a standard way at the data processing stage, rather than expecting each interviewer to do this during each interview.

21.5.5. Length and specificity of survey food lists

The lengths of survey food lists depend on the subject of the survey. Long lists and detailed questionnaires extend the length of the interviews and put the burden on respondents. On the other hand, they may aid accurate recall by respondents. Smith, Dupriez and Troubat (2014, p. 10) suggest that survey food lists should be 'sufficiently detailed to accurately capture consumption of all major food groups making up the human diet'. However, using groups that are too broad (i.e. Classification of Individual Consumption According to Purpose groups ([119]): bread and cereals, meat, fish and seafood, milk, cheese and eggs, oils and fats, fruit, vegetables, sugar, jam, honey, chocolate and confectionery, and other food products) makes the imputation of values more difficult. Therefore, food survey lists should be designed to capture the most commonly produced products for own consumption in each country, but to cover all food groups.

In Serbia in 2013 and 2014, two modules were tested for collecting data on OCPs.

1. In 2013, a module with detailed questions on quantities of products was used. The food list contained more than 20 types of products that

([119]) For further details on the Classification of Individual Consumption According to Purpose, see UNSD (2018).

are commonly produced ([120]), The difference between the retail and production prices provided by price statistics was calculated for all products to obtain the variable HY170.

2. In 2014, a module on self-assessment of the market value of all goods produced and consumed in the household and the costs of producing these was used. Unlike in 2013, the value was requested for all products together, not divided into food groups.

Further analysis of data from these two years conducted by Petrušević and Vukmirović (2016) shows that the ratio of the sum of income from OCPs of all households involved in production for own consumption to the sum of TDHI for those households (OCP/TDHI ratio) in 2013 was 2.9 %, while in 2014 this ratio was 0.9 %. The conclusion was that this decrease was not caused by a real decrease in income from OCPs but that it related to the change in methodology. Therefore, collection of data for all goods produced together underestimates income from OCPs compared with collecting data for groups of products, as people seem to undervalue their production in the former case (or overvalue the cost of production). However, the two methods also differed in whether they asked for quantities or values. This difference too may contribute to underestimation of OCP income in 2014, as we can assume that data on quantities are likely to be more accurate than market values owing to the use of recall methodology.

Similar results were found in Estonia. According to Paats and Tiit (2010), before 2007 in Estonia a simplified question was used:

> *the detailed questionnaire was implemented from 2007. Comparing the data from 2006 to 2008, an important increase appears in 2007 in the proportion of households declaring income from own-consumption: in 2006 only 11 % of households declared such income, and in 2007 this proportion jumped to 52 %.*

The conclusion was that the type of questionnaire and data collection method have an important influence on the proportion of households that declare income from own consumption.

21.6. EU-SILC versus HBS data

The latest available data from HBS across Europe are from 2010 and are therefore not fully comparable with 2015 EU-SILC data. In addition, as described in Section 21.3, HBS reports income in kind from non-salaried activities, which also includes withdrawals from an enterprise, so it is expected that HBS data will give higher values than EU-SILC data for OCPs. A comparison between 2010 HBS data and 2015 EU-SILC data was carried out for Bulgaria, Croatia, Latvia, Lithuania and Poland, as in those countries OCPs have the greatest influence ([121]) (see Chapter 22 of this book).

Estimates from both surveys are presented in Table 21.4. Although both the percentage of households with OCPs and the income from OCPs as a percentage of TDHI are lower for EU-SILC than for the HBS, no clear pattern can be seen. The difference between the percentage of households declaring income from own consumption (EU-SILC) and the percentage of households with income in kind (HBS) ranges from 6.2 percentage points in Poland to 56.9 percentage points in Lithuania. Income from own consumption as a proportion of TDHI (EU-SILC) is lower than income in kind as a proportion of TDHI (HBS) in all countries observed except for Bulgaria. However, for the methodological reasons explained above, comparisons between HBS and EU-SILC are hard to interpret.

([120]) Milk, eggs, pork meat, beef meat, poultry meat, apples, pears, plums, cherries, strawberries or raspberries or other berries, peaches/apricots, other fruits, tomatoes, cucumbers, cabbage and kale, paprika, onions, garlic, cauliflower, carrots, peas, spinach/lettuce or other greens, other fresh vegetables, melons (melons and watermelons), potatoes, honey, and other products (to be specified).

([121]) No data are available from the HBS for Serbia.

Table 21.4: Income from own consumption (2015 EU-SILC) vs income in kind (2010 HBS)

Country	% of households		% income (*)	
	Declaring income from own consumption (EU-SILC)	With income in kind (HBS)	From own consumption (EU-SILC)	In kind (HBS)
Bulgaria	23.4	61.9	5.0	3.5
Croatia	42.7	53.7	5.0	6.9
Latvia	40.8	65.4	4.3	7.0
Lithuania	17.3	74.2	5.0	7.3
Poland	18.1	24.2	3.4	4.9

(*) For households declaring OCPs / in-kind income.

Source: For HBS data, Eurostat; for EU-SILC, author's computation.

21.7. Conclusions and recommendations

The complexity of collecting OCP data increases the cost of data collection, both in material terms and when overburdening respondents is considered. In addition, the quality of the data collected is very questionable, considering the long recall period and other methodological constraints discussed earlier.

Furthermore, variation between countries in the methods of data collection causes comparability issues with regard to the HY170 variable. Variation is observed in the groups of products about which information is collected, and in whether respondents are asked to provide quantities (with subsequent conversion to monetary values) or an assessment of the monetary value of the goods produced.

However, the comparability issue could be overcome by standardising the methodology for data collection. Our analysis suggests that standardisation of methodology would require the following.

- Twelve-month recall data should be collected for the income reference period.
- The quantities of products produced (not only consumed) should be collected.
- A detailed food list should be used that covers the most commonly produced products for own consumption in each country and that covers all food groups (by COICOP classification).

Of course, these questions should be filtered out for all households that are not involved in production for own consumption; this will limit the number of households that need to be asked this set of questions.

- Imputation of values should be carried out during the data processing stage, using price statistics.

Comparing EU-SILC data with HBS data is not informative regarding the effects of differences in methods of data collection and therefore conclusions cannot be drawn on which method provides the most reliable data. Some earlier studies have shown that collecting data for groups of products separately, rather than asking a single overall question, increases the percentage of households reporting being involved in production for own consumption and also the OCP/TDHI ratio, probably because it improves recall of all products produced and consumed by households.

References

Conforti, P., Grünberger, K. and Troubat, N. (2017), 'The impact of survey characteristics on the measurement of food consumption', *Food Policy*, Vol. 72, pp. 43–52.

Eurostat (2014), *Essential SNA: Building the basics*, Publications Office of the European Union, Luxembourg, doi:10.2785/51610.

Eurostat (2016), *Methodological Guidelines and Description of EU-SILC Target Variables – DocSILC065 (2015 operation)*, Eurostat, Luxembourg.

Eurostat (2017), *Methodological Guidelines and Description of EU_SILC Target Variables – DocSILC065 – 2017 operation (version September 2017)* (https://ec.europa.eu/eurostat/documents/203647/203704/Guidelines+SILC+2018/).

Eurostat (2020), *EU-SILC Comparative Quality Report 2018* (https://circabc.europa.eu/sd/a/f0a9e3a9-0381-4678-b9f0-7def2f8682a9/2018%20EU%20SILC%20ESQRS.zip).

Eurostat (2021), *Household Budget Survey 2015 – Scientific-use files: User manual*, version 1.2 (https://ec.europa.eu/eurostat/documents/203647/7610424/HBS+User+Manual.pdf/fb5d8371-08fe-4ecf-bca6-b40984fde0b6?t=1624343433403).

Goedemé, T. and Zardo Trindade, L. (2020), 'MetaSILC 2015: a database on the contents and comparability of the EU-SILC income variables [data file]' (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TLSZ4S).

Ironmonger, D. (2001), *Household Production and the Household Economy*, research paper, Department of Economics, University of Melbourne, Melbourne.

Paats, M. and Tiit, E. M. (2010), 'Income from own-consumption', in Atkinson, A. B. and Marlier, E. (eds), *Income and Living Conditions in Europe*, Eurostat, Luxembourg, pp. 180–194.

Petrušević, M. and Vukmirović, A. (2016), 'Impact of income from manufacture of direct use goods on the degree of risk from poverty and inequality in the Republic of Serbia', in *Secondary analyses of data obtained through the Survey of Income and Living Conditions (SILC)*, Social Inclusion and Poverty Reduction Unit, Government of the Republic of Serbia, pp. 65-86.

Smith, L., Dupriez, O. and Troubat, N. (2014), 'Assessment of the reliability and relevance of the food data collected in national household consumption and expenditure surveys', *International Household Survey Network Working Papers*, No 008 (http://www.ihsn.org/sites/default/files/resources/IHSN_WP008_EN.pdf).

Social Protection Committee (2015), *Portfolio of EU social indicators for the monitoring of progress towards the EU objectives for social protection and social inclusion – 2015 update*, European Commission, Brussels.

UN General Assembly (2015), *Transforming Our World: The 2030 Agenda for Sustainable Development* (http://www.refworld.org/docid/57b6e3e44.html).

UNSD (United Nations Statistics Division) (2018), *Classification of Individual Consumption According to Purpose (COICOP) 2018* (https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf).

UNSD (n.d.), 'SDG indicators – metadata repository, Indicator 1.2.1: proportion of population living below the national poverty line, by sex and age' (https://unstats.un.org/sdgs/metadata/).

Zezza, A., Carlettoa, C., Fiedlerc, J. L., Gennarib, P. and Jolliffea, D. (2017), 'Food counts. Measuring food consumption and expenditures in household consumption and expenditure surveys (HCES) – introduction to the special issue', *Food Policy*, Vol. 72, pp. 1–6.

22 The impact of own consumption on income distributions and key EU income-based indicators

Tijana Čomić ([122])

22.1. Introduction

The purpose of this chapter is to estimate the effects of inclusion of the value of goods produced for own consumption (own consumption products (OCPs)) as a component of household disposable income on the main social indicators that have been agreed at EU level ([123]) and thereby to help inform debate about whether this source of 'income' should in future be included in the income measure used to construct social indicators.

Although data on the value of OCPs are not a part of the total disposable household income (TDHI) concept used for the computation of EU social indicators, they are collected by most European Union Statistics on Income and Living Conditions (EU-SILC) countries. However, there is no standard methodology for data collection. This causes data comparability issues, as described in Chapter 21.

The main income-based EU social indicator is the at-risk-of-poverty (AROP) rate (Social Protection Committee, 2015). This indicator is a measure of relative poverty, as the poverty risk line is set at 60 % of the national median equivalised household disposable income.

In addition to the AROP rate, the portfolio of EU social indicators includes several other income inequality measures such as the Gini coefficient and the interquartile ratio (S20/S80) of total disposable household income (TDHI). The data source for all these income-based indicators is EU-SILC.

Although the concept of total household disposable income used in these indicators includes main cash and non-cash incomes, it does not include all income components; in particular, the value of OCPs (non-monetary income) and imputed rent are not part of the EU concept of total household disposable income. Whether this is the most appropriate approach has been the focus of many discussions regarding both EU statistics and national purposes. Sauli and Törmälehto (2017) analysed the impact of imputed rent and concluded that adding imputed rent into the measure of income reduces relative inequality and increases average income levels. AROP rates and rates of at risk of poverty or social exclusion fall in a majority of countries when imputed rent is included, although there are a few countries where the opposite effect is seen.

As defined in the *Methodological Guidelines and Description of EU-SILC Target Variables* (DocSILC065) (Eurostat (2016, p. 234), which assists Member States in the preparation of the EU-SILC operation:

> *The value of goods produced for own consumption refers to the value of food and beverages produced and also consumed within the same household.*

([122]) Tijana Ćomić is a researcher at the Institute of Economic Sciences, Belgrade, Serbia. Thanks go to Sofija Suvočarev for assistance with the research presented in this chapter and also Tim Goedemé, Anne-Catherine Guio, Eric Marlier and Peter Lynn for their coordination and valuable comments and suggestions. Special thanks also go to Eurostat colleagues for kindly providing extractions from the Household Budget Survey. These individuals are not responsible in any way for the contents of this chapter and all errors are the author's responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Tijana Ćomić (tijana.comic@gmail.com).

([123]) For a detailed presentation of the EU portfolio of social indicators, see Social Protection Committee (2015).

The value of goods produced for own consumption shall be calculated as the market value of goods produced deducting any expenses incurred in the process of production.

It excludes:

- *Value of household services,*
- *Any production for sale and any withdrawals from a business by a self-employed person (these values are included under 'Gross income benefits or losses from self-employment' (including royalties) (PY050G)).*

This variable 'value of goods produced for own consumption (HY170)' only refers to alimentation products (food and beverages). Other products which can be used for own consumption, like wood, should be, according to the EU-SILC Regulations, excluded from this variable.

The motivation for collecting information on the value of OCPs is the assumption, tested in this chapter, that OCPs are concentrated in lower income quintiles, in rural areas and in less developed regions. If so, OCPs could compensate for low income in vulnerable households and push their members above the minimum threshold. There is no previous research that estimates the impact of own consumption on vulnerable subgroups of the population.

According to DocSILC065, 'the value of food and beverages shall be included [collected] when they are a significant component of the income at national level or they constitute a significant component of the income of particular groups of households'. However, although the collection of some income components has been mandatory from 2007 onwards, including HY170G/N, these components will not be included in the computation of the aggregated income variables and in the computation of the EU indicators until a final decision of the EU Social Protection Committee's Indicators Sub-Group concerning the inclusion of these components has been taken. Paats and Tiit (2010) analysed the impact of OCP data, but only at the national level. Because of the many constraints (comparability of data, lack of common methodology, refusal of some countries to collect these data, imprecision of results, the burden of collecting, cleaning, analysing and processing data) and based on the assumption that the impact of OCPs on the economic situation of households is very low in most Member States, they suggested that data collection on OCPs should not be included in the subsequent programme of EU-SILC. The main argument of Paats and Tiit (2010) for not including OCPs in TDHI was that this income is not a significant component. They found that OCPs are an important proportion of total disposable income only in Romania, where they constitute 18 % of equivalised income and reduce the risk of poverty by 3 percentage points. In other countries, OCPs have only a marginal effect on the main social indicators. However, this analysis was conducted at the national level. It is possible that the effect of own consumption should not be neglected for lower income deciles, in rural areas and in less developed regions.

The data used in this chapter are EU-SILC cross-sectional data (user database) for 2015. In order to assess the importance and distribution of OCPs, two types of analysis are conducted. First, descriptive analysis is presented that explores country variations in levels of OCPs as well as variations in levels of OCPs for different subpopulations within countries. The main purpose of this part of the analysis is to identify differences between countries in the prevalence of OCPs and to understand the profile of households involved in production for own consumption to test the hypothesis that vulnerable subpopulations are more likely to be involved as a way of improving households' economic situation. Second, methodological analysis provides estimates of the impact of OCPs on key social indicators (which is the main focus of the chapter), both at country level and at the level of subpopulations.

Based on the percentage of households reporting OCP income, countries are divided into two categories, using an arbitrary threshold:

1. countries where 15 % or more of households report OCP income;
2. countries where less than 15 % of households report OCP income.

Data presented in this chapter may differ from published data due to the exclusion of some households that had unclear data in the user database.

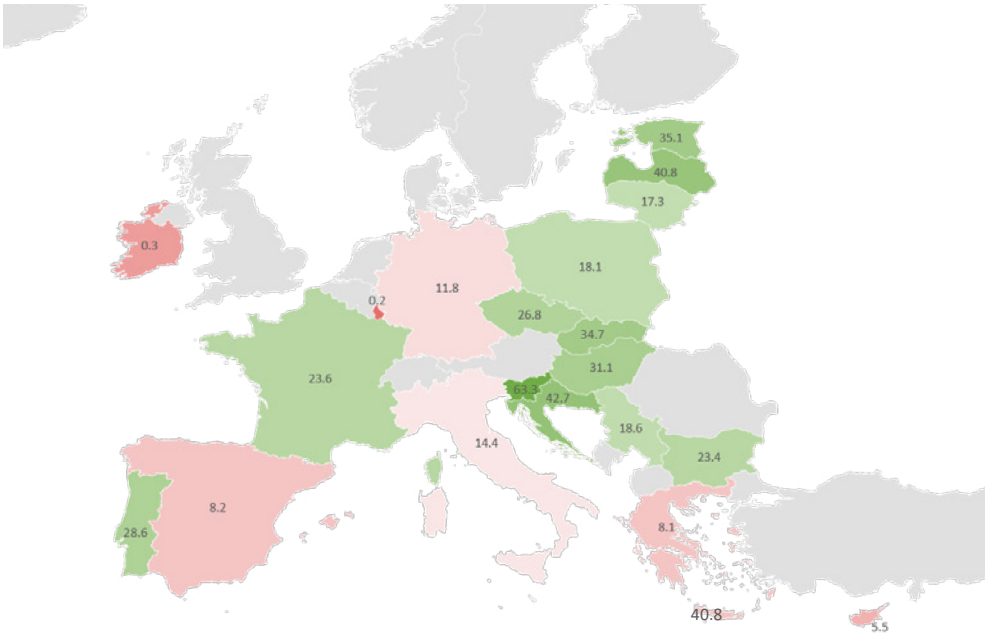
Considering the main objectives of this chapter, in Section 22.2 we describe the prevalence of OCPs in different subpopulations and the contribution of OCPs to TDHI. Moreover, we estimate the influence of OCPs on the main EU social indicators.

22.2. Profile of households reporting involvement in production for own consumption

Figure 22.1 shows how the percentage of households reporting OCP income varies between countries. The figure distinguishes between countries with less than 15 % and countries with 15 % or more of households reporting OCP income. With the exception of France and Portugal, as well as Serbia, which is a candidate country, all countries where more than 15 % of households reported OCP income are countries that joined the EU after 2004 ([124]). In these countries, income is generally lower.

Analysis presented in Figure 22.2 shows that, regardless of the percentage of households reporting some income from OCPs, the ratio of OCP income to TDHI (HY020) is very small for all countries, reaching a maximum of 2 % in Croatia.

Figure 22.1: Percentage of households reporting non-zero OCP income in countries that collected OCP data, 2015

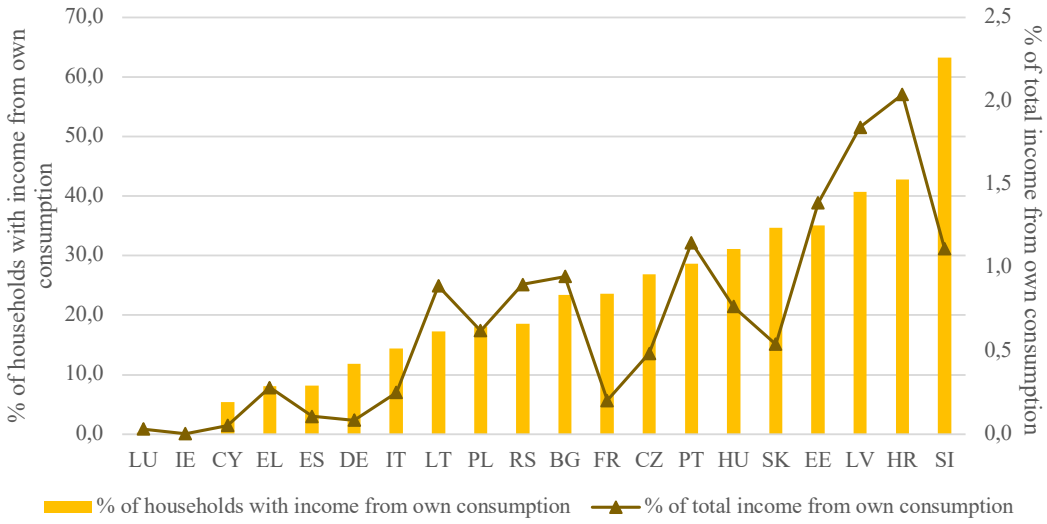


NB: Red shading represents countries with less than 15 % of households reporting OCP income, while green shading represents countries with more than 15 % of households reporting OCP income. The majority of countries marked in grey that joined the EU before 2004 would probably report low percentages of households with OCP income, as these countries did not collect OCP data as OCPs are assumed to be unimportant (as described in Chapter 21). In 2015, the percentage of households that reported OCP income was greater than 15 % in all of the countries that joined the EU in 2004 or later (except Cyprus – and Romania, which did not collect OCP data).

Source: Author's computation based on data from the EU-SILC user database, 2015.

([124]) Countries that joined the EU after 2004 are Czechia, Estonia, Cyprus, Latvia, Lithuania, Hungary, Malta, Poland, Slovenia and Slovakia (2004); Bulgaria and Romania (2007); and Croatia (2013).

Figure 22.2: Percentage of households with income from own consumption and percentage of total income from own consumption, 2015



NB: The figure is restricted to countries that collect OCP data in EU-SILC. Countries are presented in ascending order of percentage of households with income from own consumption. The figure shows, for example, that, in 2015, 63 % of Slovenian households reported that they draw an income from own consumption, but the share of own consumption income in the total income for all Slovenian households is about 1.1 %.

Source: Author's computation based on data from the EU-SILC user database, 2015.

Bearing in mind the working hypothesis that OCP income is concentrated in lower income deciles, in rural areas and in less developed regions, additional analysis aims to assess the profile of households involved in production for own consumption and to estimate the OCP/TDHI ratio for specific categories, focusing on countries where more than 15 % of households are involved in production for own consumption.

The share of TDHI derived from OCPs (s) for each household, calculated as:

$$s_i = \frac{OCP_i}{TDHI_i};\ i = 1,..,n$$

where n is the total number of households, reveals the importance of this income in TDHI (without OCP) for each household. As shown in Figure 22.3, in all countries observed, in most of the households this share is under 5 %, not including the majority of households that have no OCP income.

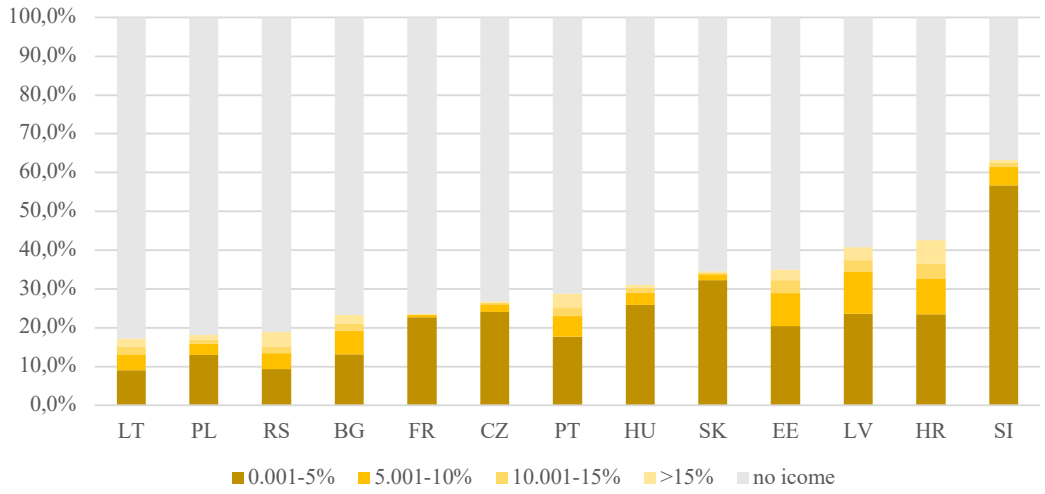
The OCP/TDHI ratio(S) at the country level – expressed as a percentage – is calculated as the ratio of total income from OCPs across all households involved in production for own consumption to total TDHI for those households:

$$S = \frac{\sum_{i=1}^{n_1} OCP_i}{\sum_{i=1}^{n_1} TDHI_i} \times 100$$

where n_1 is the number of households with income from OCPs.

The OCP/TDHI ratio for households that are involved in production for own consumption ranges from 0.8 % in France to 5.7 % in Serbia (Figure 22.4). It is obvious that this ratio is low even among households that are involved in OCP (Figure 22.3) and not only at the overall population level.

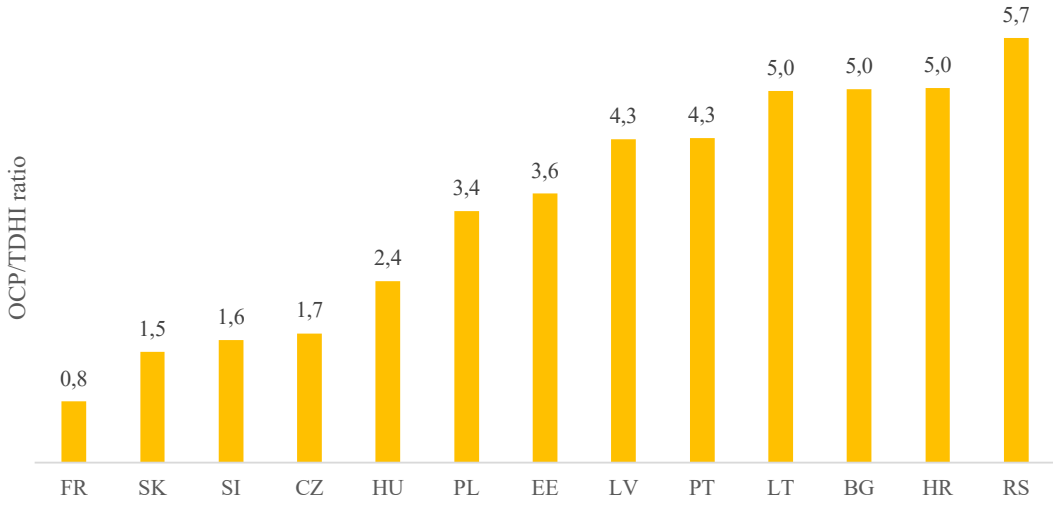
Figure 22.3: Percentage of households by share of income from OCPs in TDHI, 2015



NB: The figure is restricted to countries in which 15 % or more of households report OCP income in EU-SILC. Countries are presented in ascending order of the percentage of households with a share of income from OCPs. The figure shows, for example, that, in 2015, 63 % of Slovenian households reported that they draw an income from own consumption, while 37 % had no OCP income; 57 % of households had OCP income that was below 5 % of their TDHI. In Czechia, France, Hungary, Slovenia and Slovakia the majority of households reporting OCP income had OCP income that was below 5 % of their TDHI.

Source: Author's computation based on data from the EU-SILC user database, 2015.

Figure 22.4: OCP/TDHI ratio among households reporting OCP income, 2015



NB: The figure is restricted to countries in which 15 % or more of households report OCP income in EU-SILC. Countries are presented in ascending order of OCP/TDHI ratio.

Source: Author's computation based on data from the EU-SILC user database, 2015.

22.2.1. Production for own consumption by degree of urbanisation

Our working hypothesis is that households in thinly populated areas are more likely to be involved in production for own consumption. The 'Degree of urbanisation' (DEGURBA) ([125]) is a classification that indicates the character of an area. Based on the proportion of the local population living in urban clusters and in urban centres, DEGURBA classifies local administrative units into three types of area:

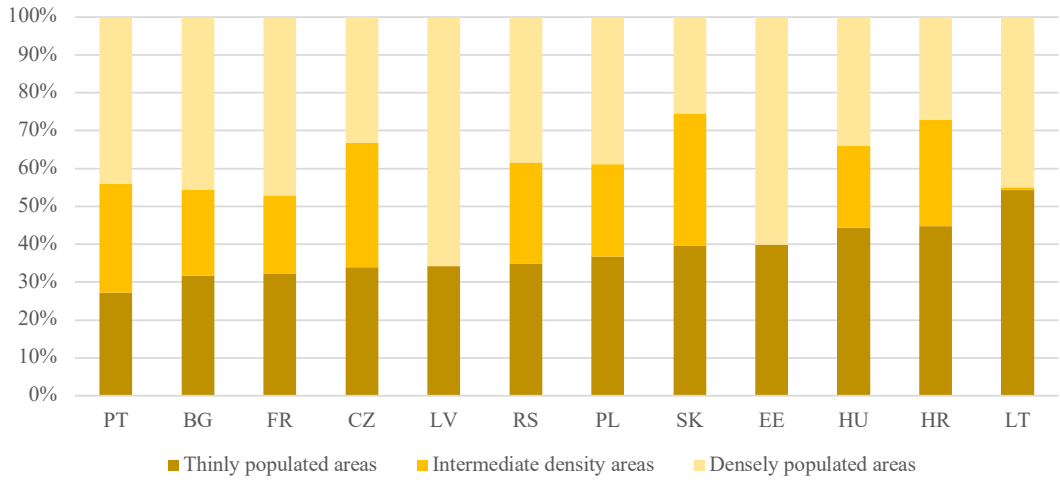
- cities (densely populated areas),
- towns and suburbs (intermediate density areas),
- rural areas (thinly populated areas).

Statistics by degree of urbanisation provide an analytical and descriptive lens on urban and rural areas.

Figure 22.5 shows that in Estonia and Latvia there are no intermediate density areas. The percentage of households living in thinly populated areas is highest in Lithuania (54 %) and lowest in Portugal (27 %).

Table 22.1 shows the percentage of households in different subpopulations involved in production for own consumption. It confirms the hypothesis that households are more involved in production for own consumption if living in thinly populated areas. This is the case in all countries. The proportion of households involved in production for own consumption is highest in thinly populated areas in Latvia, at 69 %.

Figure 22.5: Percentage of households by degree of urbanisation, 2015



NB: The figure is restricted to countries in which 15 % or more of households report OCP income in EU-SILC. The degree of urbanisation is not available in the user database for Slovenia. Countries are presented in ascending order of share of the population in thinly populated areas. The figures shows, for example, that Estonia and Latvia do not have local populations living in intermediate density areas and that more than half of households in Lithuania live in thinly populated areas.

Source: Author's computation based on data from the EU-SILC user database, 2015.

([125]) As defined by the Directorate-General for Regional and Urban Policy of the European Commission, but see also Dijkstra and Poelman (2014).

Table 22.1: Percentage of households with OCP income by degree of urbanisation, 2015

Country	Degree of urbanisation		
	Densely populated areas	Intermediate density areas	Thinly populated areas
Latvia	26.3	—	68.8
Croatia	11.0	44.9	61.1
Slovakia	13.5	22.9	58.4
Bulgaria	3.3	16.1	57.8
Portugal	15.8	26.5	52.2
Estonia	25.2	—	50.0
Hungary	10.4	26.2	49.3
Czechia	9.8	23.3	47.0
France	10.3	22.9	43.5
Serbia	2.7	15.9	39.2
Poland	2.7	11.4	38.9
Lithuania	4.1	7.6	28.4

NB: The table is restricted to countries in which 15 % or more of households report OCP income in EU-SILC. The degree of urbanisation is not available in the user database for Slovenia. Countries are presented in descending order of the proportion of households reporting OCP activity in thinly populated areas. The table shows, for example, that in Latvia 69 % of households in thinly populated areas have an income from own consumption.

Source: Author's computation based on data from the EU-SILC user database, 2015.

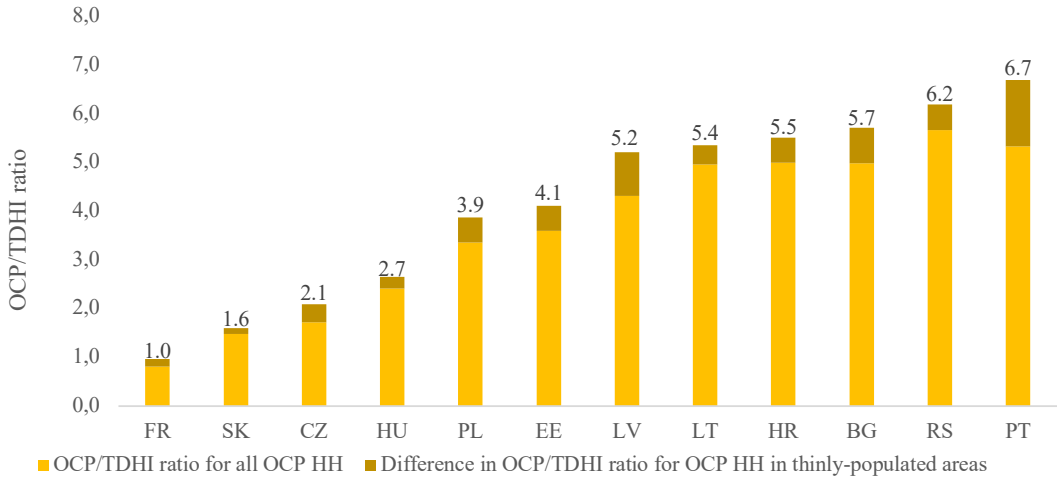
Consequently, of all households involved in production for own consumption, as expected, the majority are in thinly populated areas. The percentage ranges from 89 % in Lithuania to 49 % in Portugal.

This can be explained as follows. In Lithuania, only 28 % of households living in thinly populated areas have an income from own consumption. However, Lithuania is the country where most households (54 %) live in thinly populated areas; therefore, of all households having an income from own consumption, the proportion in thinly populated areas is actually highest in Lithuania. Therefore, in analysing the possible influence of income from own consumption on the overall income of households in different countries, we need to understand how the distribution of households between different areas varies between countries.

In all countries the concentration of households involved in production for own consumption in thinly populated areas is greater than the concentration of all households in thinly populated areas, although the extent to which this is true varies. For example, in the case of Latvia, only 34 % of all households are located in thinly populated areas but 69 % of those households have OCP income.

The OCP/TDHI ratio is slightly higher among households in thinly populated areas with OCP income than among all households with OCP income. This difference is highest in Portugal (1.4 percentage points) and Latvia (0.9 percentage points) (Figure 22.6).

Figure 22.6: OCP/TDHI ratio among households reporting OCP income living in thinly populated areas and difference from OCP/TDHI ratio among all households reporting OCP income, 2015



NB: The figure is restricted to countries in which 15 % or more of households report OCP income in EU-SILC. The degree of urbanisation is not available in the user database for Slovenia. Countries are presented in ascending order of OCP/TDHI ratio.

Source: Author's computation based on data from the EU-SILC user database, 2015.

22.2.2. Production for own consumption by region

Among the countries in which 15 % or more of households report OCP income, only a few provide data by Nomenclature of Territorial Units for Statistics 1 regions: Bulgaria, France, Hungary and Poland (Table 22.2).

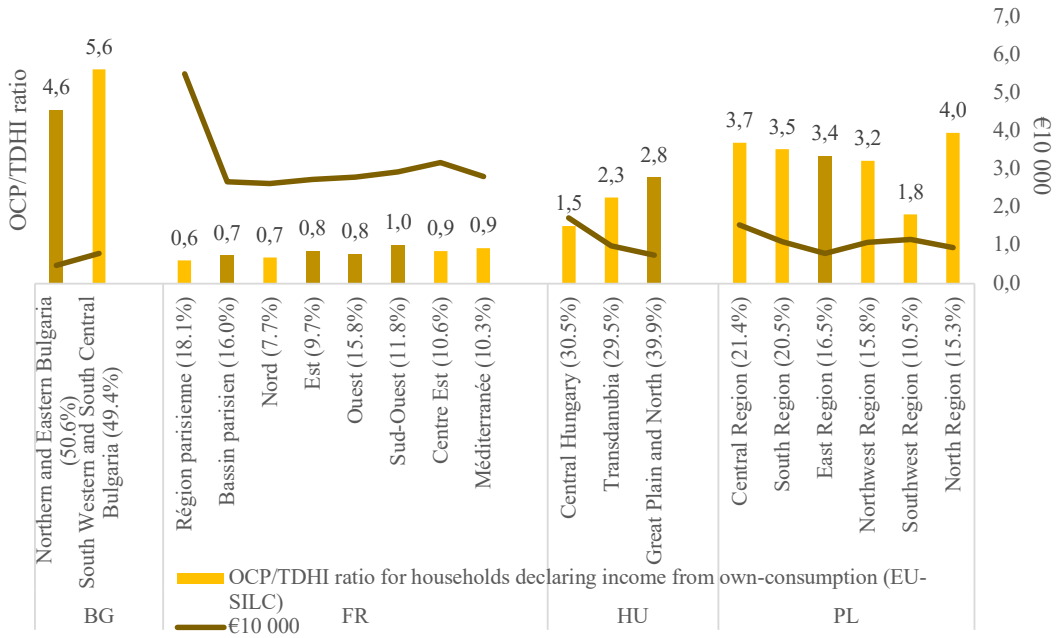
Out of all households involved in production for own consumption in Bulgaria, the majority live in Northern and Eastern Bulgaria (62 %), but the OCP/TDHI ratio is higher in South-Western and South-Central Bulgaria (5.6). Regions in the other three countries in Table 22.2 have even lower OCP/TDHI ratios. Figure 22.7 indicates that the OCP/TDHI ratio is negatively correlated with regional gross domestic product (GDP) in Hungary, meaning that the ratio is higher in less developed regions. However, this tendency is not so clear in other countries.

Table 22.2: Percentage of households with income from own consumption in different regions, 2015

Country and region		Household has OCP income (%)	
		No	Yes
Bulgaria	North and South-East Bulgaria	71.4	**28.6**
	South-West and South-Central Bulgaria	81.8	18.2
France	Paris region	93.0	7.0
	Parisian basin	68.3	**31.7**
	North	83.0	17.0
	East	69.4	**30.6**
	West	66.8	**33.2**
	South-west	69.2	**30.8**
	Centre east	72.9	27.1
	Mediterranean	87.6	12.4
Hungary	Central Hungary	87.0	13.0
	Transdanubia	65.5	34.5
	Great Plain and North	57.7	**42.3**
Poland	Central	82.7	17.3
	Southern	90.1	9.9
	Eastern	67.4	**32.6**
	North-Western	81.8	18.2
	South-Western	84.4	15.6
	Northern	83.6	16.4

Source: Author's computation based on data from the EU-SILC user database, 2015. Bold figures in 'Yes' column indicate regions with a significantly higher than average proportion of households with OCP income.

Figure 22.7: OCP/TDHI ratio among households reporting OCP income and per-capita GDP, 2015



NB: Darker colours represent regions with the highest percentages of households reporting OCP income. The percentages in parentheses represent the proportions of households in the different regions in each country. The brown lines represent GDP.

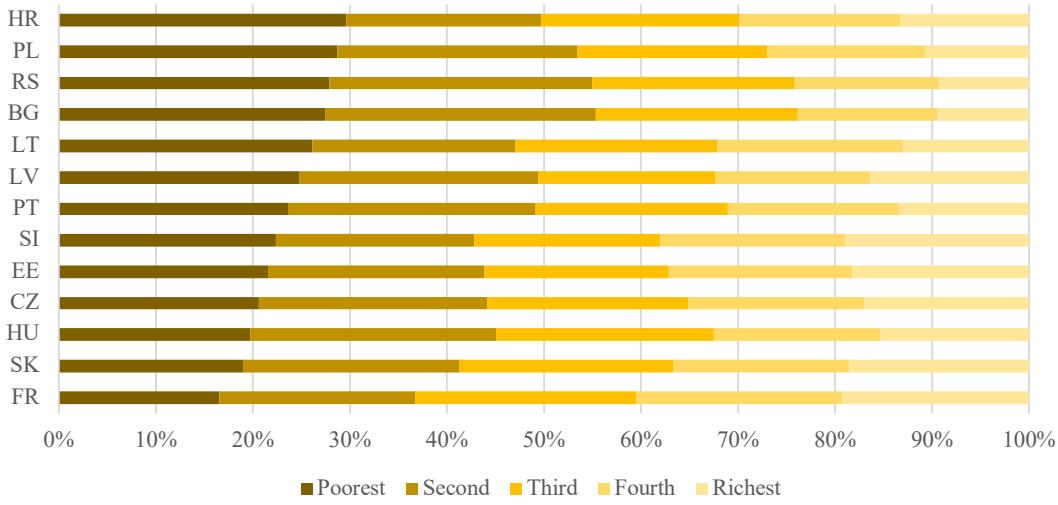
Source: Eurostat database [nama_10r_2gdp] (retrieved March 2018); author's computation based on the EU-SILC user database, 2015.

22.2.3. Production for own consumption by income quintile

Income quintiles are calculated using equivalised total household income. As calculated, equivalised total household income does not include OCP income and, therefore, when this income is calculated including OCP, the quintile distribution may change. However, here we want to identify if there is a difference in the distribution of households involved in production for own consumption by income quintile following the research hypothesis that households in lower quintiles are more frequently involved in production for own consumption. If there is no association between income and OCP activity it would be expected that one fifth (20 %) of all households involved in OCP would be in each quintile.

Figure 22.8 shows that, in the majority of countries included, over 20 % of households involved in production for own consumption are in the lowest quintile. This percentage reaches 29.6 % in Croatia. Only in France is the percentage of OCP-active households in the lowest income quintile substantially less than 20 %.

Figure 22.8: Distribution by income quintile of households reporting OCP income, 2015

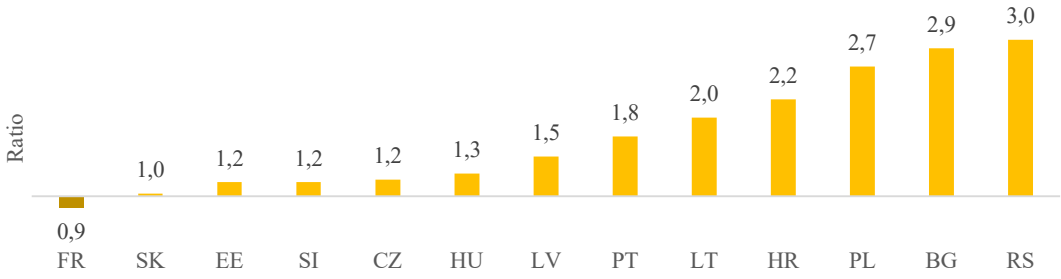


NB: Countries are presented in descending order of percentage of OCP-active households in the lowest quintile. In nine of the countries, households reporting OCP income are disproportionately likely to come from the lowest quintile.

Source: Author's computation based on the EU-SILC user database, 2015.

In addition, the difference between the proportion of OCP households in the lowest and highest quintiles can be considered an indication of whether households decide to become involved in production for own consumption for economic reasons or for some other reason (e.g. as a hobby or to grow healthy food). In Bulgaria, Croatia, Poland and Serbia this ratio is higher than two, meaning that the prevalence of household involvement in production for own consumption in the lowest quintile is more than twice that in the highest quintile (Figure 22.9).

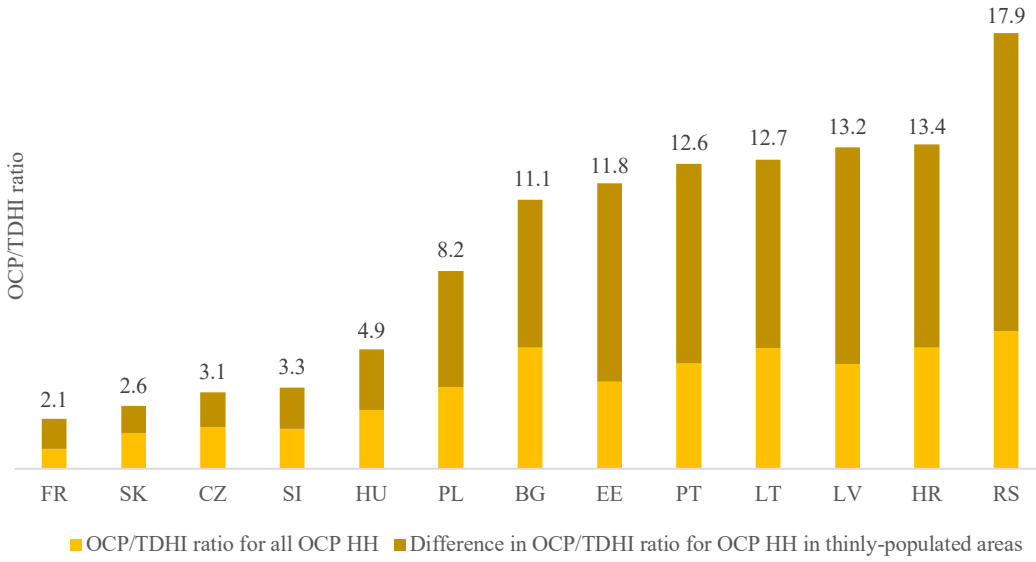
Figure 22.9: Ratio between the number of households in the lowest quintile reporting OCP income and the number of households in the highest quintile reporting OCP income, 2015



NB: Countries are presented in ascending order of ratio.

Source: Author's computation based on the EU-SILC user database, 2015.

Figure 22.10: OCP/TDHI ratio among households in the lowest quintile reporting OCP income, 2015



NB: Countries are presented in ascending order of ratio.

Source: Author's computation based on the EU-SILC user database, 2015.

For households in the lowest income quintile that are involved in production for own consumption, the OCP/TDHI ratio is substantially higher than that for all households involved in production for own consumption, indicating that OCP income tends to improve the well-being of households in the lowest quintile involved in OCP. This is especially pronounced in Croatia, Estonia, Latvia, Lithuania, Portugal and Serbia (Figure 22.10).

22.3. Influence of own consumption on EU social indicators

The influence of OCP income is estimated for the key indicators of income distribution (AROP, the Gini coefficient and the S80/S20 ratio ([126])) for countries in which OCP income represents a significant part of total income. The analysis considers both the total population and relevant subpopulations.

All key indicators analysed are actually income based and are calculated using total disposable income. In order to assess the influence of the inclusion of OCP income on these indicators, two income variables were calculated:

1. TDHI without added income from OCPs (i.e. the standard Eurostat definition, HY020),
2. TDHI with added income from OCPs.

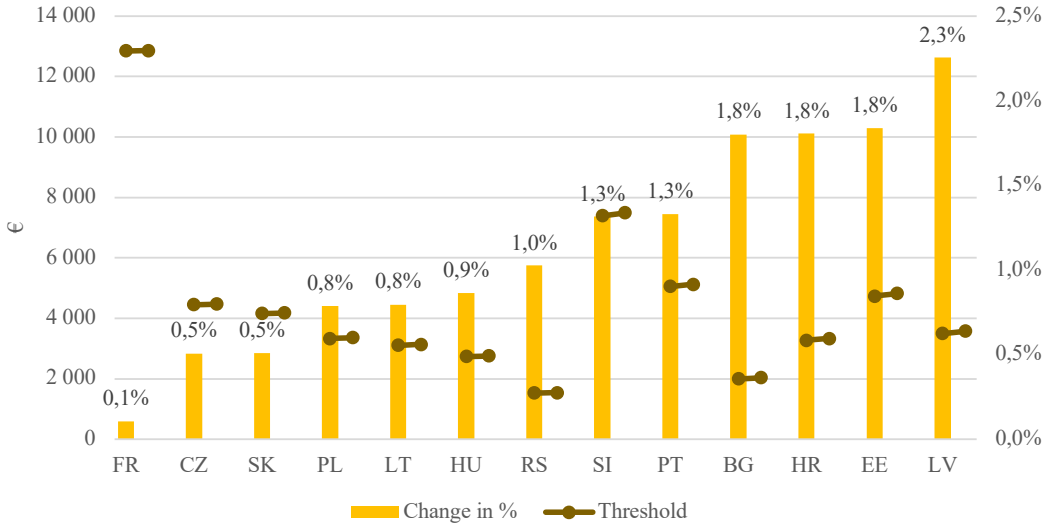
In order to eliminate the effect of household size on total household income, income was divided by the equivalised household size – the sum of the weighted number of household members – using a standard (equivalence) scale, the modified Organisation for Economic Co-operation and Development scale ([127]).

The AROP rate was calculated in the following way.

([126]) For a detailed presentation of the EU portfolio of social indicators, see Social Protection Committee (2015).

([127]) This scale gives a weight to all members of the household: 1.0 to the first adult; 0.5 to the second and each subsequent person aged 14 and over; and 0.3 to each child aged under 14. These weights are summed to provide the equivalised household size.

Figure 22.11: AROP threshold before and after inclusion of income from OCPs, 2015



NB: The first point for each country represents the AROP threshold before the inclusion of income from OCPs and the second point represents the AROP threshold after the inclusion of income from OCPs. Countries are presented in ascending order of percentage change in the threshold. The figure shows that adding income from OCPs increases the AROP threshold.

Source: Author's computation based on the EU-SILC user database, 2015.

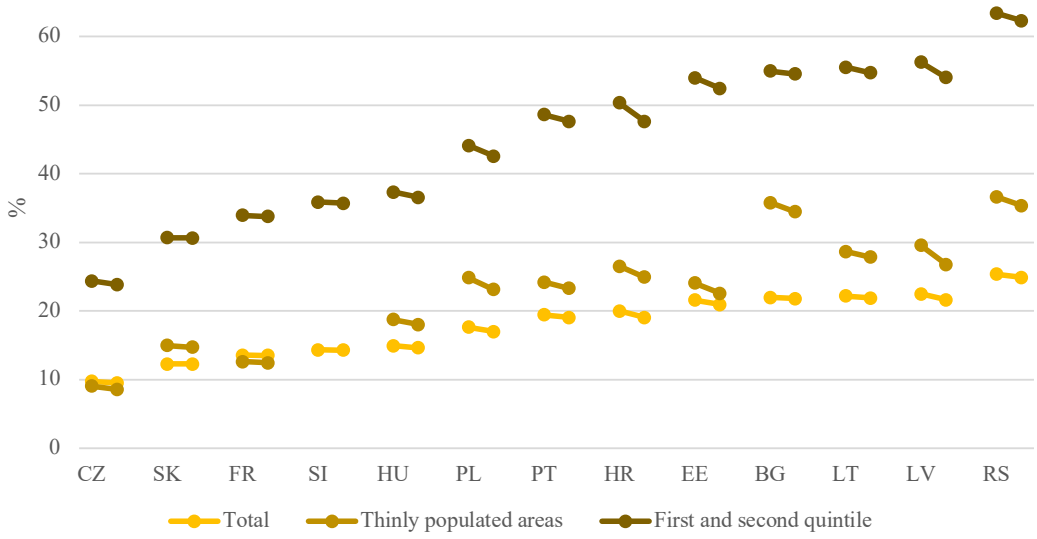
1. For each country, the AROP threshold was calculated as 60 % of the median value of the equivalised disposable income.
2. For each country, the AROP rate was calculated as the percentage of individuals having an equivalised disposable income that is less than the AROP threshold for that country.

The procedure was conducted twice: first for income without OCPs and then for income with OCPs. It is obvious that the AROP threshold should change after adding income from OCPs. When OCPs are included in the total household income, the AROP threshold slightly increases for all countries – by 0.1 % in France and by 2.3 % in Latvia. This change, in absolute terms and as a percentage, is shown for each country in Figure 22.11.

Although the AROP threshold increases, the AROP rate decreases for all countries. This means that income is more equally distributed. However, this change in rate is not significant in all countries. In all countries the reduction in the AROP rate is less than 1 percentage point. The reduction is highest in Croatia and Latvia (0.9 percentage points).

Even when only thinly populated areas are considered, the reduction in the AROP rate is modest. The greatest decreases in AROP rate in this subgroup are seen in Latvia – from 29.6 % to 26.8 % (2.8 percentage points) – and in Poland – from 24.8 % to 23.2 % (1.6 percentage points) (Figure 22.12). For lower income households (first and second quintiles), the change is greatest in Croatia – from 54.0 % to 52.4 % (2.7 percentage points) – and in Latvia – from 56.3 % to 54.1 % (2.2 percentage points).

Figure 22.12: AROP rates before and after inclusion of OCPs in total household income, for different subgroups of the population, 2015



NB: The first point for each country represents the value of the indicator when income from OCPs is not included and the second point represents the value of the indicator when income from OCPs is included. Countries are presented in ascending order by the value of the indicator in the total population when income from OCPs is not included.

Reading note: Adding the income from OCP decreases the AROP rate more substantially for people in thinly-populated areas and lower income quintiles than in the population as a whole.

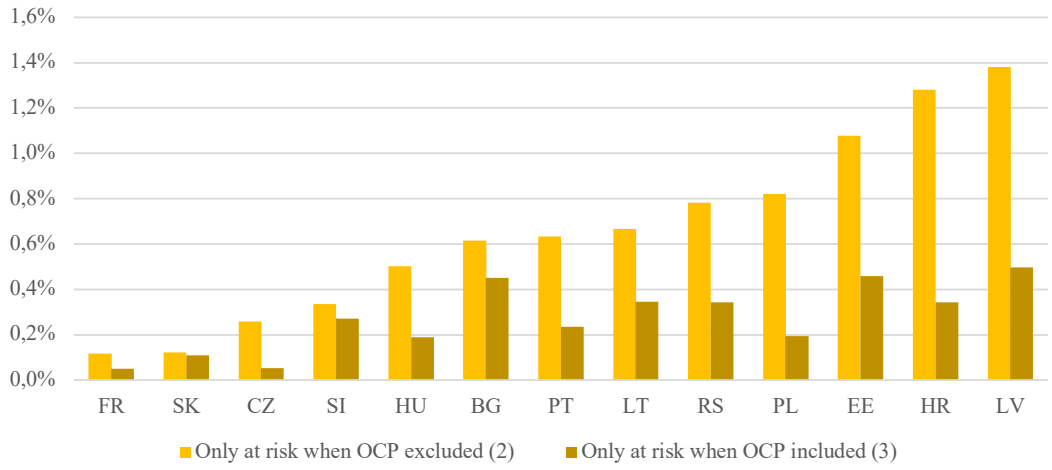
Source: Author's computation based on the EU-SILC user database, 2015.

The idea of including OCPs in total household income is to assess how this income helps people to avoid the risk of poverty. Therefore, the percentage of the population at risk of poverty before inclusion but not after inclusion of income from OCPs is estimated. There are four possible scenarios that can occur after including income from OCPs.

1. Individuals are at risk both with and without the inclusion of income from OCPs.
2. Individuals are at risk when income from OCPs is not included but not when this income is included.
3. Individuals are not at risk when income from OCPs is not included but are at risk when this income is included.
4. Individuals are not at risk in either case.

The second and third scenarios are the most interesting, as they indicate how the distributions of people at risk of poverty change when income from OCPs is included or excluded. The difference between these two gross changes, shown in Figure 22.13, reflects the net change in the AROP rate.

Figure 22.13: Percentage of individuals moving above or below the AROP threshold after inclusion of income from OCPs (scenarios 2 and 3), 2015

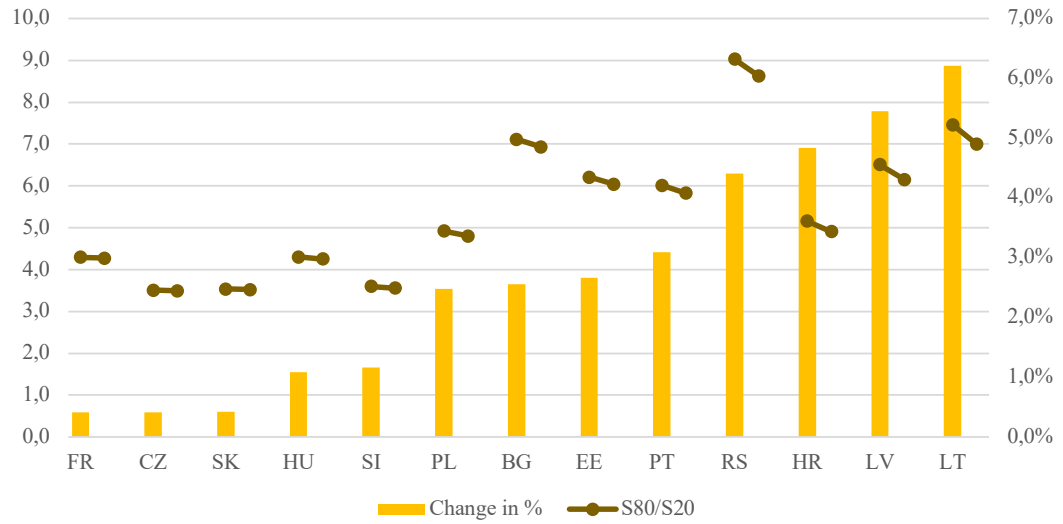


NB: Countries are presented in ascending order of proportion of individuals at risk only when income from OCPs is excluded.
Source: Author's computation based on data in the EU-SILC user database, 2015.

In all the countries included, the percentage of people who are 'taken out' of the risk of poverty by including income from OCPs is very modest, reaching a maximum of 1.4 % in Latvia. On the other hand, while OCP income pulls some people above the threshold, there are people who fall below the threshold (which increases after the inclusion of OCP income). Both categories include people who are close to the threshold and who are therefore still vulnerable.

The analysis shows that OCP income can improve the position of vulnerable people – a potential argument for including OCP income in total disposable income. Analysis of the inequality indicators (S80/S20 ratio and Gini coefficient) provides information on whether the overall income distribution is becoming more equal; greater equality would indicate that OCP income has a bigger impact for people in the lower part of the income distribution. The S80/S20 ratio decreases after the inclusion of OCP income for all countries, confirming the reduction in inequality. This reduction is highest in Lithuania – from 7.46 to 7.00 (6.2 % reduction) – and in Latvia – from 6.51 to 6.15 (5.5 % reduction) (Figure 22.14).

Figure 22.14: S80/S20 ratio before and after inclusion of OCP income, 2015



NB: The first point for each country represents the value of the indicator when income from OCPs is not included and the second point represents the value of the indicator when income from OCPs is included. Countries are presented in ascending order of percentage change in the value of the indicator. Including income from OCPs decreases the S80/S20 ratio for all countries.

Source: Author's computation based on data in the EU-SILC user database, 2015.

Although the Gini coefficient for equivalised TDHI also decreases after the inclusion of OCP income, this decrease is very modest, ranging from 0.2 % in France to 2.3 % in Latvia and 2.4 % in Croatia (Figure 22.15).

The analysis shows that there are some changes in the apparent well-being of households when OCP income is included. However, this effect is very modest, even for the subpopulations considered the most vulnerable. Although in countries that joined the EU after 2004 the percentage of households involved in production for own consumption is above 15 %, the share of income from OCPs in those households reporting production for own consumption reaches a maximum of 5.7 % at the national level in Serbia, meaning that more than 94 % of total household income is from other sources. The share of income from OCPs is higher when specific subpopulations are considered, such as households in thinly populated areas or households in the lowest quintile. However, the share does not exceed 20 %, reaching a maximum in Serbia among households in the lowest quintile reporting OCP income.

The value of the EU social indicators does not change substantially after the inclusion of OCP income, neither is equality improved substantially at national level, even in the most vulnerable subpopulations.

Figure 22.15: Gini coefficient before and after inclusion of OCP income, 2015

NB: The first point for each country represents the value of the indicator when income from OCPs is not included and the second point represents the value of the indicator when income from OCPs is included. Countries are presented in ascending order of percentage change in the value of the indicator. Including income from OCPs decreases the Gini coefficient for all countries.

Source: Author's computation based on data in the EU-SILC user database, 2015.

22.4. Conclusions and recommendations

Though there are some changes in the apparent well-being of households when OCP income is included, this effect is very modest, even for the subpopulations that were considered the most vulnerable.

The analysis shows that the benefit of including OCP income in total disposable income is marginal not only at the level of the total population, but also at the level of vulnerable subpopulations. However, although marginal, changes in the AROP rate for vulnerable subpopulations cannot be ignored. Inclusion of OCP income decreases the AROP rate in thinly populated areas in Latvia from 29.6 % to 26.8 % and in Poland from 24.8 % to 23.2 %. This means that a simple methodological change (i.e. 'correction') lifts 2 % of the population in rural areas out of poverty. This impact could be considered equivalent to a rather effective policy.

However, as described in Chapter 21, the collection of data on OCPs is very complicated and very sensitive to the methodology used. Considering, on the one hand, the obstacles to data collection and existing concerns about overburdening respondents with the volume of questions in EU SILC and, on the other hand, the modest impact of OCP income on the inequality indicators, it is suggested that OCP income should not be included in the calculation of TDHI.

References

Dijkstra, L. and Poelman, H. (2014), 'A harmonised definition of cities and rural areas: the new degree of urbanisation', *Regional Working Papers*, WP 01/2014, European Commission Directorate-General for Regional and Urban Policy, Brussels.

Eurostat (2016), *Methodological Guidelines and Description of EU-SILC Target Variables – DocSILC065 (2015 operation)*, Eurostat, Luxembourg.

Paats, M. and Tiit, E. M. (2010), 'Income from own-consumption', in Atkinson, A. B. and Marlier,

E. (eds), *Income and Living Conditions in Europe*, Eurostat, Luxembourg, pp. 179–194.

Sauli, H. and Törmälehto, V. M. (2017), 'The distributional impact of imputed rent in EU-SILC 2007–2012', in Atkinson, A. B., Guio, A. C. and Marlier, E. (eds), *Monitoring Social Inclusion in Europe*, Eurostat, Luxembourg, pp. 141–157.

Social Protection Committee (2015), *Portfolio of EU social indicators for the monitoring of progress towards the EU objectives for social protection and social inclusion – 2015 update*, European Commission, Brussels.

Survey modes, data collection and survey processing

23 Mode issues in comparative surveys

Lars Lyberg, Peter Lynn and Barry Schouten ([128])

23.1. Introduction

The mode of data collection refers to the medium that is used in a survey for obtaining sample members' responses to the survey questions. Modes include personal interviews, telephone interviews, various forms of self-administered modes, and variants and combinations of these. The data collection mode does not have to be the same as the mode of initial contact, but they are often naturally aligned. It has long been recognised that the choice of data collection mode can affect survey data quality along several dimensions, including accuracy, timeliness and costs. Different modes have their specific error structures and other characteristics, which make them more or less suitable for collecting data on a given subject matter in a given context. Early studies including those by Hochstim (1967) and Woltman, Turner and Bushery (1980) compared different strategies for data collection by contrasting single modes and mixes of modes in order to determine whether different modes might result in reasonably similar outcomes.

Timeliness and costs can be seen as constraints in the choice of mode, whereas accuracy components such as coverage, non-response and measurement error have effects that are more difficult to assess. The literature on what is referred to as 'mode effect' has expanded steadily during recent decades. A mode effect is the systematic effect of mode on a particular estimate for a specific target population: mode effects can (and do) therefore differ between estimates from the same survey. Attempts to assess the marginal effect of mode can be very complicated due to the varying design features that may be used in different mode situations (Groves, 1989). Therefore, it can be helpful instead to compare the outcome of entire data collection systems. The effect of mode on bias cannot be assessed directly, but information about the direction of the bias can help determine which mode is the better one (Lyberg and Kasprzyk, 1991). However, in order to understand the causes of a difference between modes, researchers need to disentangle the components that can contribute to the difference, prime among which are selection effects (different kinds of people may tend to respond in each mode) and measurement effects (some people may tend to give different answers in each mode). Recently, new methods for disentangling and even adjusting for mode effects have been developed (e.g. Vannieuwenhuyze, Loosveldt and Mohlenberghs, 2010; Schouten et al., 2013; Kolenikov and Kennedy, 2014; Klausch, Schouten and Hox, 2017), although these rely on unverifiable assumptions about the nature of pure mode effects. Confounding is a big problem, and careful design decisions aimed at preventing or at least mitigating mode effects seem to be a practical way to handle this error source.

During recent decades, the number of modes has increased, although most of them are developments of basic modes already used and mentioned above. Most notably, we have witnessed an increase in web surveys and alternative big data sources

([128]) Lars Lyberg was with Demoskop, Inc., Sweden; Peter Lynn is with the University of Essex, Colchester, United Kingdom; and Barry Schouten is with Statistics Netherlands and Utrecht University. All errors are the authors' responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Peter Lynn (plynn@essex.ac.uk) and Barry Schouten (jg.schouten@cbs.nl).

such as social media data and administrative registers (Couper, 2017; Tourangeau, 2017). Although it is in theory possible to choose the ideal data collection mode for a given study goal, in practice this is not achievable. Costs and other constraints call for some compromise or a mix of modes to accomplish good results regarding respondent recruitment, coverage, non-response and measurement error. For instance, it has become increasingly difficult for face-to-face interviewers to recruit survey respondents. Mixed-mode strategies have often been part of survey design (de Leeuw, 2005), but now they have become the norm, mostly due to costs and response rate concerns.

In comparative and international surveys such as European Union Statistics on Income and Living Conditions (EU-SILC), the issues related to survey mode are magnified. The nature of mode-specific error structures and constraints can differ between countries, as can cost differentials. This can lead to differences in preferred modes. Although Eurostat provides guidelines on which modes are allowed in EU-SILC, countries are still free to use combinations of modes if they see fit. In contrast, the European Social Survey (ESS) has for a long time managed to maintain face-to-face interviewing as the sole mode across all countries. However, as costs and non-response have increased, and survey field capacity has reduced in some countries, the pressure from participating countries has triggered experimentation with mixed-mode strategies and the introduction of the Cross-national Online Survey (CRONOS) probability-based web panels (see Chapter 27).

Thus, when it comes to comparative surveys we have to accept that the typical scenario is that we have a mix of modes within some countries and varying combinations between countries. Some countries may use a single mode, whereas others may use a mix of two or three modes. This kind of variation and its consequences for comparability are hard to grasp but will be discussed below. The actual effect of mode on EU-SILC between countries is unknown. It seems that a stricter protocol deserves a place in the guidelines. Chapter 24 describes the variety of modes used within and between countries for EU-SILC.

The impact of modes of data collection on comparative surveys becomes even more relevant when considering statistics over time, namely comparability over time. This holds for EU-SILC, with it being a survey that is repeated over many years. Mode coverage, mode preferences and mode familiarity are not time stable, so in time statistics may become incomparable.

In Section 23.2, we discuss modes and common mode characteristics such as error structures. In Section 23.3, various mixed-mode strategies are presented. Section 23.4 discusses mode issues in comparative surveys. Section 23.5 examines how to prevent, assess and adjust for mode effects. Section 23.6 puts mode effects into a total survey error (TSE) framework, and Section 23.7 provides some recommendations on how to handle mode effects in EU-SILC. Section 23.8 concludes the chapter.

23.2. Modes and mode characteristics

Modes can be classified in terms of three dimensions, namely the degree of contact with the respondent or sample member, the degree of data collector / interviewer involvement and the degree of computer assistance. Various modes using this classification scheme are presented in Table 23.1. These dimensions can be developed into features that differ between modes and are believed to be able to affect measurement, such as whether response options are presented aurally or visually (Lynn et al., 2012) and whether the interview is conducted at a faster or slower pace (Holbrook, Green and Krosnick, 2003). In principle, there are just two basic modes: those that involve interviewers or data collectors and those that are self-administered. All other modes are variants of these. Some of the variants have been short-lived, whereas others have prevailed. Some are complicated, and a simple statement of the mode used is seldom an adequate description of a survey research protocol. Thus, one has to be very specific and detailed when describing and documenting a mode. Modes are intertwined with the way samples are drawn, how respondents are recruited and how different contact devices function together. For instance, a web survey in one country may entail the respondent answering all questions,

Table 23.1: Data collection modes as a function of respondent contact, data collector / interviewer involvement and computer assistance

Type of contact	High data collector involvement		Low data collector involvement	
	Paper	Computer	Paper	Computer
Direct contact with the respondent	Face-to-face (PAPI)	CAPI, video web	Diary	CASI, ACASI, text CASI, video CASI
Indirect contact with the respondent	Telephone (PAPI)	CATI	Mail, SAQ	Web, video web, CSAQ, IVR
No contact with the respondent	Direct observation	CADE	Administrative records, big data	Webscraping, gadgets for passive data collection

NB: ACASI, audio computer-assisted self-interviewing; CADE, computer-assisted data entry; CAPI, computer-assisted personal interviewing; CASI, computer-assisted self-interviewing; CATI, computer-assisted telephone interviewing; CSAQ, computerised self-administered questionnaire; IVR, interactive voice response; PAPI, paper and pencil interviewing; SAQ, self-administered questionnaire.

whereas in another country income data may be taken from registers. It could be argued that there is also a third principal mode that includes data collection scenarios in which respondents play a passive role or do not even exist. Examples include direct observation of behaviours, housing environments and littering, and alternative data sources such as big data, scanners, sensors and administrative registers (Hill et al., 2019). However, in many of these scenarios a data collector must eventually be involved; thus, we return to the interviewer / data collector mode.

All modes are associated with advantages and disadvantages. These are described in detail in many places (e.g. Biemer and Lyberg, 2003; de Leeuw, Dillman and Hox, 2008). Here, we emphasise a few key characteristics of the most common modes.

Mail surveys

- There is no interviewer support.
- All respondent support must be visible in the questionnaire and other written materials.
- They are suitable for sensitive topics.
- They are relatively inexpensive.
- There is no control of the response process.
- The questionnaires should not be too long or complicated.
- There is a medium level of non-response.
- The timing of the interview is determined by the respondent.
- The pace of the interview is determined by the respondent.
- They are classified as a visual mode, but there is limited flexibility.

Web surveys

- There is uneven access within and between countries.
- They are suitable for sensitive topics.
- Data collection is relatively inexpensive.
- Data collection period can be much shorter than with other modes.
- The questionnaires should not be too long or complicated.
- There is a high level of non-response.
- The timing of the interview is determined by the respondent.
- The pace of the interview is determined by the respondent.
- They are classified as a visual mode.

Face-to-face interviewing

- This is the only mode that suits all populations.
- It is expensive.
- It benefits from interviewer support.
- Interviewer errors and inflated variances can occur.
- It is not suitable for sensitive questions.
- There is a relatively low level of non-response.
- The timing of the interview is constrained by the interviewer.
- The relatively quick pace of the interview is moderated by the interviewer.
- It is classified as an audio mode with an option for visual display.

Telephone interviewing

- It shares many of the face-to-face interviewing characteristics.

- Questionnaires cannot be too long.
- There is interviewer support, but it is limited.
- There is a medium to high level of non-response.
- Interviewer errors can occur.
- It is not particularly suitable for sensitive questions.
- Interviewing in central locations is good for monitoring and feedback.
- The timing of the interview is constrained by the interviewer but is relatively flexible.
- The pace of the interview is quick and is moderated by the interviewer.
- It is classified as an audio mode.

The mode dimensions form the starting point for attempts to harmonise features of modes. Examples are showcards to add a visual dimension to interviewer-assisted modes, chatbots and interactive voice response (IVR) to provide an interviewer feel to self-administered modes, and prompts to slow down the pace of the interview in the web mode.

The modes displayed in Table 23.1 do not constitute an exhaustive list of modes that can be used in national and comparative social surveys, nor are all listed modes always an option. For instance, a time use survey or a household expenditure survey needs a diary even if it has to be complemented with interviews. In addition, face-to-face surveys may be the only viable option among populations with limited literacy.

It has become increasingly difficult to conduct a large-scale survey using a single mode, and this is especially true for comparative surveys. To our knowledge, the ESS is the only comparative survey that has managed to maintain a single-mode design since its inception, although this may soon change due to increasing costs, non-response and a pandemic that largely prevents face-to-face interviewing. As a result, mixed modes and computerised modes have become the norm in many national surveys and almost all comparative surveys. However, this scenario is not new. For decades, one mode has routinely been followed up by some other mode to increase response rates. What is new is the fact that many additional factors such as costs, coverage, measurement issues and privacy concerns, together with technology development and new data sources, have made the mode choice more complicated, especially since data users want richer, more timely and cheaper insights (Holt, 2007). Therefore, developments in data collection possibilities continue to be of great interest to the survey industry (Couper, 2011, 2017). The decline in response rates is almost universal and has affected telephone surveys more than any other mode. People can easily screen telephone numbers they do not recognise. The development of mobile phones has been extremely rapid, and landlines are increasingly being abandoned. At the same time, telephone interviewing (computer-assisted telephone interviewing (CATI)) is the natural alternative to the more expensive face-to-face interviewing, which is becoming more and more difficult due to problems with recruiting and retaining interviewers. Telephone interviewing is still widely used in longitudinal studies.

Web surveys of the general population continue to be a challenge. They have many attractive features and probably a great future as a function of growing internet use. They are inexpensive (at least relatively) and fast, and allow complex questionnaires to be used while keeping social desirability biases at bay. They are also a great vehicle for experimentation. For instance, Haan et al. (2017) studied response behaviour using a form of video web in which human-like interviewers replaced human interviewers and asked respondents recorded questions. Video web was compared with traditional web surveys, computer-assisted personal interviewing (CAPI) and CATI regarding disclosure of sensitive information and satisficing behaviour including 'don't know' responding, and primacy and recency behaviour. This specific video web application contained aural and visual stimuli and was non-responsive; in other words, there was no interaction between the human-like interviewer and the respondent. If the level of engagement were to increase, the human-like features would have to include responsiveness. That in turn would probably introduce new error structures. In this experiment, respondents displayed behaviours similar to those displayed in any regular mode.

The choice of mode or the mix of modes is an important part of the survey design process. As men-

tioned above, the various dimensions of mode must be taken into account, and so must information about study specifics and what to expect. For instance, can we expect the presence of others during a personal interview and does that vary between countries? Should we use the same channel of communication across countries? Is the locus of control the same across the board? Do the perceptions of privacy and what may be considered sensitive questions vary across countries or other subgroups? These design considerations illustrate that mode cannot be defined in a simple sentence: careful documentation of all the relevant dimensions of mode is necessary.

23.3. Mixed-mode strategies

Mixed-mode surveys have been around for decades. The US Census of Population and Housing has used mixed-mode strategies since 1970, combining mail with follow-up by personal enumeration if necessary. Many other surveys use a mix of self-administered modes and interviews to boost response rates. However, it was not until 2005 that mixed-mode strategies were promoted more systematically by providing typologies and frameworks that could help decisions regarding mode choice and disentangle various mode effects (de Leeuw, 2005; Voogt and Saris, 2005). Since then, the literature on various aspects of mixed-mode strategies, including review articles, has grown considerably (e.g. Tourangeau, 2017; de Leeuw, 2018), as has the number of mixed-mode surveys. Since EU-SILC was launched in 2003, many countries have used mode combinations, which vary as described in Chapter 24. And since 2012, all social surveys conducted by Statistics Netherlands have used a mix of web, telephone and face-to-face modes in that order.

Mixed-mode strategies are not easily defined. A mix can be a combination of modes for collecting data, but it can also mean the different ways of contacting and soliciting respondents together with the various methods used to collect data from them. For instance, respondents may be recruited over the telephone and then asked to go online and answer questions using the web. In this case, two modes are used during the communication, but only one of them is used to ask survey questions.

Dillman, Smyth and Christian (2009) list the most common reasons for considering a mixed-mode approach:

- establishing contact and delivering materials;
- enhancing trust;
- offering sample members a choice of selected modes;
- reducing the coverage error when there is no single mode that covers the entire target population;
- improving response rate and reducing non-response bias;
- reducing measurement error;
- reducing overall data collection costs.

However, offering a choice of mode can be counterproductive. In a study by Medway and Fulton (2012) on the effect of concurrent web options on mail survey response rates the response rates went down. The authors speculate that, in such a scenario, the respondent must make two decisions: one concerns participation and the other concerns the choice of mode. For some sample members, this is too much of a cognitive burden and they prefer being directed to a specific mode.

de Leeuw (2005), Dillman, Smyth and Christian (2009) and Groves et al. (2009) provide typologies of mixed-mode surveys. During the contact phase, various activities including advance notification such as a mail invitation for a web survey, telephone recruitment for an IVR survey and screening can be carried out using one mode while data are collected using another. During the response phase several mixed-mode scenarios exist.

- **One sample, one time period and one questionnaire.** Here, one mode is used for some respondents and one or more modes are used for others in the sample. This is the classic scenario in which modes can be administered sequentially, typically starting with a relatively inexpensive mode and finishing with a more expensive one, with the purpose of increasing

response rates while minimising costs. It is also possible to administer the modes concurrently, which means that more than one mode is offered at the same time. If the intention with this approach is to let respondents choose in order to boost response rates, this strategy can sometimes backfire, as mentioned above. The Medway and Fulton (2012) finding has since been confirmed by other studies (de Leeuw, 2018). An alternative form of concurrent mixed-mode design is where the mode offered differs between sample members, based either on (assumed) preference or on availability of contact information.

- **One sample and one time point, but different modes for different parts of the questionnaire.** This means that respondents answer some questions in one mode and other questions in another mode. The strategy is mostly used for questions or segments of questions that are at risk of social desirability bias. With this approach, there must be direct contact between the interviewer and the respondent, as in CAPI, so that the computer can be physically passed to the respondent, who then responds using computer-assisted self-interviewing or audio computer-assisted self-interviewing (ACASI).
- **One sample and multiple time points.** This is the typical scenario in longitudinal or panel studies, in which data on respondents are collected on multiple occasions. For example, in short-term rotating panels such as the Labour Force Survey, it is common to start with a face-to-face interview followed by subsequent telephone interviews. The mix does not necessarily have to be face-to-face and telephone interviewing, but this is a common combination. The reasons for using a mix include cost reductions, maintaining reasonable response rates and general efficiency. The multiple time points permit the collection of additional contact details (email addresses, phone numbers) that can be used to make contact at subsequent time points.
- **Different samples and different modes.** This is the multipopulation or comparative survey situation seen in international surveys. The administrative, financial and methodological situations vary greatly between countries, making it very difficult to prescribe a single mode to be used across all participating countries. The only mode that is universal enough to accomplish this is face-to-face interviewing, but the costs involved make it an impossible option in many countries. A vast majority of comparative surveys use mixed-mode designs both within and between countries.

Early mixed-mode designs aimed to improve coverage and response rates. A dual-mode design was used to gradually increase response rates by improving coverage. For example, after 1993, once face-to-face fieldwork had been exhausted, the British Household Panel Survey would attempt telephone interviews with remaining non-respondents at each wave. This would tend to increase the response rate by between 1 and 4 percentage points. Dual-mode designs of this kind could have problems, however. For example, in some surveys the mail questionnaires for non-respondents were passed to a group of telephone interviewers who were asked to use them in a telephone interview. In the 1960s and 1970s, the cognitive aspects of the response process were not fully recognised – unless a questionnaire is developed to fit two or more modes measurement problems will occur. The cognitive response process (Tourangeau, Rips and Rasinski, 2000) is different depending on mode and mix of modes. Examples of cognitive phenomena include social desirability bias in interview surveys, acquiescence (the tendency for respondents to agree with statements), interviewer errors including fabrication of data, the order of response alternatives, memory and recall, visual aids, and the presence of others during an interview. It goes without saying that mixes of modes with their individual error structures can affect survey data quality and even more so comparative survey quality.

We have seen that a sequential offering of modes can lead to improved response rates (de Leeuw, 2005). The effects of offering a mode choice are more debateable since results are mixed. Overall, however, offering a choice seems to generate an increased cognitive burden that actually tends to increase non-response rates. Perhaps the surprising

thing is that this effect of an increased burden sets in so early. We have witnessed the same phenomenon when it comes to primacy and recency effects, when respondents are presented with lists of response alternatives. Information presented at the beginning and the end tends to be better retained than information presented in the middle (Krosnick and Alwin, 1987). By and large, concurrent mixed-mode designs do not seem to have a substantial positive effect on response rates or coverage. Mixed-mode designs seem to generate better representativeness than single-mode designs, and a mixed-mode design can really help to address various problems associated with single-mode designs, such as inherent problems with coverage, non-response, costs and measurement, which can vary greatly between modes. However, designing a mixed-mode survey or moving an organisation to a mixed-mode platform does not come without problems. For instance, questionnaire design and overall quality control of the data collection process through paradata are more demanding, and measurement will change (Dillman and Edwards, 2016). Should a mixed-mode design aim to improve comparability by using a unimode questionnaire that preserves equivalence as much as possible or should each mode be utilised to its full potential to minimise the TSE? As comparability is the ultimate goal in comparative surveys such as EU-SILC, unimode instruments would seem to be desirable. However, EU-SILC and other EU surveys are squeezed between the goal of comparability and each country's need for a national estimate. A unimode approach may not be the best approach if a country wants the best possible national estimate. The results of the consultation indicate that many countries put more emphasis on getting good national estimates.

The mode effect is the net effect of non-observation and observation error differences between modes. Using a variety of modes tends to bring in different and more respondents, which is something that is desired, whereas measurement effects are undesired and should be adjusted. Measurement effects are serious since they threaten comparability. Effects can be mode inherent, context specific or implementation specific. Therefore, prior to any adjustment, certain preventive measures can be taken, including the provision of guidance on developing a reasonably harmonised core questionnaire and translating it into the various EU languages. Translation methods in EU-SILC are currently completely opaque. One should also avoid mode-specific question structures and reduce differences between aural and visual presentation of questions.

A sequential offering of modes is an efficient mixed-mode implementation strategy, since it can improve coverage and response rates and can be implemented in a way that minimises costs. However, an element of concurrent choice can also be introduced. In an invitation to perform the survey on paper or online, it can be announced that an interviewer will contact the person/household if the survey is not completed after a specified time period. Although concurrent offerings have some drawbacks, as already noted, we cannot escape concurrent strategies. First, web surveys come with a multitude of opportunities for data collection through various devices, including smartphones and tablets, with varying screen sizes and other characteristics. If the web survey allows the use of a variety of response devices and is offered as part of a mix of modes, this will generate comparability problems. A mixed-device scenario could be thought of as a subgroup of concurrent mixed-mode design (de Leeuw and Toepoel, 2018), which complicates design and implementation since questions and questionnaires must be shorter than in other modes, and new types of response options and formats are needed (no grids or matrix questions). Second, in comparative surveys such as EU-SILC, the central organisation must accept that different modes are used across mobile devices in a mixed-mode implementation. Third, the future of mobile devices in EU social surveys has been discussed within Eurostat's Mixed Mode Designs for Social Surveys (MIMOD) project, and, although no EU-SILC country is currently ready to start implementing such data collection devices explicitly, some countries believe they will implement them eventually. Fourth, in the era of big data and multiple data sources it is possible to enhance the outputs from social surveys by combining survey data with register data, interviewer observations and passive collection through wearables. The survey landscape as we know it is changing rapidly, and the choice of data collection modes and mixes of modes deserves attention.

23.4. Mode issues in a comparative setting

The primary goal of international social surveys such as EU-SILC, the Survey of Health, Ageing and Retirement in Europe (SHARE), the European Social Survey, the International Social Survey Programme (ISSP), the World Values Survey and the European Quality of Life Survey is to compare populations, often countries, regarding social phenomena. The concept of comparability is highly complex and relies on the assumption that it is possible to develop stimuli, for example in terms of survey questions, that result in equivalent measures that can be compared across the populations. To achieve comparability, a certain level of input harmonisation is needed. Countries need to implement at least some survey processes in ways that are considered best practice. The choice of mode or mix of modes is one of those decision points that can affect data quality, either in terms of TSE or in terms of comparability. Admittedly, as we have seen above, different modes generate different responses but, with a good mix, both good coverage and good response rates can be achieved. Measurement errors are, however, more crucial. For instance, data collection by interviewers can generate errors, especially social desirability bias, but also an inflated variance due to interviewers' non-standardised behaviours. With a few exceptions, the data sets of comparative surveys have been generated using different modes and different mixes of modes, and under different general survey conditions. There is no reason to believe that these data are automatically comparable. As Jowell (1998) pointed out, comparative surveys must strive for equivalence, apply the highest methodological standards and conduct the survey under basically the same essential survey conditions across countries. He suggests the use of input harmonisation, but in most comparative surveys this is not possible due to local conditions. Input harmonisation can, however, be applied in various ways without being too rigid. In most comparative surveys, it is not feasible to prescribe a certain mode mix for all countries, but it would be possible to recommend that countries investigate possible mode effects and their sizes, directions and significance. Minimum standards regarding interviewer training and monitoring are also needed; some cross-country variation in process steps is due to a lack of methodological resources and varying research traditions. The EU-SILC guidelines provide very little advice regarding mode choice and its consequences. They prescribe using paper and pencil interviewing, CAPI, CATI and web mode, but there is no mention of how to choose or implement a mixed-mode design. The guidelines also discourage the use of proxy interviewing; abstaining from using proxy respondents where possible is certainly good advice.

It must be stressed that comparability is not guaranteed over time when sticking to the same mode design. Telephone and web population coverage rates have changed quite drastically over time. The choice of online device is still a moving target. At Statistics Netherlands, the proportion of mobile device respondents has increased by approximately 3 percentage points per year since 2012.

It seems that running a large comparative survey is so demanding that typically the available resources and efforts are barely enough to get the survey under way. Furthermore, most countries put more effort into coverage and non-response issues than measurement issues, even though the latter can hurt comparability the most. The ESS is the only large comparative survey that has studied mode issues in detail (see Chapter 27 of this book). The ESS has used face-to-face interviewing as its single mode since its inception in 2002; however, from the outset it was clear that this strategy would not be sustainable in the long run, primarily due to increasing non-response in most countries, then rising costs and finally a lack of survey organisations that can handle large-scale face-to-face interviewing. The ESS mixed-mode study programme began in 2005 and has involved experiments regarding alternatives to face-to-face interviewing, by using the telephone or web or a mix of modes (Roberts, Jäckle and Lynn, 2006; Jäckle, Roberts and Lynn, 2010; Martin, 2011; Martin and Lynn, 2011; Villar and Fitzgerald, 2015, 2017). Mode research has also been conducted for the ISSP (de Leeuw, Suzer-Gurtekin and Hox, 2019; Suzer-Gurtekin et al., 2019).

The extensive long-term ESS mode experimental programme involved six studies investigating

measurement differences across modes, causes of measurement differences and the feasibility of alternatives to face-to-face interviewing including mixed-mode designs. Over the years, the web option has been increasingly studied not as a single mode, but as a complement to interviewing. Here are some highlights from the programme (Martin, 2011).

- No single mode can outperform the face-to-face mode regarding data quality, since all alternatives, single mode and mixed mode, generate higher non-response rates.
- Many individual items (27–72 %) were affected by mode.
- Mode effects vary depending on country and topic, which makes them very difficult to predict.
- Cost savings may be difficult to realise due to complexity, even though substantial savings seem possible in some countries. Some countries may actually face increased costs with a new mode scenario.
- The telephone mode has increasingly become a mode of the past but will probably become interesting again when countries start working on data collection using smartphones.
- Time series will be disrupted, but this is unavoidable whenever major design changes take place.
- Different mode designs lead to measurement differences that are inconsistent across countries. Systematic differences may be observed between countries within a round and between rounds within a country. Thus, mode design is an error source within the TSE framework. Loss of comparability due to use of different mode designs is unacceptable, and if mixed-mode designs are to be implemented in the ESS a critical number of countries will need to start implementing them at the same time for the effort to be cost-effective.
- It is difficult to disentangle selection effects from measurement effects. Numerous differences in item measurements exist, even after controlling for demographics. An ESS future in which some countries stick to the current face-to-face model and some use their own mix with a variation in the proportion of respondents participating in each mode over time is something that the ESS scientific leadership has a hard time accepting, since there is no cost-effective way to adjust for measurement differences stemming from mode.
- A probability-based web panel from which face-to-face respondents can be recruited may be a solution to the cost problem and some of the data quality problems. Work on developing the methodology for such a panel is under way.

Roberts, Jäckle and Lynn (2006) investigated mode sensitivity of the ESS face-to-face questionnaire by looking at the potential impact on data quality of a switch to telephone interviewing. Indicators included social desirability bias and satisficing, and there were statistically significant mode effects for approximately one third of the questionnaire items. Mode experiment results differed between countries due to study design and actual differences. Mixed-mode design confounds coverage, non-response and measurement errors, and then further confounding takes place because of possible cross-national differences. As the authors point out, non-sampling errors are not replicated across countries. Within-country mode variability can cause differential mode effects between countries. For instance, the effect of face-to-face interviewing can differ between countries due to variability in interviewer workload and in the nature and extent of monitoring and training and other quality assurance measures. The bottom line is that, unless there is a set of agreed-upon requirements or specifications, each country will use the methods it thinks are the most appropriate, leading to unintentional variation in survey errors.

Discussions of mixed-mode design in a comparative setting have also arisen in the context of the ISSP, which has a concurrent design (Suzer-Gurtekin et al., 2019) in which specific modes are used for specific subgroups. For example, those without internet access are provided with either internet access or a paper self-administered questionnaire (SAQ) if that is preferred. Those with a high response propensity are allocated to the cost-effective web or mail mode. In SHARE, the German data are combined with data from other sources (de Leeuw, 2018). There is now a rapid development regarding data sources that can complement survey data, not

just administrative data and smartphones data, but also data from gadgets, devices and platforms such as smart weighing scales, accelerometers, scanners, Google Maps, social media, webscraping and the Global Positioning System (Williams and Ghimere, 2019). Although the large comparative social surveys have and must have a mixed-mode design, they are not very well suited to it, since questionnaires tend to be long, which causes an increased respondent burden and makes telephone and SAQ modes less attractive. In addition, for pre-existing surveys, attempts at making questions comparable across modes can hurt equivalence over time. Mixed-mode designs introduce an extra layer of uncertainty to the comparability of survey data.

23.5. Preventing, assessing and adjusting mode effects

To some extent, mode effects can be prevented. For instance, sensitive questions should be measured using self-completion modes (web, paper SAQ or ACASI) to avoid social desirability bias. What is considered sensitive can, however, vary across countries, cultures and subgroups. Similarly, through careful design, question format differences between modes can be avoided, even though the result may be that the unimode format is not the optimum format in all modes. It is important to establish whether the goal of the survey is comparability across populations, whereby sizes of errors are equalised in a unimode fashion, or whether the goal is to produce population estimates with errors that are as small as possible using best practices. In practice, comparative surveys have both goals, and a set of input harmonisation requirements that do not interfere with countries' best practices is required.

Mode effects concern who responds and how they respond. They are the net effects of the differences between modes regarding the non-observation errors and the observation errors. The differences are not errors per se; rather, they reflect differences between mode-specific biases. Mode assessment studies become insightful when selection and measurement are separated. Separation could be achieved using a validation study, mode groups could be made comparable by weighting or by using regression methods, or errors could be estimated using modelling. Jäckle, Roberts and Lynn (2010) observed some challenges associated with mode effect assessments. First, the experimental design must be such that the only difference between samples is the mode that is assigned to the sample members. Second, if the sample composition differs between modes because of differential non-response, then tests for mode effects must control for respondent characteristics. Third, differences in responses across modes may affect some types of estimates but not others. The authors refer to testing for mode effects as a 'fishing' exercise. Conducting many tests within a single experiment can increase the risk that false positives are interpreted as significant mode effects. Mixing modes entails a trade-off between errors and costs, as in all survey design decisions. Assessment can be carried out by comparing data collection systems, but if the systems produce estimates that are significantly different we will not know which system produces results that are closest to the truth. Assumptions regarding the likely direction of the bias or a comparison with a gold standard independent survey can help establish which system is the better one. However, it is entirely possible that one system is better for selection (representativeness), whereas another is better for measurement. Comparisons of quality indicators such as degree of item non-response, length of responses to open-ended questions, reliability, extent of straight-lining and comparisons with external data may assist in identifying preferred systems. Designing informative experiments on mode effects is hard to accomplish: it is difficult to fully control all relevant factors, leading to the need to make strong assumptions. Still, the amount of literature on mode assessment is relatively large, although outcomes have not been extensively picked up by national statistical institutes (NSIs) in the EU. Survey designers have a number of options regarding the assessment of mode effects. The effects can be ignored, which we know is quite common. Notification of the direction and magnitude of the effect at a question level using descriptive methods comes next. Randomised mode assignments for a portion of the sample with subsequent comparisons of re-

sponse distributions are more ambitious, whereas statistical modelling to make responding samples equivalent across modes is more complicated. Adjustments are not as common in practical survey work or in the production of statistics. Complexity, lack of auxiliary data and the fact that adjustments lead to increased variances and costs and possibly increased logistics may be important deterrents.

Within Eurostat's ESSnet MIMOD project, Buelens, van den Brakel and Schouten (2018) provide an overview of methodologies that can be used in mode effect assessment and adjustment. Some of the points they make are summarised here.

- Mode assessment studies tend to quantify the total mode effects rather than separate selection and measurement effects.
- Separation of selection and measurement effects is only possible under strong assumptions or when specific data are available (Schouten et al., 2013; Vannieuwenhuyze and Loosveldt, 2013; Klausch, Schouten and Hox, 2017).
- Measurement effects arise when the same respondent gives different answers to the same question in different modes. Measurement effects are sometimes referred to as pure mode effects.
- Mode effects may vary across items, countries and designs.
- It is hard to find consistent results that can be turned into default design principles or best practices. Furthermore, studies are sometimes conducted despite the likelihood of mode effects.
- Advanced experiments are not common. These would entail using parallel independent surveys, embedded experiments and reinterview studies.
- Causes of mode effects can be found in some of the elements of the response process (absence or presence of an interviewer, presence of others, speed of the interview, literacy, computer literacy, perceived confidentiality, type of questions and questionnaire design).
- Confounding of selection and measurement is a key problem (Biemer, 1988; Jäckle, Roberts and Lynn, 2010).
- Assessments can be made using experimental designs (embedded experiments, split samples, repeated measurement designs) or non-experimental designs (observational studies, weighting, and regression-based methods to control for selection effects).
- The literature on adjustment methods is limited. Techniques include reweighting, calibration, imputation and prediction, which partly overlap with adjustment methods for coverage and non-response in one-population surveys.
- Until now, there has been no great interest in developing or applying mode adjustment methods within EU Member States, even though the web option has sparked considerable new interest.
- Often, mode effects are assessed relative to a benchmark mode, sometimes regarded as the gold standard, which would allow an estimate of bias, but sometimes it could be just the standard mode, which does not allow bias estimation.
- The MIMOD project conducted a consultation and found that about one third of EU Member States did not conduct any assessments of mode effects in their social surveys. Those that claim they do apply rather standard test procedures. About two thirds of countries have not made any attempts to adjust mode effects. Only half of the countries report plans for future mode or adjustment studies.

Methods that have been used to adjust for mode effects include regression with mode as binary predictor, multigroup structural equation modelling analysis, predictive mean matching, the potential outcomes approach and comparisons with reference or gold standard surveys. To assess and adjust for mode effects requires auxiliary data. Those data may already be available in a reference survey, in a longitudinal data set or in respondent demographics. Otherwise, new data must be collected by, for example, experiments or reinterviews. That complexity may be one reason why countries abstain from more sophisticated attempts and instead rely on simpler attempts or choose to ignore mode effects. The literature on mode effects is not always easily digestible, and various studies show diverging or inconclusive results depending on method choice and survey topic. Some studies consistently

show that selection and measurement effects are confounded (Vannieuwenhuyze, Loosveldt and Mohlenberghs, 2010). The order of modes within a sequence of modes can make a difference regarding bias reduction, but it is not possible to make any generalisations (Sakshaug, Cernat and Raghunathan, 2019). Some studies confirm accepted design principles regarding, for example, self-completion modes and reduced social desirability bias (Laaksonen and Herskanen, 2014). Kolenikov and Kennedy (2014) evaluated three approaches to adjust for mode effects. Their work illustrates how quickly efforts become complicated when trying to adjust for these effects. First, under a regression modelling approach, adjustments were computed by regressing survey responses on mode, demographics and other relevant variables. Second, under a multiple imputation approach, mode effects were conceptualised as a missing data problem, and standard multiple imputation techniques were used to impute responses in additional modes. Third, a new imputation approach based on an econometric framework of implied utilities in logistic regression modelling was proposed. The multiple imputation approach produced estimates with better apparent accuracy based on better internal consistency of the estimates and only moderate increases in the standard errors.

Schouten et al. (2013) and Klausch, Schouten and Hox (2017) describe reinterview design experiments used to disentangle mode effects. They assumed that the reinterview had a negligible impact on the mode-specific measurement error model. They adopted two strategies. One strategy was to calibrate response based on the reinterview variables. Another strategy was to treat the reinterview as a repeated measure and apply observed mode differences to participants who responded in only one mode. The underlying assumptions can, however, be quite strong and unrealistic.

23.6. Mode and total survey error

The decision about mode or mix of modes is an important part of the survey design process, since the choice constrains other design decisions such as access to frames, sampling strategy, survey topic, confidentiality, timing, non-response follow-up and costs. Thus, mode choice involves a cost–error trade-off.

The uncertainty of an estimate can, in theory, be measured by the mean squared error (MSE), which is the sum of the sampling error and all the non-sampling errors. This sum is the TSE. Non-sampling errors are due to mistakes or system deficiencies. A simple definition of MSE is the following (Biemer and Lyberg, 2003), where the subscripts denote the various sources of error and these sources constitute a TSE framework:

$$\begin{aligned}\text{MSE} = \text{variance} + \text{bias}^2 = \text{Var}_{\text{sampling}} \\ + \text{Var}_{\text{measurement}} + \text{Var}_{\text{data processing}} + (\text{B}_{\text{specification}} \\ + \text{B}_{\text{non-response}} + \text{B}_{\text{frame}} + \text{B}_{\text{measurement}} + \text{B}_{\text{data processing}})^2\end{aligned}$$

Over time, alternative frameworks for handling TSE decomposition have been suggested. Factors that have triggered this include the emergence of additional components such as model/estimation error and revision error, and new or not so new types of study. For instance, Smith (2011) has developed a framework for comparative surveys in which he introduces the concept of comparison error. Statistics Netherlands has started to develop error frameworks for sensor data (Beinhauer, Snijkers and Bakker, 2020), and Statistics Norway have started to develop them for integrated survey and register data (Zhang, 2011). These are frameworks for studies that use multiple data sources, which is reflected by the fact that they are examples of what may be referred to as total error frameworks rather than TSE frameworks.

We have seen that the mode or mode mix can generate various coverage, non-response and measurement errors. If mode effects are adjusted, the variance of estimates increases. Thus, the mode contributes to the TSE through several components, both variances and biases. If the mode is chosen and implemented properly, then each mode should be utilised to its full potential, even though comparisons of countries may suffer. The unimode approach can compensate for this, provided the unimode instrument is carefully reviewed, translated and adapted, and not overly

simplified in some countries in order to achieve equivalence.

If mode effects are ignored, comparisons of estimates for some variables, such as sexual habits, attitudes towards abortion, political sympathies and economic activity, may be way off target (Tourangeau and Smith, 1998; Groves et al., 2009; Krumpel, 2013) and result in information that may be quite misleading or even useless. However, one should bear in mind that, most of the time, mode effects are quite moderate and what is considered sensitive varies across countries and cultures and over time.

de Leeuw (2018) summarises the TSE perspective on achieving mode equivalence as comprising three steps. First, equivalent questionnaires are developed, tested, translated and, if necessary, adapted. Second, mode effects are estimated by separating intended mode measurement effects and unintended mode measurement effects. Third, if necessary, one should adjust for unintended differential mode measurement error. The second and third steps need auxiliary data, such as demographic data, frame data, record data, reinterview data or reference or gold standard survey data.

23.7. Recommendations for EU-SILC

- The current guidelines for designing and implementing EU-SILC do not address mode issues. There is only one page on permitted modes. This is insufficient regarding a potentially serious error source that can damage comparability. The guidelines should be expanded and include elements of light input harmonisation. 'Light' means that the guidelines, instead of being requirements to which countries must adhere, can identify a number of default mode design alternatives for countries to choose from and best practice methods regarding processes that are associated with mode choice, such as core questionnaire development, testing, adaptation and translation.
- Eurostat should offer stakeholders webinars and other capacity-building activities to enhance know-how regarding mode issues.
- Eurostat should strengthen the infrastructure around EU-SILC so that the design and implementation gap between countries shrinks, with the ultimate goal being a survey conducted under approximately the same essential survey conditions. This is a bold ambition given the current situation, in which we have a very large amount of variation with regard to modes and their implementation across countries. This goal can be achieved only if Eurostat establishes some kind of control centre or scientific leadership team much like surveys such as the ESS and SHARE have.
- Countries should be encouraged to perform methodological studies so that they have an idea about the nature of mode effects and how they could be addressed.
- Given that EU-SILC also needs to be comparable over time, countries should closely watch changes in coverage and response rates of modes over time.

23.8. Endnote

The variation in modes and mixes of modes in EU-SILC within and across countries is worrying and difficult to handle. Having said that, it should be noted that the guidelines contain only one page explicitly devoted to mode. That page tells countries what modes they can use, but there is nothing on mixes and how modes may be combined. Another worrying gap is the lack of guidelines on the core questionnaire and its adaptation and translation.

Mode effects are difficult to handle in single countries, and the problem magnifies in a situation in which the number of countries is increased. For instance, the World Values Survey covers more than 100 countries. Just contemplating applying a method to assess mode effects is overwhelming. Perhaps, it is more realistic to work on mode effects at a country level or within small groups of collaborating countries, and do our best to minimise those given constraints. It is important not to ignore this error source.

There are literally hundreds of published articles and conference papers, many manuals and book chapters, a weighty set of guidelines for cross-cultural surveys published by the University of Michigan (Survey Research Center, 2016) and a task force report on the quality of comparative surveys (Lyberg et al., 2021) that provide information on mode issues. Unfortunately, very little of this know-how has made its way into official statistics design principles and statistics production. The reasons for this state of affairs probably include a prioritisation of costs and of the more familiar TSE components of sampling, coverage and non-response. Of all methodological reports published by Eurostat, a minor number cover measurement errors. This probably has to do with methodological resources and know-how. Insights regarding measurement errors rest on theories of social cognition and social norms, and it seems as if those skills are played down in many NSIs despite the fact that they are so important for accuracy of measurement.

An endeavour such as EU-SILC is too big for individual countries to handle. A central coordinating centre that oversees the design implementation in individual countries and that can assist countries with methods for within-household selection, proxy interviewing, questionnaire testing, adaptation, translation and related issues is needed. The centre could also organise training events and provide materials that explain and discuss modes (and other methodological issues) and what needs to be done to enhance comparability.

References

Beinhauer, L, Snijkers, G. and Bakker, J. (2020), 'Towards a total error framework for sensor and survey data', paper presented at BigSurv20, 6 November.

Biemer, P. (1988), 'Measuring data quality', in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls, W. and Waksberg, J. (eds), *Telephone Survey Methodology*, Wiley, New York, pp. 321–340.

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality*, Wiley, New York.

Buelens, B., van den Brakel, J. and Schouten, B. (2018), 'Current methodologies to deal with mode effects in mixed mode designs', ESSnet MIMOD (Mixed Mode Designs for Social Surveys) deliverable 1 (https://www.istat.it/en/research-activity/international-research-activity/essnet-and-grants).

Couper, M. (2011), 'The future of modes of data collection', *Public Opinion Quarterly*, Vol. 75, No 5, pp. 889–908.

Couper, M. (2017), 'New developments in survey data collection', *Annual Review of Sociology*, Vol. 43, pp. 121–145.

de Leeuw, E. (2005), 'To mix or not to mix data collection modes in surveys', *Journal of Official Statistics*, Vol. 21, No 2, pp. 233–255.

de Leeuw, E. (2018), 'Mixed-mode: past, present and future', *Survey Research Methods*, Vol. 12, No 2, pp. 75–89.

de Leeuw, E. and Toepoel, V. (2018), 'Mixed-mode and mixed device surveys', in Vannette, D. and Krosnick, J. (eds), *The Palgrave Handbook of Survey Research*, Palgrave Macmillan, London, pp. 51–61.

de Leeuw, E., Dillman, D. and Hox, J. (2008), 'Mixed mode surveys: when and why', in de Leeuw, E., Hox, J. and Dillman, D. (eds), *International Handbook of Survey Methodology*, Lawrence Earlbaum Associates, Taylor & Francis, New York and London, pp. 299–316.

de Leeuw, E., Suzer-Gurtekin, T. and Hox, J. (2019), 'The design and implementation of mixed-mode surveys', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 387–408.

Dillman, D. and Edwards, M. (2016), 'Designing a mixed mode survey', in Wolf, C., Joye, D., Smith, T. W. and Fu, Y. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London, pp. 255–267.

Dillman, D., Smyth, J. D. and Christian, L. M. (2009), *Internet, Mail and Mixed Mode Surveys: The tailored design method*, Wiley, Hoboken, NJ.

Groves, R. (1989), *Survey Errors and Survey Costs*, Wiley, New York.

Groves, R., Fowler, F. J. Jr, Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009), *Survey Methodology*, 2nd edition, Wiley, New York.

Haan, M., Ongena, Y., Vannieuwenhuyze, J. and LeGloppa, K. (2017), 'Response behavior in a video-web survey: a mode comparison study', *Journal of Survey Statistics and Methodology*, Vol. 5, No 1, pp. 48–69.

Hill, C. A., Biemer, P., Buskirk, T., Callegaro, M., Cordova Cazar, A. L., Eck, A. et al. (2019), 'Exploring new statistical frontiers at the intersection of survey science and big data: convergence at "BigSurv18"', *Survey Research Methods*, Vol. 13, No 1, pp. 123–135.

Hochstim, J. R. (1967), 'A critical comparison of three strategies of collecting data from households', *Journal of the American Statistical Association*, Vol. 62, pp. 976–989.

Holbrook, A. L., Green, M. C. and Krosnick, J. A. (2003), 'Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability response bias', *Public Opinion Quarterly*, Vol. 67, No 1, pp. 79–125.

Holt, D. T. (2007), 'The official statistics Olympic challenge', *The American Statistician*, Vol. 61, No 1, pp. 1–8.

Jäckle, A., Roberts, C. and Lynn, P. (2010), 'Assessing the effect of data collection mode on measurement', *International Statistical Review*, Vol. 78, No 1, pp. 3–20.

Jowell, R. (1998), 'How comparative is comparative research?', *American Behavioral Scientist*, Vol. 42, pp. 168–177.

Klausch, L. T., Schouten, B. and Hox, J. J. (2017), 'Evaluating bias of sequential mixed-mode designs against benchmark surveys', *Sociological Methods and Research*, Vol. 46, No 3, pp. 456–489.

Kolenikov, S. and Kennedy, C. (2014), 'Evaluating three approaches to statistically adjust for mode effects', *Journal of Survey Statistics and Methodology*, Vol. 2, No 2, pp. 126–158.

Krosnick, J. and Alwin, D. (1987), 'An evaluation of a cognitive theory of response-order effects in survey measurement', *Public Opinion Quarterly*, Vol. 51, No 2, pp. 201–219.

Krumpel, I. (2013), 'Determinants of social desirability bias in sensitive surveys: a literature review', *Quality & Quantity*, Vol. 47, No 4, pp. 2025–2047.

Laaksonen, S. and Herskanen, M. (2014), 'Comparison of three modes for a crime victimization survey', *Journal of Survey Statistics and Methodology*, Vol. 2, No 4, pp. 459–483.

Lyberg, L. and Kasprzyk, D. (1991), 'Data collection methods and measurement error: an overview', in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N. and Sudman, S. (eds), *Measurement Errors in Surveys*, Wiley, Hoboken, NJ, pp. 238–257.

Lyberg, L., Pennell, B.-E., Hibben, K. C. and de Jong, J. (2021), *AAPOR/WAPOR task force report on quality in comparative surveys* (https://www.aapor.org/Education-Resources/Reports/AAPOR-WAPOR-Task-Force-Report-on-Quality-in-Compar.aspx).

Lynn, P., Hope, S., Jäckle, A., Campanell, P. and Nicolaas, G. (2012), 'Effects of visual and aural communication of categorical response options on answers to survey questions', *Institute for Social and Economic Research Working Papers*, No 2012-21, University of Essex, Colchester.

Martin, P. (2011), 'What makes a good mix? Chances and challenges of mixed mode data collection in the ESS', *Centre for Comparative Social Surveys Working Paper Series*, No 2, City, University of London, London.

Martin, P. and Lynn, P. (2011), 'The effects of mixed mode survey design on simple and complex analyses', *Centre for Comparative Social Surveys Working Paper Series*, No 2, City, University of London, London.

Medway, R. and Fulton, J. (2012), 'When more gets you less: a meta-analysis of the effect of concurrent web options on mail survey response rates', *Public Opinion Quarterly*, Vol. 76, No 4, pp. 733–746.

Roberts, C., Jäckle, A. and Lynn, P. (2006), 'Mixing modes in the European Social Survey: implications for data quality', paper presented at the American Association for Public Opinion Research, 18–21 May, Montreal.

Sakshaug, J., Cernat, A. and Raghunathan, T. (2019), 'Do sequential mixed-mode surveys decrease nonresponse bias, measurement error bias and total bias? An experimental study', *Journal of Survey Statistics and Methodology*, Vol. 7, No 4, pp. 545–571.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J. and Klausch, L. T. (2013), 'Disentangling

mode-specific selection and measurement bias in social surveys', *Social Science Research*, Vol. 42, pp. 1555–1570.

Smith, T. W. (2011), 'Refining the total survey error perspective', *International Journal of Public Opinion Research*, Vol. 23, No 4, pp. 464–484.

Survey Research Center (2016), *Guidelines for Best Practice in Cross-Cultural Surveys*, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor (http://ccsg.isr.umich.edu/).

Suzer-Gurtekin, T., Valliant, R., Heeringa, S. and de Leeuw, E. (2019), 'Mixed-mode surveys: design, estimation, and adjustment', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 409–439.

Tourangeau, R. (2017), 'Mixing modes: tradeoffs among coverage, nonresponse, and measurement error', in Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. et al. (eds), *Total Survey Error in Practice*, Wiley, Hoboken, NJ, pp. 115–132.

Tourangeau, R. and Smith, T. W. (1998), 'Collecting sensitive information with different modes of data collection', in Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls, W. and O'Reilly, J. (eds), *Computer Assisted Survey Information Collection*, Wiley, New York, pp. 431–453.

Tourangeau, R., Rips, L. J. and Rasinski, K. (2000), *The Psychology of Survey Response*, Cambridge University Press, Cambridge.

Vannieuwenhuyze, J. and Loosveldt, G. (2013), 'Evaluating relative mode effects: three methods to disentangle selection and measurement effects', *Sociological Methods & Research*, Vol. 42, No 1, pp. 82–104.

Vannieuwenhuyze, J., Loosveldt, G. and Mohlenberghs, G. (2010), 'A method for evaluating mode effects in mixed-mode surveys', *Public Opinion Quarterly*, Vol. 74, No 5, pp. 1027–1045.

Villar, A. and Fitzgerald, R. (2015), *Mixed-Mode Synthesis Report – ESS ERIC deliverable: Package no 5*. City University, London.

Villar, A. and Fitzgerald, R. (2017), 'Using mixed modes in survey research: evidence from six experiments in the ESS', in Breen, M. (ed.), *Values and Identification – Evidence from the European Social Survey*, Routledge, New York, pp. 273–309.

Voogt, R. and Saris, W. (2005), 'Mixed mode designs: finding the balance between nonresponse bias and mode effects', *Journal of Official Statistics*, Vol. 21, No 3, pp. 367–387.

Williams, N. and Ghimere, D. (2019), 'Mixed methods in a comparative context: technology and new opportunities for social science research', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 431–454.

Woltman, H., Turner, A. and Bushery, J. (1980), 'A comparison of three mixed-mode interviewing procedures in the National Crime Survey', *Journal of the American Statistical Association*, Vol. 75, No 371, pp. 534–543.

Zhang, L.-C. (2011), 'Topics of statistical theory for register-based statistics and data integration', *Statistica Neerlandica*, Vol. 66, No 1, pp. 41–63.

24 Preventing and mitigating the effects on data quality generated by mode of data collection, coding and editing

Sophie Psihoda, Nadja Lamei and Lars Lyberg ([129])

24.1. Introduction

This chapter focuses on two of the most critical stages in the life cycle of a survey: data collection and post-survey processing. It aims to provide new insights into the different practices of mode use, interviewing, coding and editing used by national statistical institutes (NSIs) for European Union Statistics on Income and Living Conditions (EU-SILC). An online questionnaire was sent out in June 2019 asking the NSIs about their concrete practices for these four tasks ([130]). In addition to the results of this expert survey, quality reports and literature have also been drawn on in the analyses that are provided here. In this chapter, after briefly discussing the methodological background of data collection, editing and coding, the issue of harmonisation in cross-country surveys and error concepts, we provide an overview of current NSI practices in data collection, coding and editing for EU-SILC. As will be seen, practices in data collection techniques and post-data editing steps differ significantly between NSIs – which is in line with EU-SILC being oriented towards output harmonisation. Because of different organisational or institutional backgrounds or survey traditions, this seems legitimate and helps to produce high-quality statistics for each country. However, with little input harmonisation it is doubtful if results from different countries can be interpreted in the same way. We have to assume that the different practices used have serious impacts on cross-country comparability as well as unnoticed effects in terms of the total survey error (TSE) framework. One of several conclusions is to view the increasing use of web interviewing as a chance to introduce elements that are input harmonised. In this way, some of the pending issues of comparison and survey error in the area of data collection and post-collection processing could be tackled.

([129]) Sophie Psihoda and Nadja Lamei are with Statistics Austria. Lars Lyberg was with Demoskop Inc. (Sweden) until his death in April 2021. The authors would like to thank Marlene Blüher (Statistics Austria) for her very useful comments. All errors are the authors' responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Sophie Psihoda (sophie.psihoda@statistik.gv.at) and Nadja Lamei (nadja.lamei@statistik.gv.at).

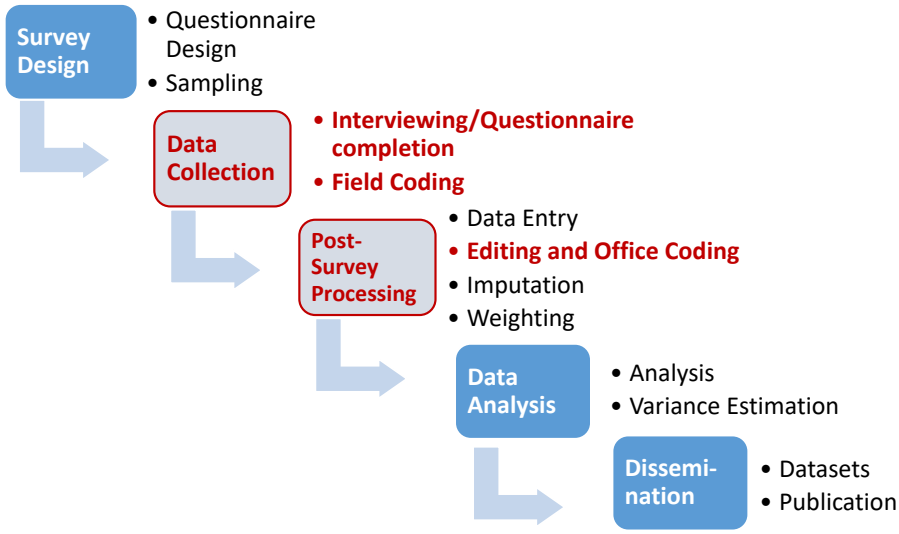
([130]) Altogether, the questionnaire contained 52 questions. The EU-SILC experts in the participating NSIs were asked to fill out the online questionnaire between May and July 2019, which mostly related to the 2018 EU-SILC (if not mentioned otherwise). In total, 30 NSIs provided answers that we could process.

24.2. Methodological background

24.2.1. Data collection, editing and coding in the life cycle of a survey

Figure 24.1 shows the different processing steps of a survey, including the main tasks of a computer-assisted survey, which are (i) survey design, (ii) data collection, (iii) post-survey processing, (iv) data analysis and (v) dissemination. Although interviewing takes place only in the data collection phase, editing and coding can take place in two of the five phases, namely post-survey processing and data collection. Coding can be separated into 'field coding', which is performed by the interviewers in

Figure 24.1: Data collection, editing and coding in the different stages of a survey



NB: Stages and tasks in red are dealt with in this chapter.

a face-to-face or telephone interview setting, and 'office coding', which is performed by specialised staff in the NSIs, subsequent to completion of the interviews (see Section 24.5). Similarly, data editing may be carried out in the data collection phase (by interviewers) or in the post-survey-processing phase (by office staff). These two forms of editing are often referred to as decentralised editing and centralised editing (see Section 24.6). For paper surveys (paper and pencil interviewing (PAPI)), the survey tasks are largely the same except for the tasks considered in detail in this chapter, namely data collection, coding and editing, which have undergone considerable changes in computer-assisted modes. To take a closer look at these three tasks in PAPI collections is, however, still necessary in the context of EU-SILC, since nine NSIs ([131]) (still) use PAPI as the main mode of data collection in any (or all) of the four waves.

([131]) Bulgaria, Czechia, Greece, Cyprus, Latvia, Lithuania, Poland, Romania and Slovakia.

24.2.2. Input harmonisation versus output harmonisation

EU-SILC is mainly output harmonised. By allowing different methods of data collection and post-survey processing in participating countries, we achieve high-quality country results that might not have been achieved had a standardised design been enforced across countries, since such a design may deviate from organisational or institutional backgrounds or survey traditions of individual NSIs. Ideally, the same underlying concepts would be measured by the methods that are most suitable in each country. For cross-country comparative studies, however, output harmonisation is challenging (Johnson et al, 2019). Some comparability issues were resolved following the introductory phase of EU-SILC through cooperation and communication. However, some issues remain. These include the large amount of variability in the design of household/address samples versus selected respondent design, which is maintained in the integrated European social statistics regulation (Euro-

pean Commission, 2019), the increasing number of survey modes and data sources (the introduction of web interviews in some countries, questionnaire versus register data use), and the differences in the design of the rotational schemes (4 years versus 6 years versus a longer panel duration).

Input harmonisation is an alternative paradigm to output harmonisation. Social surveys such as the European Social Survey and the Survey of Health, Ageing and Retirement in Europe are deliberately designed to accomplish comparability across countries. This is achieved using a mix of standardised procedures that each country must adhere to, such as the same data collection mode, similar interviewer training, a central questionnaire design and similar translation procedures. Small steps towards input harmonisation were also agreed by the EU Member States for EU-SILC (e.g. common guidelines for modules with example questionnaires, harmonisation of concepts and question wording for deprivation variables). However, there is a difference between guidelines and requirements. The comparative quality reports published by Eurostat (2018, 2019a, 2020) as well as in-depth reviews on comparability and methodological issues from the Network for the Analysis of EU-SILC (Net-SILC) and Second Network for the Analysis of EU-SILC (Net-SILC2) ([132]) have proved to be very valuable insights into the different settings and practices of EU-SILC in the participating countries. They show that process variability and its effects are considerable in EU-SILC.

24.2.3. Measurement and processing errors as sources of non-sampling error

Measurement errors are defined as the difference between the value of a characteristic provided by the respondent and the true (but unknown) value of that characteristic. The topic of measurement errors in surveys is well described in the literature (Biemer et al., 1991, 2017; Lyberg et al., 1997; Lepkowski et al., 2008; Wolf et al., 2016). Measurement errors can give rise to both bias (systematic errors) and variance (variable errors) in a survey estimate. Four sources of measurement error can be distinguished: the questionnaire, the data collection mode, the interviewer and the respondent. Hence, an 'observation error', as it is also called by Groves et al. (2009), occurs in the data collection stage of the survey. Similarly, measurement errors are defined by the EU-SILC implementing regulation for the quality reports (European Commission, 2004, Annex 1(h)) as 'errors that occur at the time of data collection. There are a number of sources for these errors, such as the survey instrument, the information system, the interviewer and the mode of collection'. Although there is acceptance that measurement errors can be large, it is rare that they are quantified for a given survey. The reason is that such assessments require special studies, such as randomised experiments, cognitive research studies, repeated measurement studies or record check studies. Assessment of measurement errors is always costly, and a preferred strategy is to prevent and mitigate these errors by using reliable processes for questionnaire design, pretesting, interviewer training, translation, and choice of mode and mixes of mode. Control of interviewers and their workloads is also important.

Data processing takes place after the data have been collected and comprises all activities aimed at converting the 'raw' survey data to a cleaned and corrected state so that they can be used in analysis, presentation and dissemination. It includes tasks such as data entry, editing, coding, file preparation and dissemination of an output file (Biemer and Lyberg, 2003). Likewise, processing errors, according to EU-SILC legislation, are 'errors in post-data-collection processes such as data entry, keying, editing and weighting' (European Commission, 2003). Sources of processing errors may be the individuals performing the data-processing task or sources specific to the technology used for automation of the data processing. Together with coverage and non-response errors, measurement and processing errors contribute to the non-sampling error of a survey. These errors, together with the sampling error, can be seen as parts of a TSE framework (Groves and Lyberg, 2010).

([132]) Earlier Net-SILC project documentation is available via the Collaboration in Research and Methodology for Official Statistics portal (https://ec.europa.eu/eurostat/cros/content/net-silc_en).

24.3. Mode issues

The mode of data collection refers to the medium that is used in a survey to (i) contact the sample members and (ii) obtain responses to the questions. The modes used for these two actions do not have to be the same. Modes have different characteristics, for instance the degree of contact with the respondent and the degree of computer assistance. Furthermore, each mode has its own general error structure. Thus, the choice of mode can have an impact on the TSE. It is generally known that self-administered modes such as mail and web modes generate more truthful responses to sensitive questions, such as those on drug use, sexual preferences and income. Nonetheless, satisficing and motivated misreporting may be more prevalent with self-administered modes. Modes involving interviewers, however, tend to trigger under-reporting of sensitive behaviours and characteristics due to social desirability bias (De Maio, 1984). Other things that can vary across modes include cost structures and response rates (Dillman, Smyth and Christian, 2014; Tourangeau, 2017; de Leeuw, 2018; de Leeuw, Suzer-Gurtekin and Hox, 2019). It can be difficult to isolate the mode effect – that is, the effect that is purely the result of using one mode rather than another. Often several modes are used in combination to benefit from each mode's comparative advantages (de Leeuw, 2005). Such designs are called mixed-mode designs (de Leeuw, 2018).

According to the EU-SILC methodological guidelines (Eurostat, 2019b), six modes of data collection are allowed: PAPI, computer-assisted personal interviewing (CAPI), computer-assisted telephone interviewing (CATI), self-administered questionnaires, web questionnaires and proxy interviewing, whereby a family member or some other knowledgeable person responds on behalf of the designated respondent. Proxy interviewing is always combined with a technical mode listed above; in EU-SILC it is generally discouraged but has to be used to a certain extent (see Section 24.4.4). The EU-SILC guidelines give priority to PAPI and CAPI, since face-to-face interviewing is considered the ideal mode for this kind of survey. Self-reports are not encouraged but are used anyway in some cases. The use of register data can be seen as a special mode, with the register data replacing data that otherwise would have been generated from interviews.

Our online consultation of NSIs taking part in EU-SILC, which related to the 2018 EU-SILC operation ([133]), shows that sending an advance letter is common when contacting sample members. Twenty countries use this approach before contacting respondents personally or in other ways. Seven countries use the face-to-face contact mode right away, whereas three use email as a first contact mode. If these initial contacts fail (e.g. in the event of advance letters being returned), countries tend to use interviewers or multiple modes. There is a tendency to move from less expensive to more expensive contact modes in cases of an initial failure in contacting. Nine countries mention that they switched contact strategy between waves. However, we do not have any information on which combinations are the most successful, as no NSI seems to keep track of this. The maximum number of contact attempts varies between one and nine, and this variation is probably due to varying levels of non-response and effort.

We further asked NSIs what modes (PAPI, CAPI, CATI, computer-assisted web interviewing (CAWI)) were used in the four interviewing waves (Table 24.1) ([134]). There is no mention of mixes of modes in the EU-SILC guidelines, but many countries use more than one mode within and between waves. Seventeen NSIs use more than one mode within waves and 16 use more than one mode between waves, and 14 NSIs use multiple modes both within and between waves. The reasons for using multiple modes are mainly to reduce costs and increase response rates. Some countries use specific preassigned modes for certain subgroups. Examples of such subgroups are those in which telephone numbers or email addresses are available or in which respondents have expressed mode preferences in previous waves. In some countries, interviewers can decide when to switch mode. This large variation in mode adminis-

([133]) For details, see footnote 130.

([134]) Throughout this chapter, we use the same abbreviations for interviewing modes as Eurostat uses in the EU-SILC guidelines: PAPI, CAPI, CATI and CAWI (see Eurostat, 2019b). Other modes, such as paper self-completion questionnaire, are mentioned in full. It should be mentioned that 'CAWI', which is used as a term in the EU-SILC documents, is increasingly referred to as 'web survey'.

tration has an impact on comparability and should be investigated more thoroughly.

There seems to be a general trend in statistics production towards using mixed-mode designs more frequently (Dillman, Smyth and Christian, 2014; de Leeuw, 2018). This is also the case for EU-SILC. Sixteen countries still use one mode, but nine use a sequential design (in which modes are administered one after the other), two use a concurrent design (in which modes are administered simultaneously) and three use both sequential and concurrent designs. The reasons for using mixed-mode designs include increased response rates, limited fieldwork capacity, respondent preferences, improved coverage and decreased fieldwork costs. It is not uncommon for NSIs to offer respondents a choice of mode (seven NSIs do this).

Six countries were already using CAWI in 2018: Denmark, Estonia, Latvia, Lithuania, Hungary and the Netherlands (see Table 24.1). The reasons for using CAWI include an increased use of web mode in general in the NSIs, difficulties obtaining telephone

Table 24.1: Interviewing mode per panel wave and country, 2018

Country	Wave 1	Wave 2	Wave 3	Wave 4
Austria	CAPI	CAPI, CATI	CAPI, CATI	CAPI, CATI
Belgium	CAPI	CAPI	CAPI	CAPI
Bulgaria	PAPI	PAPI, CAPI	PAPI, CAPI	PAPI, CAPI
Croatia	CAPI	CAPI	CAPI	CAPI
Czechia	PAPI, CAPI	PAPI, CAPI	PAPI, CAPI	PAPI, CAPI
Denmark	CATI, CAWI	CATI, CAWI	CATI, CAWI	CATI, CAWI
Estonia	CAPI	CAPI, CATI, CAWI	CAPI, CATI, CAWI	CAPI, CATI, CAWI
Finland	CAPI, CATI	CAPI, CATI	CAPI, CATI	CAPI, CATI
France	CAPI	CAPI	CAPI	CAPI
Germany	Paper self-completion	Paper self-completion	Paper self-completion	Paper self-completion
Greece	PAPI, CATI	PAPI, CATI	PAPI, CATI	PAPI, CATI
Hungary	CAPI, CAWI	CAPI, CAWI	CAPI, CAWI	CAPI, CAWI
Ireland	CAPI	CAPI	CAPI	CAPI
Italy	CAPI, CATI	CAPI, CATI	CAPI, CATI	CAPI, CATI
Latvia	PAPI, CAPI	PAPI, CAPI, CATI, CAWI	PAPI, CAPI, CATI, CAWI	PAPI, CAPI, CATI, CAWI
Lithuania	PAPI, CAPI, CAWI	PAPI, CAPI, CATI, CAWI	PAPI, CAPI, CATI, CAWI	PAPI, CAPI, CATI, CAWI
Luxembourg	CAPI	CAPI	CAPI	CAPI
Malta	CAPI	CAPI, CATI	CAPI, CATI	CAPI, CATI
Netherlands	CATI, CAWI	CATI, CAWI	CATI, CAWI	CATI
Norway	CATI	CATI	CATI	CATI
Poland	PAPI, CAPI	PAPI, CAPI	PAPI, CAPI	PAPI, CAPI
Portugal	CAPI	CAPI	CAPI	CAPI
Romania	PAPI	PAPI	PAPI	PAPI
Serbia	CAPI	CAPI	CAPI	CAPI
Slovakia	PAPI	PAPI	PAPI	PAPI
Slovenia	CAPI	CAPI, CATI	CAPI, CATI	CAPI, CATI
Spain	CAPI, CATI	CAPI, CATI	CAPI, CATI	CAPI, CATI
Sweden	CAPI, CATI	CAPI, CATI	CAPI, CATI	CAPI, CATI
Switzerland	CAPI, CATI	CAPI, CATI	CAPI, CATI	CAPI, CATI
United Kingdom	CAPI	CAPI	CAPI	CAPI

NB: 'Paper self-completion' means that respondents fill in the paper questionnaire themselves without the help of an interviewer.
Source: Net-SILC3 online consultation, June and July 2019 (only completed cases). Answers relate to the 2018 EU-SILC operation.

numbers, respondent preferences, decreased costs for the producers, a deliberate push towards web mode through incentives and to increase response rates. It is also planned to use CAWI in the near future or it is already in the testing phase in four other countries (Austria, Finland, Norway and Sweden). Linked to the use of web questionnaires are issues regarding adaptation to mobile devices (Couper, Antoun and Mavletova, 2017). Only two countries have reported pretesting the web questionnaire to increase its suitability for multiple devices and browsers, and only three countries have made such adaptations. No NSI has reported that an EU-SILC app is currently being developed. Based on the lack of examples of adaptations, our interpretation is that any adaptation problems are of a technical nature or due to a lack of resources rather than cognitive issues related to the response process.

The survey asked all NSIs whether or not they use web mode and what they see as the advantages and disadvantages of using mobile devices for data collection in EU-SILC. Table 24.2 lists the answers given.

There are two ways to mitigate mode effects in comparative surveys. One can try to enforce the use of one common mode (or one common mixed-mode protocol, meaning that all countries use the same mode per wave), as is the case in the European Social Survey, or one can try to adjust for mode effects if more than one mode is administered (de Leeuw, 2018; de Leeuw, Suzer-Gurtekin and Hox, 2019). For EU-SILC, rather than ask for more conformity regarding mode combinations across Member States, which is not very realistic, we could ask for mode studies to get some sense of the size of any mode effects and their relative importance compared with other specific error sources. However, these studies would need to be comparative to establish whether the effect of a particular mode differs between countries. Only six NSIs reported having attempted to investigate mode effects. These attempts seem to have been rather basic and included studies comparing income data from EU-SILC and registers, comparisons of CATI and CAPI subsamples, comparisons of indicators of selectivity and precision across modes, and analyses of mode paradata such as interview duration and number of 'don't know' responses. Just two NSIs mentioned that they had attempted to adjust for mode effects. Overall, the lack of studies and the fact that many questions about mode issues in our consultation were left unanswered indicate that NSIs should pay more attention to mode issues. Methods for mitigating and adjusting for mode effects may be rather unfamiliar territory in many NSIs, and there is probably a need for guidelines on how to design EU-SILC so that errors emanating from mode choice or mixed-mode designs are minimised given the resources available.

Table 24.2: Advantages and disadvantages of the use of mobile devices

Advantages of mobile device use	Disadvantages of mobile device use
• Growing expectation of some population groups for modern devices • Increases the response rate for younger people • No interviewers • Mode flexibility • Decreased costs • May improve timeliness • May improve coverage • Can accommodate many response alternatives • Greater perceived sense of anonymity • Easier data transmission • May send a message that the NSI is up to date and modern	• No control over who is responding • Technical problems • Difficult for respondents • Questions have to be short • Adaptation is costly • No interviewer assistance • Limits regarding mobile device capacity • Limited space for instructions • Increased item non-response • Need for a mixed-mode design • EU-SILC is too complicated for use with mobile devices

Source: Third Network for the Analysis of EU-SILC online consultation, May–July 2019 (only filled-in cases).

24.4. The interviewing process

The interviewing task can be broken down into several subtasks: contacting the respondents, informing them about the subject and persuading them to participate, conducting the interview, recording non-response and its reasons, and recording actual information and meta-information (e.g. 'contactability' for the next wave). EU-SILC has some specific complexities due to its rotational panel structure and follow-up rules. For each panel wave and/or mode the procedures may differ. Subtasks range from those for which the interviewer has to follow a very structured protocol (e.g. in administering the questionnaire) to those that rely on the interviewer's judgement (e.g. how to contact respondents) (Lessler, Eyerman and Wang, 2008). Some countries use self-administered modes either with paper questionnaires (Germany) or with web questionnaires (six countries in 2018, as mentioned above).

The interviewing task in EU-SILC depends greatly on the circumstances of the whole survey process, for example whether interviewers are responsible for collecting information on the entire household or on selected respondents, or for collecting information on all household income components or only some (due to register data use). Ideally, the interview mode should not have an effect on the measurements, but due to different levels of workload and varying methods for supporting or supervising the interviewers this may be the case. Therefore, interviewer training is crucial for collecting high-quality and comparable data. The quality of the data is largely an outcome of the interviewers' understanding and implementation of the questions. The circumstances of their work, such as type of contract, payment, workload and interviewer turnover, affect the quality of the survey. However, we found that, besides teaching the interviewers certain standards, quality control of interviews and interviewer conduct is not widely implemented, not well documented and not standardised enough to allow for consistent reporting.

24.4.1. Interviewer numbers, workloads and payment schemes

The number of interviewers involved in data collection depends on the number of interviews conducted in each country (which is set out by EU regulations or triggered by national needs) and is reflected in the workload. In terms of costs, organisational aspects, response rates, cooperation of households and quality, stable, reliable and well-trained interviewer staff are preferable. The initial investment in recruitment and training is costly.

Regarding the number of interviewers, a distinction by mode was made in our survey. Seven Member States reported that they used PAPI in the 2018 EU-SILC. A minimum number of 83 PAPI interviewers in Poland and a maximum of 792 in Romania were involved. The average number of PAPI interviewers was 362. The share of PAPI interviewers new to EU-SILC in the 2018 operation was reported to be very low – on average 4 %.

With regard to CAPI , 22 out of 30 responding countries used this mode in 2018. The lowest numbers of interviewers were seen in Switzerland (one) and Sweden (four); the maximum number of 869 was seen in Poland; and on average 196 interviewers were involved. Of these, about 15 % were new to EU-SILC in 2018.

CATI was used by 15 of the responding countries. The number of interviewers involved ranged from 6 in Malta to 132 in Finland (four countries reported 'don't know' for the number of CATI interviewers). An average of 52 CATI interviewers were working on the 2018 EU-SILC. Of these, about 35 % started working on EU-SILC only in 2018.

Compared with the survey carried out for Net-SILC2, which referred to the 2009 EU-SILC (Glaser et al., 2015), some changes can be seen in the average number of interviewers by mode: the average number of CATI interviewers across countries has decreased (from 104 to 52), the average number of CAPI interviewers has increased (from 123 to 196), and PAPI has remained the mode with the highest number of interviewers (311 in the 2009 EU-SILC – this does not include Poland and Italy, which at that time each had more than 1 000 PAPI interviewers). It is important to note that the increasing

number of countries using CAWI as an additional if not the main mode for EU-SILC of course reduces the numbers of interviewers required. During that period of nine EU-SILC operations, nearly all countries changed their mode structures or their default modes. A full overview of these year-on-year changes can be found in the Eurostat comparative quality reports (Eurostat, 2018, 2019a, 2020).

On average, across Member States and depending on the mode, a large majority of the interviewers may be re-employed from the last wave(s). The fluctuation is higher for CATI than for CAPI or PAPI, which may be due to the types of contracts used and/or characteristics of the staff ([135]). Compared with the earlier survey, in which higher shares of 'new' interviewers were recorded, this can be interpreted as an increase in the stability of the field implementation of EU-SILC in recent years.

A supervisor in any mode should control, organise and support the staff and give feedback to the interviewers and sometimes correct them. The numbers of supervising staff vary considerably by mode, country and number of interviewers employed in each mode: from 8 to 236 for PAPI (on average 65 across countries), from 2 to 236 for CAPI [136] (on average 30) and from 1 to 66 for CATI (on average 12).

Payment schemes often show common characteristics across countries by mode: nearly all countries using PAPI pay their interviewers per responding sample household, sometimes per person or adding a factor for household size. Rarely, travelling time or distance and number of contacts with households also affect remuneration. Dependency of pay on data quality is explicitly mentioned only by one country (Romania). Latvia additionally reports the 'degree of qualification of the interviewer' as a criterion for payment differences (in all modes).

For CAPI interviewers, Czechia, Finland and Switzerland pay hourly wages. France reported that its payment scheme for CAPI is grade related. In Croatia, the CAPI fieldwork was reported to be conducted partly by internal interviewers (47 employees in 2018) and partly by external interviewers (73), who work for the NSI on a civil contract. Although the external interviewers are reimbursed per questionnaire, including all costs (e.g. travel), for internal interviewers data collection is considered their regular work, including overtime work, and they receive a wage. Ireland provides a basic salary for CAPI interviewers. In addition, it uses a tiered bonus system, with the amount of the bonus based on the percentage of allocated households successfully interviewed. Poland also reported a regular salary for PAPI and CAPI interviewers. In the United Kingdom, the interviewers receive an agreed annual salary based on the fixed number of hours they are expected to work each week. The other countries pay per sample unit (household or person) and additionally take into account the number of contacts (Bulgaria), travel burden (Bulgaria, Hungary and Slovenia) and non-response documentation (Bulgaria, Latvia and Austria). Data quality and response rates are also considered (Hungary, Malta and Austria).

CATI staff are paid per hour by five countries (Austria, Finland, Latvia, Slovenia and Switzerland). Estonia mentioned that its CATI interviewers conduct not only EU-SILC but all other kinds of surveys in CATI mode and that they receive a fixed wage. Italy and Malta pay their CATI staff per household.

The workload of interviewers differs considerably between countries and modes. From a survey theory perspective, extensive clustering of the sample by interviewers should be avoided. If the workload for single interviewers is too extensive, the error associated with interviewers increases (Groves, 2004; West, Kreuter and Jaenichen, 2013; West and Blom, 2017; Mneimneh et al., 2019). On average across countries, the workload is 16 assigned sample units (households or, for selected respondent countries, individuals) for PAPI, 73 for CAPI and 169 for CATI. As CATI is usually under tighter control and quicker to conduct than CAPI or PAPI, a bigger workload may seem acceptable. The highest number of assigned sample units is reported by Italy for both CAPI (698 assigned sample units) and CATI (1 091). As for the number of completed interviews, Italy reported the highest number of interviews carried out by one CAPI interviewer (611), but the average for CAPI is considerably lower (96), meaning that very high

([135]) In Austria, for example, the CAPI staff are relatively stable and comprise freelance interviewers, whereas CATI personnel are employed only during the EU-SILC data collection and consist mainly of students, who do this as a part-time job.

([136]) Both instances of 236 refer to Poland, which used the same staff for CAPI and PAPI.

workloads are rare. In addition, the highest number of CATI interviews successfully completed was reported in Italy, where one CATI interviewer carried out 656 interviews. The average was 319 successful interviews per CATI interviewer in Italy ([137]). An even higher average was reported for Latvia (371 interviews). These are very large workloads by any standard and they raise questions about data quality. The goal must be reasonably even workloads within countries that are not too small or too large.

The effects of workload on interviewer variance could therefore be very large in some countries. We therefore recommend that the number of sample units assigned to one interviewer should be restricted, as should the variation in assignment size between interviewers. The ability to do this will be constrained by practical aspects of recruitment, training and interviewer organisation and the fact that interviewing staff typically work on more than one survey. As Biemer and Lyberg (2003, p. 166) put it, deciding on the optimal number of interviewers for a survey should take into account both interviewer variance concerns and logistical factors related to fieldwork.

24.4.2. Interviewer training and quality control

Interviewers require support, and their training and supervision are key to the quality of the data collected. As the module questionnaire is new each year and other things in the questionnaire or the methodology can change, it is advisable that all interviewers – even experienced ones – are trained or briefed before starting fieldwork.

As the example of Belgium shows, the details and length of a training session depend on the overall interviewer experience and their experience in EU-SILC: a briefing of approximately 2 hours is carried out for experienced EU-SILC interviewers; a full day of training is carried out for those not familiar with EU-SILC; and 2 days of training are carried out for those new to the job. Czechia highlighted the organisational aspects of supervisors' training, since they in turn train the interviewers at regional level; the process in Serbia and Spain is similar. Learning from each other and passing on knowledge from experienced interviewers to colleagues new to the project are also methods of choice (e.g. in Serbia and the United Kingdom). Training does not always need to be provided using classroom sessions: Estonia reported that updates by email are used; guidelines and manuals are often sent to the interviewers (e.g. in Greece, the Netherlands and Austria); video conferences are held (in Slovakia and Finland); and online refresher courses and assessment tests (in Italy, the Netherlands and the United Kingdom) are provided to those who have received training previously. Asking the interviewers to test the survey instrument on themselves or on a fellow interviewer is encouraged by many NSIs. Norway mentioned carrying out a second round of briefings after interviewers have conducted approximately half of the interviews. To ensure the high quality of the next data round and understand the quality issues in the current round, individual or group feedback sessions (debriefings) can also be valuable. This was mentioned by Austria and Poland.

The length of interviewer training depends on the survey mode, with CAPI interviewers receiving the most training – on average 13 hours, varying between 2 hours (Belgium and Estonia) and 42 hours (Latvia) – followed by PAPI interviewers – on average 10 hours, ranging from 3 hours (Slovakia) to 18 hours (Greece) – and CATI interviewers – on average 6 hours, varying between 1 hour (Finland) and 28 hours (Spain). Inexperienced interviewers received, on average, 1 hour (PAPI, CATI) or 3 hours (CAPI) of additional training.

We also asked a general question about quality control procedures used for interviewers. Greece and France reported not having any, whereas the other countries (except for Denmark, Germany and Luxembourg, which did not provide any answers) answered affirmatively. The variety of checks and methods used are interesting and are listed in Table 24.3. They range from monitoring response rates, interview duration and other process

([137]) It should be noted that the (gross and net) sample sizes in Italy are especially large compared with other countries. The 2017 quality report (the latest available report) shows a net sample size of 22 226 households. In addition, the latest available comparative quality report (the 2016 operation) shows that Italy has the highest sample size among Member States (followed by Spain with a net sample of about 3 000 households fewer).

data at the interviewer level to technical checks of the data (all of which are carried out by nearly all countries), monitoring the Global Positioning System (GPS) data of the interviewer laptops (Czechia and Hungary) and carrying out callbacks at household level to verify that interviews took place and that the data are correct (Hungary, Ireland, Poland, Serbia and Slovenia). Some techniques change during the course of fieldwork (as mentioned by Spain). Some are mode specific (the Netherlands mentions the ability of supervisors to listen to interviews).

The literature on interviewer quality control is quite extensive, covering methods such as monitoring, reinterviews, callbacks, keeping track of workload sizes, collecting interview paradata and detecting signs of fabrication (e.g. University of Michigan, 2016; West and Blom, 2017; Blasius, 2018; Lyberg, Japec and Tongur, 2019; Ongena, Hahn and Dijkstra, 2019; Sharma, 2019). The literature on training, however, is less extensive. Some examples of training in comparative settings are provided by Robbins (2019), Weiss, Sakshaug and Börsch-Supan (2019) and Ackermann-Piek et al., 2020. For EU-SILC, it can be concluded that a lot is done in terms of training and quality control of measurement (i.e. data editing); however, procedures for data collection are very diverse, and quality control with an emphasis on field procedures was not reported extensively.

Table 24.3: Quality control procedures used, 2018

Country	Methods mentioned in the questionnaire
Austria	Non-response at unit and item levels is checked. There are detailed checks of data plausibility, duration of interview and response rates, all of which are also carried out at interviewer level.
Belgium	During the fieldwork, interviewer performance and response rates are followed. If necessary, there is intervention during the fieldwork. At the end of the fieldwork it is considered whether a collaboration can be continued next year.
Bulgaria	Procedures are not centralised. The supervisor from each district checks for completeness of the questionnaires, coding and household coverage, and inspects extreme income values .
Croatia	Detailed logical controls and checks of the collected survey data are performed. Some of the programmed controls include the following: • controls relating to activity status / economic status of the household members, • controls of logical relationships between household members, • controls of all income categories that were recorded at the household level and the individual level, • household costs exceeding certain value limits, • periods of maternity and parental leave as well as maternity and parental benefits, • logical connections between individual questions. All potential errors are listed for each interviewer and checked if corrections of the initially entered data are needed. If correction is really needed, then the data will be corrected, and a note will be made about the correction. Furthermore, in the CAPI questionnaire there are defined automatic controls that result in a warning message when the answer to a question does not seem to be correct.
Czechia	GPS monitoring.
Estonia	All the respondents receive a feedback questionnaire afterwards by email. In addition, the person who is in charge of quality control checks the CATI interview recordings. We also check the CAPI, CATI and CAWI interviews if there are any errors. We always check the interviews that are done by new interviewers.
Finland	Validation calls and text message surveys are used for some parts of the sample (both respondents and non-respondents).

Country	Methods mentioned in the questionnaire
Hungary	The GPS data of laptops are registered during the interview. Interview times and dates are also monitored. In total, 5 % of successful interviews are monitored by a supervisor by phone – they ask respondents whether they were interviewed by a representative of the statistical office, what the topic was, whether they received any incentive, if they were content with the incentive, and what their general perception of the interviewer was. For CAWI, the IP addresses of the computers are checked.
Ireland	The interviewers are divided into teams of 10 and each team reports to a field coordinator, who monitors the work of the interviewers with regard to number of hours worked, number of households visited and distance covered in order to reach the households. Coordinators spot check households to ensure that interviewers genuinely did call. When the data are received in the office, individual interviewers are contacted if there are consistent or recurring faults in their work.
Italy	A subsample of interviewed households is re-called in order to clarify issues pertaining to the interviews.
Latvia	Measurement errors are detected by logical checks and verification of received data, including verification online during the fieldwork. During the fieldwork, distribution of the main variables is compared between all interviewers. In cases of significant differences in results, the interviewer is asked to peruse the methodology of the variable again and to update the answers for the households interviewed (if needed). Compliance of the database with Eurostat requirements is checked using the SAS program.
Lithuania	Supervisors check partly or fully completed questionnaires. Logical and arithmetical checking rules are integrated into the data entry programme.
Malta	A number of validation checks are integrated into the data collection programme; however, when the data are received, a sample is taken by the interviewer and further validation checks and quality control procedures are carried out.
Netherlands	The results per interviewer are recorded by a tool and compared. Supervisors can listen in.
Norway	Interviews per hours worked, the percentage of contacts that lead to interviews, etc., are checked.
Poland	Telephone calls and visits to households are carried out where the interviews have already been completed. Real-time control of data received from interviewers.
Serbia	We contact households by telephone during fieldwork and check some of the basic data collected.
Slovakia	We use manuals for control questionnaires.
Slovenia	Control letters are sent twice during fieldwork to respondents to find out whether the interviewer visited the household. In the letter, there are also some key questions that we can compare with the answers in the survey. The interviewers send the data twice a week and controllers monitor some key variables. If discrepancies occur, they discuss them with the EU-SILC methodologist, and after that we contact the interviewer to find out what the problem was.
Spain	In the first week of fieldwork, there is an exhaustive control of the work, verifying that the methodology, definitions, etc., have been adequately assimilated by the interviewers. In the following weeks, there will be different controls and monitoring of the work.
Sweden	Methods used are in accordance with the International Organisation for Standardisation's International Standard 20252, including training of staff, validation of interview data and monitoring of interviews.
Switzerland	The fieldwork is carried out under contract by an external private institute. We listen to live interviews, and there are plausibility checks online.
United Kingdom	There are checks built into the Blaise code, which runs the questionnaire. The interviewer will be alerted if any responses fail these checks and will be instructed either to fix the error or to explain why the check failure is genuine and should be allowed to remain. Response and contact rates are regularly monitored and field managers will liaise with interviewers to ensure that these rates do not get too low.

NB: Answers are edited from the original entries for the sake of brevity and clarity.

Source: Net-SILC3 online consultation May–July 2019 (only filled-in cases).

24.4.3. Language of interviews

Four countries reported using three different languages in their EU-SILC interviews: Belgium (Dutch, French and German), Bulgaria (Bulgarian, English and Turkish), Finland (Finnish, Swedish and English) and Switzerland (German, French and Italian). This is also usually the case for Austria (German, Turkish and Bosnian/Croatian/Serbian), but the 2018 EU-SILC was an exception, as due to the introduction of a new survey tool it was not possible to implement foreign language versions that year. Another six countries reported using a second language: Estonia (Estonian and Russian), Greece (Greek and Turkish), Italy (Italian and German), Latvia (Latvian and Russian), Malta (Maltese and English) and Norway (Norwegian and English). Sometimes, providing interviews in more than one language is decided by law (in countries with more than one official language). In other cases, it can be assumed that this was a deliberate decision to achieve good response rates and representativeness of the sample, and that the additional languages were chosen by weighing up the population who would potentially be excluded and the additional costs of translation and surveying.

A considerably lower share of interviews was conducted in a second language than in a first language. For face-to-face interviews, this ranged from 0.5 % in Greece (interpreters of the Turkish language conducted these) to 22 % in Estonia ([138]), and for CATI interviews the share ranged from 1.7 % in Norway to 34.5 % in Switzerland. Regarding interviews in a third language, Switzerland reported a rate of 7.2 % and Finland reported a rate of 0.7 %.

It is striking that, according to the information from the consultation, far fewer than half of all countries use more than one interviewing language. The question arises as to how those parts of the population that are not able to speak the survey language(s) are handled proficiently by the survey organisations. Information on ad hoc translations by individuals present at the interview may be a relevant quality criterion; proxy interviewing or a mix of proxy answers and interpretation is also possible and can be seen as a risk to quality. As concluded in Net-SILC2 (Glaser et al., 2015), the large portion of proxy interviewing in some countries seems to suggest that this is how they have solved the language problem, despite the fact that the EU-SILC guidelines discourage the use of proxy interviewing and that, when it is used, some kind of accuracy assessment should be made.

([138]) Please note that only Estonia, Greece, Italy and Finland provided numbers of interviews by language and mode; Italy and Finland reported that all respondents were native speakers; the others did not specify this.

24.4.4. Proxy interviews

When using the term 'proxy', we refer to an interview situation in which another person answers the survey questions instead of the target person. This is not uncommon as a means to reduce non-response and costs. Intuitively, proxy interviewing should have a negative effect on data quality, and general research shows that this is the case (Cobb and Krosnick, 2009; King, Cook and Hunter Childs, 2012; Sudman et al., 1994; see also Chapter 7 of this book). Proxy responses lead to so-called encoding problems; in other words, the proxy respondent simply does not know certain facts associated with the sampled person (Biemer and Lyberg, 2003). The quality of proxy interviews – or the size of the resulting measurement error – is likely to depend on the subject of the question, the relationship between the proxy respondent and the target respondent, how much information they usually share and the survey mode (Blair, Menon, and Bickart, 1991, p. 145). It has also been shown that proxy responses can lead to fewer socially desirable responses.

Proxy interviews in EU-SILC are supposed to be exceptions to the rule that states that personal interviews should be carried out for all individuals aged 16 and over. Three situations are accepted as reasons for proxy interviews based on the *Methodological Guidelines and Description of EU-SILC Target Variables* (DocSILC065; Eurostat, 2019b).

1. The person was physically or cognitively unable to respond.
2. The person was not available during the data collection phase.
3. The person had language problems (no interpretation possible).

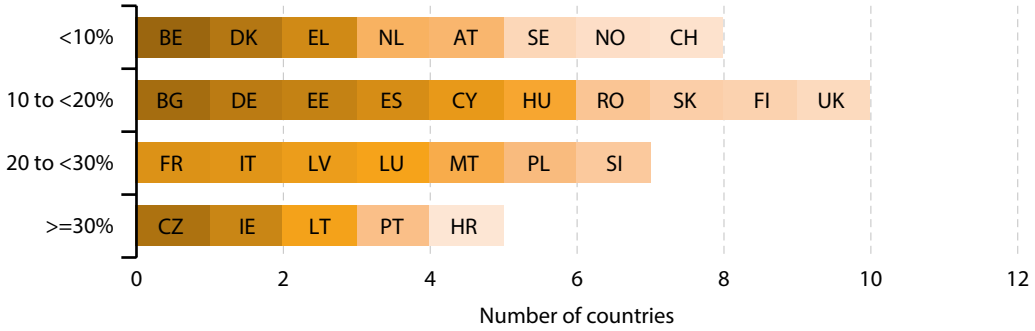
All these impediments must endure for the duration of the fieldwork period; for example, if a person is on a business trip, this is no reason to accept a proxy unless she or he is away for the whole fieldwork period. DocSILC065 further specifies that proxy interviews are to be especially avoided for income variables, health information and detailed labour information.

For countries using the 'selected respondent model', the household respondent (in most cases the selected respondent) is asked for information about all household members. Therefore, these countries cannot be fully compared with other countries, since proxy means something else for them. The EU-SILC comparative quality report for 2018 (the latest available year) shows that proxy rates differ considerably. Most countries (10) have a proxy rate of between 10 % and 20 %. Eight countries present a proxy rate of below 10 % (selected respondent model countries Denmark, the Netherlands, Norway and Sweden, plus Austria, Belgium, Greece and Switzerland). A proxy interview rate between 20 % and 30 % is reported for seven countries, and five countries exceed a rate of 30 % (Figure 24.2). As can be seen from complementary data from our online questionnaire, it is not always easy to code a proxy situation correctly, so not all situations are reflected in the data (see also Chapter 7).

In our questionnaire, we collected the reasons countries accept proxy interviews. Czechia, Serbia and Slovakia mention non-availability (the second reason given in the list above) as the sole reason. Malta accepts health problems (the first reason listed above) as the only reason. Estonia, Croatia, Hungary and Romania report that the first and second reasons listed above are valid reasons for replacing a personal interview with a proxy interview. The Netherlands names health problems (the first reason listed above) and language problems (the third reason listed above). The majority of countries (16) accept all three reasons in that list. Norway specified that it allows proxies for the work questions only when the target person is not available. Slovenia reported that module questions are never asked in proxy interviews. Austria conceded that, in very restricted cases, including if the designated person refuses, it is accepted that another person in the household can provide the proxy responses needed. However, the questions on well-being and satisfaction are never asked in proxy interviews.

In contrast to other decisions associated with fieldwork practices, situations when a proxy should be accepted are consistent across modes within countries. Most countries, namely 15, do not use a mix of personal and proxy questions in interviews. Ten countries report doing this occasionally (Bulgaria, Denmark, Greece, Spain, Italy, Lithuania, Hungary, Slovenia, Finland and Sweden). Sometimes, this happens when a respondent does not have the relevant knowledge. In this case, a personal interview is completed with some questions completed by proxy (e.g. parents answer for younger respondents). Sometimes, it is the other way around: the questionnaire is initiated with a proxy interview, and a telephone personal interview is used to check and complete the questionnaire (e.g. in

Figure 24.2: Proxy interview rates by country, 2018



Source: Eurostat (2020).

Spain). Switching interview mode (e.g. from personal interview to proxy interview) during one interview cannot be documented, since there is only one Eurostat target variable for type of interview (RB260). As explained by Italy, in the data it is not indicated whether a mix of proxy and personal interviews took place. The only information provided is how the interview started (proxy or personal).

Only three countries have changed their practices regarding proxies in EU-SILC during the past 5 years. This has been due to changes in data collection (in Spain the CATI mode is used to complement the CAPI mode to complete missing personal interviews) or non-acceptance of proxies for certain topics (health in Finland, some modules in Poland). For CAWI (or any other self-administered mode), it is much harder to know who is actually filling out the questionnaire. Denmark, Lithuania, Latvia and the Netherlands explicitly ask who is filling out the questionnaire in multi-person households; the Netherlands even mentioned control questions. In Estonia and Hungary, there are no direct checks on who is answering.

24.5. Coding

24.5.1. Definition, goals and tasks of coding

Coding is defined as a process in which raw survey data are classified and assigned code numbers or categories in order to use them for estimation, tabulation and analysis (Biemer and Lyberg, 2003; Lyberg and Kasprzyk, 1997). This definition may be applied to numeric and alphanumeric data. More specifically, in this section we want to focus on the classification of textual information, and thus 'the translation of nonnumeric material into numeric data' (Groves et al., 2004, p. 305). Examples of variables that typically require coding are level of education, academic field of study, industry worked in, occupation, place of work, socioeconomic status and geospatial information. Moreover, coding is also needed for answers to open-ended questions, such as respondent or interviewer remarks. A prerequisite for the successful translation of textual information into numeric data is a classification system or category framework, which is referred to as a 'nomenclature', a 'code structure' or a 'code list' with the purpose of grouping similar responses of group objects. The codes must reflect the intended uses of the variable and not just theoretical aspects. Many nomenclatures have a hierarchical structure with several levels. This is particularly relevant to classification systems for industry or occupation, such as the general industrial classification of economic activities within the European Union (NACE) or the International Standard Classification of Occupations (ISCO). A detailed description of classification systems and their necessary attributes is provided by Groves et al. (2004, p. 306).

In general, we can differentiate between three kinds of coding systems: (i) strictly manual coding, whereby the coder relies on paper, pencil and manuals, (ii) computer-assisted coding, whereby a human coder assigns code numbers while working interactively with a computer, and (iii) fully computer-automated coding without any human intervention. In addition, we can distinguish between 'office coding', which refers to coding pursued by NSI officers at their desks in the post-survey-processing phase, and 'field coding', whereby the interviewer codes the verbal answer of the respondent into a numeric category during the interview – a distinction which can affect coding quality (Collins and Courtenay, 1985).

The quality of coding can be judged by looking at two aspects: coding structure and coder variance. Systematic coding errors arise when the code structures are poorly conceptualised, for example combining two responses that have different analytical implications into one category, or if classification systems are not maintained according to changing realities (e.g. coding of occupations). Both introduce measurement error. Coder variance 'is a component of the overall variance of a survey statistic arising from different patterns of use of code structures by different coders' (Groves et al., 2004, p. 316). Differences in the use of a coding system may arise as a result of, for example, varying interpretations of response words for a given code category or the coders' tendency to use a residual code such as 'other'. Coder variance follows the same logic that applies to interviewer variance (Jabine and Tepping, 1973; Sturgis, 2004; Conrad, Couper and Sakshaug, 2016) and asks for similar countervailing measures, name-

ly training and control of coders, the development of clear coding instructions and limiting the number of cases coded by each coder.

24.5.2. Looking inside the black box of EU-SILC coding

In DocSILC065, recommendations concerning coding relate only to the coding of education level and occupation. For education level, the variables PE020 (level of current education attended) and PE040 (highest level of education attained) are both coded using the classification system of the national International Standard Classification of Education integrated mapping (Eurostat, 2019b, 260f, 263f). For occupation, which is collected by the variable PL051, coding is performed using the classification system of ISCO-08, which succeeded the previous nomenclature ISCO-88 in 2011 (p. 291).

The Net-SILC3 online consultation shows that, of the 30 NSIs that took part in the survey, altogether 22 NSIs reported asking open-ended questions (16 NSIs) or allowing for respondent remarks (15 NSIs) in the 2018 EU-SILC data collection; nine NSIs included both in their questionnaires. A similar number (22) allowed interviewer remarks to be collected.

Although NSIs with open-ended questions in their questionnaires were equally likely to code the answers manually (eight NSIs) or by using a combination of automated and manual coding (seven NSIs), only 8 of the 15 NSIs collecting respondent remarks also coded them: seven manually and one using a combination of manual and automated coding. For the automated coding, only one NSI reported using commercial software, whereas the other six NSIs reported using an in-house system. With regard to interviewer remarks, 14 NSIs used them to edit answers in the current data collection and/or considered them in the next wave, whereas eight NSIs did not use them any further. Thus, although the text information of answers to open-ended questions is used quite well by NSIs, information from respondent or interviewer remarks remains unused in many NSIs (Table 24.4). This may be attributed to a lack of resources and time pressure in the post-survey-processing phase.

Table 24.4: Coding techniques/practices for answers to open-ended questions and respondent and interviewer remarks, 2018

Open-ended questions			Respondent remarks			Interviewer remarks		
Manual coding	Combination of automated and manual coding	No coding	Manual coding	Combination of automated and manual coding	No coding	Edited answers using information	Considered for the next wave	Did not use them
Belgium	Austria	United Kingdom	Austria	Switzerland	Croatia	Austria	Austria	Estonia
Estonia	Bulgaria		Belgium		Estonia	Belgium	France	Greece
Germany	Hungary		Czechia		Greece	Croatia	Malta	Hungary
Greece	Italy		Malta		Latvia	Czechia	Netherlands	Ireland
Latvia	Netherlands		Slovakia		Poland	Finland	Poland	Italy
Slovenia	Norway		Serbia		Slovenia	France	Slovakia	Romania
Serbia	Sweden		Spain		United Kingdom	Latvia	Switzerland	Spain
Spain						Serbia		United Kingdom
						Slovenia		

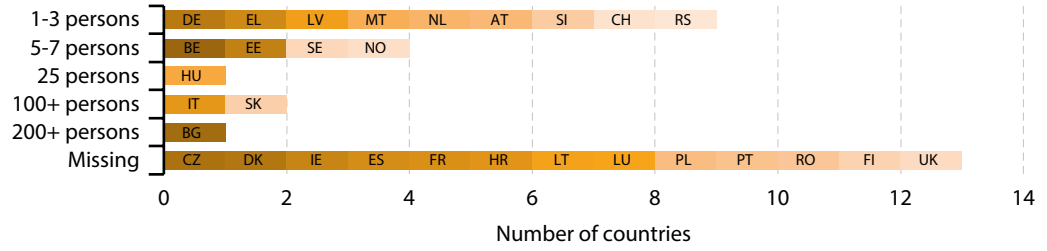
NB: In total, 16 countries collected answers to open-ended questions, 15 collected respondent remarks and 22 collected interviewer remarks.

Source: Net-SILC3 online consultation May–July 2019. Answers relate to the 2018 EU-SILC operation.

In terms of the personnel involved in coding activities (coders), seven NSIs reported involving between one and three people, and four NSIs reported involving five to seven people. Rather high numbers were reported by Hungary (25), Slovakia (100), Italy (153) and Bulgaria (293). The remaining 13 countries did not provide answers to these questions (Figure 24.3). The large numbers in Bulgaria and Italy can be explained by the fact that both countries rely on field interviewers for coding answers to open-ended questions. Spain, Latvia and Austria also reported relying on field interviewers; however, Latvia and Austria seem to count only the number of office staff involved in coding, while Spain did not answer this question.

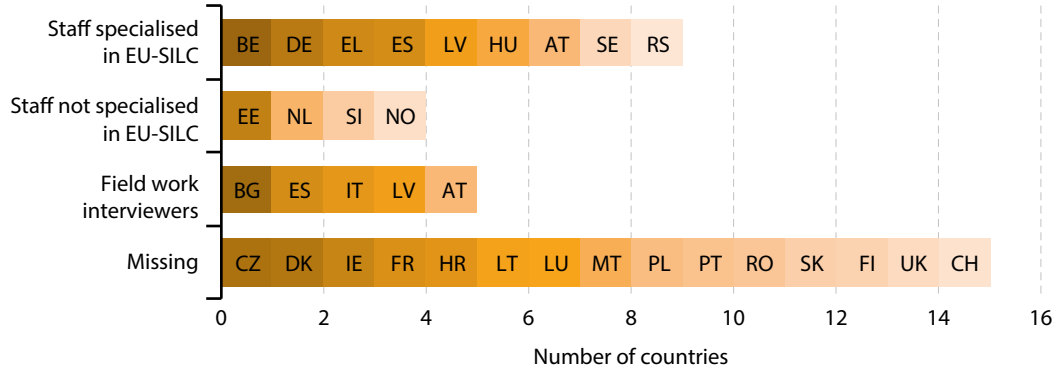
Coding by staff specialised in EU-SILC was reported most frequently, namely by nine NSIs. Four NSIs reported that the coding of answers to open-ended questions and respondent remarks is carried out by staff not specialised in EU-SILC (Figure 24.4). We also asked NSIs whether the people involved in the coding of answers to open-ended questions or respondent remarks receive specific training: 12 reported that they did, six reported that they did not and 12 did not answer. No relationship can be seen between the provision of training and whether the coders are specialised in EU-SILC or are field interviewers. Quality checks of the coding operation such as logical checks or supervision of the coding during the fieldwork and re-contacting households are implemented by only four NSIs.

Figure 24.3: Number of staff involved in coding of answers to open-ended questions and respondent remarks, 2018



Source: Net-SILC3 online consultation May–July 2019 (only filled-in cases (n = 30)). Answers relate to the 2018 EU-SILC operation.

Figure 24.4: Type of staff involved in coding of answers to open-ended questions and respondent remarks, 2018



Source: Net-SILC3 online consultation May–July 2019 (only filled-in cases (n = 30); multiple answers possible). Answers relate to the 2018 EU-SILC operation.

Furthermore, we asked the NSIs to provide us with a rough estimate of the number of working hours used for coding activities in EU-SILC. Of the 14 NSIs that answered this question, seven reported using fewer than 100 working hours for coding, four reported using between 100 and 300 working hours, and three reported using between 700 and 1 000 working hours. In Malta, the Netherlands, Austria and Slovenia, the relatively low numbers of coders correspond to a low number of working hours, whereas Bulgaria, Estonia, Hungary, Norway and Sweden reported both more working hours and a higher number of coders. However, because of the many missing values it is not possible to draw conclusions about the relationship between working hours and number of coders.

24.6. Data editing

24.6.1. Definition, goals and tasks of data editing

Data editing can be defined as 'the inspection and alteration of collected data, prior to statistical analysis' in order to 'improve the quality of the data' (Groves et al., 2004, p. 319). As '[s]ome uses of the term editing also include coding and imputation' (p. 319), it becomes difficult to determine when editing ends and imputation starts. Sometimes, the two are distinguished by the method used to replace an unreliable or missing value. If the value were to be estimated based on the respondents' values for other items in the survey, this would be called imputation. In the case of replacing a missing or unreliable value with a true value received from register data or data based on logical rules, the procedure would be closer to editing.

Although the goal of improving the quality of the data can be achieved in different ways, some definitions of editing focus on the goal of cleaning and correcting the data. The US Federal Committee on Statistical Methodology (1990) defines editing as:

> *procedures designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures.* (Biemer and Lyberg, 2003, pp. 226–227).

Similarly, the definition of the Economic Commission for Europe, which has been endorsed by the International Work Session on Statistical Data Editing, focuses on 'the procedure for detecting ... and for adjusting ... errors resulting from data collection or data capture' (United Nations Economic Commission for Europe, 1997, p. ix). However, apart from cleaning and correcting the data, Granquist (1984) notes that editing also aims to deliver information about the quality of the data and provide the basis for future improvement of the survey.

The goals of data editing are pursued by different kinds of checks, of which the following six are most frequently used (Groves et al., 2004, p. 319): (i) range edits; (ii) ratio edits; (iii) balance edits limiting the input of a certain value by a minimum and maximum value; (iv) consistency edits referring to logical consistency, for example by age and marital status; (v) checks of implausible outliers; and (vi) comparisons with historical data in the case of panel surveys.

Furthermore, there are several ways to distinguish between different dimensions of editing. One dimension can be seen in the differentiation between editing as a validating procedure following a micro-editing approach and aimed at 'identifying inconsistencies and suspicious values and then, if deemed necessary, correcting the value' (Lyberg and Kasprzyk, 1997, p. 355) of the individual record, and editing as a statistical procedure following a macro-editing approach with 'between-record checks aimed at detecting outliers in univariate or multivariate distributions, reviews of aggregate data' (p. 355). A second dimension of editing is proposed by Biemer and Lyberg (2003, p. 228), who differentiate between deterministic edits, carried out in the case of certain errors, and query edits, carried out when values are identified as suspicious, and may be in error, but further investigation is needed. If query edits have a substantial probability of affecting the estimates, they are called stochastic edits. Moreover, Biemer and Lyberg (2003) distinguish between fatal or critical edits, which must be corrected in order to make the data record usable, and non-critical edits which need not necessarily

be corrected. A third dimension of editing can be seen in differentiating between explicit edits and implicit edits. An edit is explicit if there are 'rules to which data must conform in each survey record' (Groves et al., 2004, p. 321). In contrast, edits are implicit if there are 'similar rules, logically deduced given [there is] some explicit edit rule that must be followed' (p. 321).

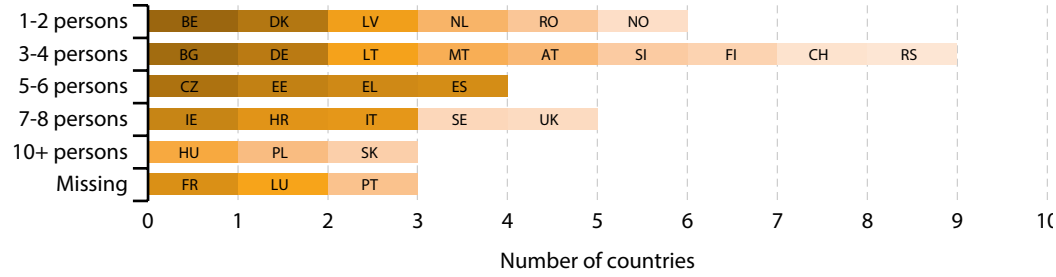
Furthermore, editing can also be distinguished according to the stages of the survey process. In the case of editing at the stage of data collection, the interviewer conducts the edit checks alone or in cooperation with the respondent according to interviewer instructions. However, most data editing is performed during the post-survey-processing stage. This can happen prior to data entry when the survey is carried out by PAPI and questionnaires arrive at the statistical office and are classified as accepted, rejected or action needed. Editing can also be pursued during data entry, for example when CATI is used and a supervisor corrects the data after the interview but before the data are sent for further data-editing steps. The majority of data editing will, however, happen after data capture, when a system of edits and checks is applied with very limited manual intervention and mostly automated intervention through the use of software. Lastly, output editing is the final stage during the post-survey-processing phase, when editing is performed with a focus on the values that are presented to users (Biemer and Lyberg, 2003).

As with any stage of the life cycle of a survey, there are certain drawbacks associated with editing. In particular, editing can be time-consuming and costly. According to the Federal Committee on Statistical Methodology (1990) and Gagnon, Gough and Yeo (1994), the monetary costs of editing have been estimated at between 20 % and 40 % of the total survey budget. Although these estimates are comparatively old and continuous technological developments have contributed to efficiency gains during data collection and post-survey processing, as shown in Section 24.6.2, many working hours are still used for data editing in many NSIs participating in EU-SILC. In this respect, concerns regarding so-called overediting – the overuse of editing – have been raised in the literature (Granquist and Kovar, 1997; Lyberg and Kasprzyk, 1997). Granquist and Kovar (1997, p. 419) note that 'time and resources spent on editing might have a higher quality payoff if allocated to other tasks, for example, to efforts related to raising the response rates'. Indeed, several studies have shown that extensive editing does not contribute to overall improved data quality, since editing follows the Pareto principle – 'i.e., few errors are responsible for the majority of actual value changes' (Lyberg and Kasprzyk, 1997, p. 358). However, with EU-SILC being a longitudinal survey, it also needs to be considered that editing in terms of minor changes may have no influence on the cross-sectional results, but it may be relevant for the longitudinal study for consistency reasons. Therefore, a judgement on when overediting takes place becomes even more difficult with EU-SILC.

24.6.2. Looking inside the black box of EU-SILC data editing

Recommendations on editing are found only four times in DocSILC065. In rather general terms, the word 'editing' is mentioned in the same sentence with imputation and is described as 'taking into account auxiliary values from the current wave, previous and future waves (countries using a rotational or long-term panel will apply a common imputation method for the cross-sectional and longitudinal component)' (Eurostat, 2019b, p. 54). In terms of concrete variables, the word 'editing' is mentioned only three times in DocSILC065. In the case of variable HY170G/HY170N, which measures the value of goods produced for own consumption, editing procedures are referred to as applying a market price to the type of goods consumed after respondents report the quantity of the goods (p. 233). For the variables HY080G/HY080N (regular inter-household cash transfer received, p. 222) and HY130G/HY130N (regular inter-household cash transfer paid, p. 227), 'editing' is mentioned regarding its function to 'limit measurement error (for both alimonies and others) and to avoid capital transfer'. Moreover, in this instance Eurostat recommends collecting 'capital transfers in parallel so as to avoid having to collect them in regular transfers' (p. 227). Thus, with all three variables editing is defined as the alteration of data by using auxiliary data that are not collected by the survey itself.

Figure 24.5: Number of staff involved in data editing, 2018



Source: Net-SILC3 online consultation May–July 2019 (only filled-in cases (n = 30)). Answers relate to the 2018 EU-SILC operation.

The Net-SILC3 online consultation shows that, for the majority of the NSIs, one or two people (six NSIs) or three or four people (nine NSIs) are involved in the data-editing process. Four NSIs reported involving five or six people in EU-SILC data editing, and five NSIs reported involving seven or eight people. Rather large numbers of people active in the data-editing process were reported by Hungary (25 people), Poland (44) and Slovakia (100). France, Luxembourg and Portugal reported that they do not know how many people are involved (Figure 24.5).

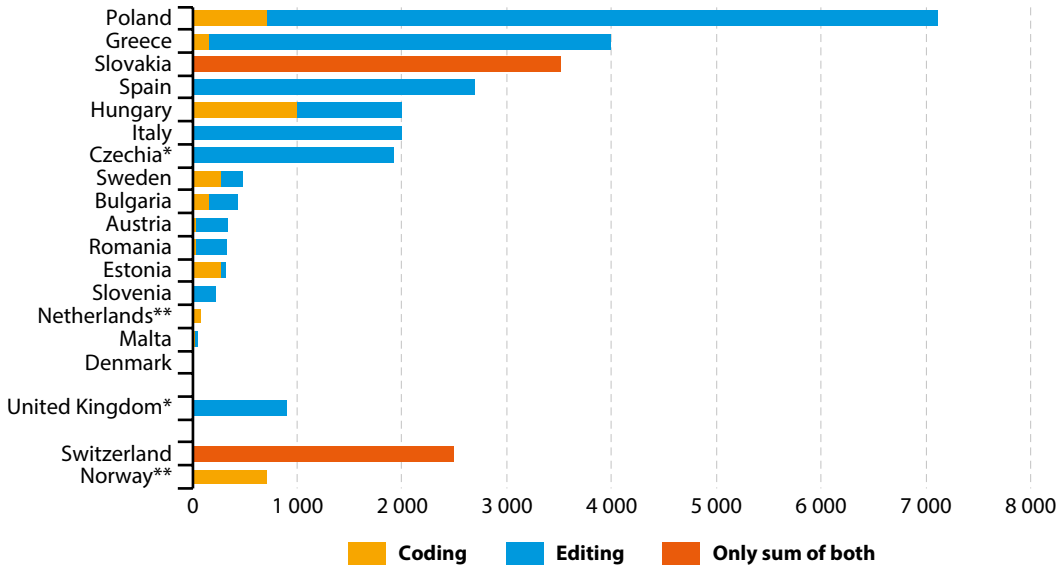
Even more variability can be observed with regard to the reported working hours used for data editing. Of the 17 NSIs that reported the number of hours used for data editing, those of four countries (Estonia, Malta, the Netherlands and Norway) use fewer than 200 hours since they report not doing any editing. Another five NSIs (in Bulgaria, Austria, Romania, Slovenia and Sweden) reported that they use between 200 and 300 working hours for data editing countries are Bulgaria, Austria, Romania. Four NSIs use between 900 and 2 000 working hours, with Hungary and the United Kingdom reporting between 900 and 1 000 working hours and Czechia and Italy reporting roughly 2 000 working hours. The highest numbers of working hours were reported by Spain (2 700 working hours), Greece (3 840) and Poland (6 405). Although the variability is very large between the 17 NSIs providing answers, it should be mentioned that the difference in reported working hours for editing also depends largely on the interpretation of the activities that are referred to as editing (as opposed to imputation, programming of target variables, integrating of register data, and the like). As pointed out in Section 24.6.1, this can be quite challenging.

Most NSIs (26) reported that the data editing is carried out by staff working on EU-SILC. Slovakia reported that the editing process is carried out by the fieldwork interviewers – this explains their large number of data editors (100). The Netherlands reports not doing any editing at all, and three NSIs did not answer the question. In addition, two thirds of the NSIs participating in the survey reported that their data editors receive specific training on EU-SILC; only eight NSIs reported that this is not the case, and three did not answer the question. However, quality checks are not only important in terms of the data-editing activities; data editing should also provide information on the data quality of the survey in general and on potential improvements for the survey. Hence, we also asked the NSIs if data editors provided feedback to questionnaire designers on EU-SILC. This was the case for 19 NSIs, of which 17 also reported that questions are not changed as a result of data editors' feedback. Nine NSIs reported that data editors do not provide feedback to the questionnaire designers, and three did not answer the questions.

24.7. Connecting modes, coding and editing

Considering the sum of working hours spent on editing and coding, it is obvious that, of those countries that reported separate numbers of hours for both tasks, the majority of the post-survey-processing time is used for editing in most NSIs. Only Estonia and Sweden reported using more time for coding than for editing. In addition, Hungary

Figure 24.6: Working hours used for coding and data editing, 2018



NB: Countries are ordered by the sum of working hours for coding and editing. * indicates no hours reported for coding. ** indicates no hours reported for editing.

Source: Net-SILC3 online consultation May–July 2019 (only filled-in cases (*n* = 19)). Answers relate to the 2018 EU-SILC operation.

reported using the same amount of time for editing as for coding. In terms of editing, Greece and Poland reported the largest numbers of hours in total as well as in comparison with coding. The NSIs in those two countries also reported the highest total number of hours for coding and editing (Poland: 7 100 hours; Greece: 3 990 hours). Fewer but still considerably large numbers were reported by Slovakia; it reported using 500 working hours for editing and coding, without providing information on the distinction between hours used for coding and hours used for editing. Spain and Switzerland reported that they use a total of 2 700 and 2 500 working hours, respectively, for coding and editing of EU-SILC data. Particularly few working hours, namely below 100, were reported by Denmark, the Netherlands and Malta (Figure 24.6).

In order to determine if the numbers of working hours used for editing and coding are related to the data collection mode, we grouped NSIs that use the same modes of data collection. NSIs that use CAWI in their data collection reported the lowest total numbers of hours for coding and editing (Denmark, Estonia, the Netherlands and Norway). The exception is Hungary. In contrast, NSIs using PAPI in their data collection reported the highest numbers of hours for coding and editing, particularly if a second non-self-administered mode is used, such as in Czechia, Greece and Poland. In between are the NSIs that use CAPI and/or CATI for their data collection, namely Austria, Italy, Malta, Slovenia, Spain, Sweden and the United Kingdom.

24.8. Conclusions and recommendations

Conclusion I: Different data collection modes – can consistency of data collection be increased by interviewer training and fieldwork quality controls?

- As shown in Sections 24.3 and 24.4, there is a lot of variation, in terms of the use of data collection modes and mixes of modes, among the NSIs participating in EU-SILC. This can be ascribed to different research traditions,

methodological and financial resources, organisational practices in the NSIs, and national regulations. For example, NSIs that have access to register data do not need to ask respondents all the questions, which can influence the decision to use a certain mode.

- The potential to enhance data quality and comparability in EU-SILC may lie, to a great extent, in the effective training of interviewers and having quality control measures in place during the data collection phase. An exchange of best practices and training materials could be a valuable contribution in this regard. In addition, using a default design option for the data collection process and standardising procedures would be a useful step towards greater comparability.
- The large variation in the quality control measures for fieldwork is not acceptable. Different things are checked in different countries, and sometimes one gets the impression that these are the methods used in the individual NSIs, which can make a method mix suboptimal. Instead, we believe that Eurostat should include a set of minimum fieldwork controls in the EU-SILC guidelines for all countries to perform, to enhance comparability across countries. It is then up to individual NSIs to complement these controls with other controls that may benefit the quality of their national surveys. Concepts such as interviewer variance do not seem to be well known in many NSIs. This should be corrected by Eurostat by adding a quality control module to its guidelines.

Conclusion II: Computer-assisted web interviewing and its potential for quality improvement – the time is now!

- Data collection in the form of CAWI is becoming increasingly important for EU-SILC. It provides a way to deal with budget constraints and decreasing response rates. However, so far there have been almost no general guidelines from Eurostat on data collection using CAWI. Hence, NSIs pursue the implementation of CAWI mostly on their own. This causes additional work, as previous data collection modes also need to be supported.
- Both technical challenges (e.g. use of mobile devices, different technical systems and respondent behaviour, and login data) and challenges with regard to content (a new question design due to the need for self-completion by respondents) must be mastered. Experience in many NSIs is limited or relevant to other kinds of surveys (shorter surveys, cross-sectional surveys only, and so on).
- More coordination and cooperation in terms of official guidelines, best practice workshops and a harmonised CAWI strategy may not only reduce the burden for each NSI but also encourage further consistency between countries before country-specific organisational settings and values fully define CAWI practices.

Conclusion III: Diverging practices of coding and editing – an often-neglected field in which improvements are needed!

- NSIs differ in terms of coding practices and in terms of the people responsible for coding. The majority of NSIs asking open-ended questions perform the coding manually. A similar level of variability can be found in terms of the qualifications of coders. Although in some NSIs coding is carried out by people specialised in EU-SILC, in others it is carried out by fieldwork interviewers or other office staff. This variability could affect measurement error, as explained in Section 24.5.2. A way to counteract this effect would be to provide an official coding and editing manual. Results show that only some NSIs can refer to a manual for post-survey-processing work. It seems that problems with coding and editing are underestimated by many NSIs. Guidance from Eurostat would therefore be welcome.
- Country-specific differences in the practices of coding and editing may also be related to the large differences in the number of working hours used for coding and editing. These differences may partly be due to the modes

of data collection used and the availability of register data.

- Some country differences are due to different perceptions of what coding and editing entail and how this work is organised. At this stage, it is hard to compare efforts across countries. More detailed information on and documentation of coding and editing is needed. Moreover, exchange of best practices in coding and editing could be valuable in the enhancement of further standardisation across the participating NSIs and the reduction in measurement errors.

References

Ackermann-Piek, D., Silber, H., Daikeler, J., Martin, S. and Edwards, B. (2020), 'Interviewer training guidelines of multinational programs: a total survey error perspective', *Methods, Data, Analyses*, Vol. 14, No 1, pp. 35–60.

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality*, Wiley, Hoboken, NJ (http://onlinelibrary.wiley.com/book/10.1002/0471458740).

Biemer, P., Groves, R., Lyberg, L. Mathiowetz, N. and Sudman, S. (eds) (1991), *Measurement Errors in Surveys*, Wiley, Hoboken, NJ.

Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. et al. (eds) (2017), *Total Survey Error in Practice*, Wiley, Hoboken, NJ.

Blair, J., Menon, G. and Bickart, B. (1991), 'Measurement effects in self vs. proxy responses to survey questions: an information-processing perspective', in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N. and Sudman, S. (eds), *Measurement Errors in Surveys*, Wiley, Hoboken, NJ, pp. 145–166.

Blasius, J. (2018), 'Fabrication of interview data', *Quality Assurance in Education*, Vol. 26, No 2, pp. 213–226.

Cobb, C. and Krosnick, J. (2009), 'Experimental test of the accuracy of proxy reports compared to target report with third-party validity', paper presented at the American Association for Public Opinion Research annual meeting, 16 May 2009, Hollywood, FL.

Collins, M. and Courtenay, G. (1985), 'A comparison study of field and office coding', *Journal of Official Statistics*, Vol. 1, pp. 221–227.

Conrad, F., Couper, M. and Sakshaug, J. (2016), 'Classifying open-ended reports: factors affecting the reliability of occupation codes', *Journal of Official Statistics*, Vol. 32, No 1, pp. 75–92.

Couper, M., Antoun, C. and Mavletova, A. (2017), 'Mobile web surveys', in Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. et al. (eds), *Total Survey Error in Practice*, Wiley, Hoboken, NJ, pp. 133–154.

de Leeuw, E. (2005), 'To mix or not to mix: data collection modes in surveys', *Journal of Official Statistics*, Vol. 21, pp. 233–255.

de Leeuw, E. (2018), 'Mixed-mode: past, present, and future', *Survey Research Methods*, Vol. 12, No 2, pp. 75–89.

de Leeuw, E., Suzer-Gurtekin, Z. and Hox, J. (2019), 'The design and implementation of mixed-mode surveys', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 387–408.

De Maio, T. (1984), 'Social desirability and survey measurement: a review', in Turner, C. F. and Martin, E. (eds), *Surveying Subjective Phenomena*, Vol. 2, Russell Sage Foundation, New York, pp. 257–282.

Dillman, D., Smyth, J. and Christian, L. M. (2014), *Internet, Phone, Mail, and Mixed-mode Surveys: The tailored design method*, 4th edition, Wiley, Hoboken, NJ.

European Commission (2003), Commission Regulation (EC) No 1981/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the fieldwork aspects and the imputation processes, OJ L 298, 17.11.2003, p. 23 (https://eur-lex.europa.eu/eli/reg/2003/1981/oj).

European Commission (2004), Commission Regulation (EC) No 28/2004 of 5 January 2004 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the detailed content of inter-

mediate and final quality reports, OJ L 005, 9.1.2004, p. 42 (https://eur-lex.europa.eu/eli/reg/2004/28/oj).

European Commission (2019), Commission Regulation (EU) No 2019/1700 of 10 October 2019 establishing a common framework for European statistics relating to persons and households, based on data at individual level collected from samples, OJ L 005, 9.1.2004, p. 42 (https://eur-lex.europa.eu/eli/reg/2019/1700/oj).

Eurostat (2018), *EU-SILC Comparative Quality Report 2016* (https://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports).

Eurostat (2019a), *EU-SILC Comparative Quality Report 2017* (https://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports).

Eurostat (2019b), *Methodological Guidelines and Description of EU-SILC Target Variables – DocSILC065 – 2019 operation*, Eurostat, Luxembourg.

Eurostat (2020), *EU-SILC Comparative Quality Report 2018* (https://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports).

Federal Committee on Statistical Methodology (1990), 'Data editing in federal statistical agencies', *Statistical Policy Working Papers*, No 18. US Office of Management and Budget, Washington DC.

Gagnon, G., Gough, H. and Yeo, D. (1994), *Survey of Editing Practices in Statistics Canada*, unpublished report of Statistics Canada.

Glaser, T., Kafka, E., Lamei, N., Lyberg, L. and Till, M. (2015), 'European comparability and national best practices of EU-SILC: a review of data collection and coherence of the longitudinal component', *Net-SILC2 Working Papers*, No 5/2015, Statistics Austria, Vienna.

Granquist, L. (1984), 'On the role of editing', *Statistical Review*, Vol. 2, pp. 105–118.

Granquist, L. and Kovar, J. (1997), 'Editing of survey data: how much is enough?', in Lyberg, L., Biemer, P., Collings, M., de Leeuw, E., Dippo, C., Schwarz, N. and Trewin, D. (eds), *Survey Measurement and Process Quality*, Wiley, Hoboken, NJ, pp. 415–435.

Groves, R. (2004), *Survey Errors and Survey Costs*, Wiley, Hoboken, NJ.

Groves, R. and Lyberg, L. (2010), 'Total survey error: past, present, and future', *Public Opinion Quarterly*, Vol. 74, No 5, pp. 849–879.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2004), *Survey Methodology*, Wiley, Hoboken, NJ.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009), *Survey Methodology,* 2nd edition, Wiley, Hoboken, NJ.

Jabine, T. and Tepping, B. (1973), 'Controlling the quality of occupation and industry data', invited paper presented at the International Statistical Institute meeting, Vienna, Austria.

Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds) (2019), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ.

King, T., Cook, S. and Hunter Childs, J. (2012), 'Interviewing proxy versus self-reporting respondents to obtain information regarding living conditions', paper presented at the American Association for Public Opinion Research annual meeting, 17–20 May, Orlando, FL.

Lepkowski, J., Tucker, C., Brick, M., de Leeuw, E., Japec, L., Lavrakas, P. et al. (eds) (2008), *Advances in Telephone Survey Methodology*, Wiley, Hoboken, NJ.

Lessler, J. T., Eyerman, J. and Wang, K. (2008), 'Interviewer training', in de Leeuw, E., Hox, J. J. and Dillman, D. A. (eds), *International Handbook of Survey Methodology,* Taylor & Francis / Lawrence Erlbaum Associates, New York, pp. 442–460.

Lyberg, L. and Kasprzyk, D. (1997), 'Some aspects of post-survey processing', in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. and Trewin, D. (eds), *Survey Measurement and Process Quality*, Wiley, Hoboken, NJ, pp. 353–370.

Lyberg, L., Japec, L. and Tongur, C. (2019), 'Prevailing issues and the future of comparative surveys', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 1055–1082.

Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. and Trewin, D. (eds) (1997), *Survey*

Measurement and Process Quality, Wiley, Hoboken, NJ.

Mneimneh, Z., Lyberg, L., Sharma, S., Vyas, M., Bal Sathe, D., Malter, F. and Altwaijri, Y. (2019), 'Case studies on monitoring interviewer behavior in international and multinational surveys', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 731–770.

Ongena, Y., Hahn, M. and Dijkstra, W. (2019), 'Multinational event history calendar interviewing', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 643–664.

Robbins, M. (2019), 'New frontiers in detecting data fabrication', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 771–805.

Sharma, S. (2019), 'Paradata, interviewing quality, and interviewer effects', PhD dissertation, University of Michigan, Ann Arbor.

Sturgis, P. (2004), 'The effect of coding error on time use surveys estimates', *Journal of Official Statistics*, Vol. 20, No 3, pp. 467–480.

Sudman, S., Bickart, B., Blair, J. and Menon, G. (1994), 'The effect of participation level on reports of behavior and attitudes by proxy reporters', in Schwarz, N. and Sudman, S. (eds), *Autobiographical Memory and the Validity of Retrospective Reports*, Springer, New York.

Tourangeau, R. (2017), 'Mixing modes: tradeoffs among coverage, nonresponse, and measurement error', in Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. et al. (eds), *Total Survey Error in Practice*, Wiley, Hoboken, NJ, pp. 115–132.

United Nations Economic Commission for Europe (1997), *Statistical Data Editing Volume No. 2: Methods and techniques*, United Nations, New York and Geneva (https://unece.org/fileadmin/DAM/stats/publications/editing/SDE2.pdf).

University of Michigan (2016), *Cross-Cultural Survey Guidelines* (https://ccsg.isr.umich.edu).

Weiss, L., Sakshaug, J. and Börsch-Supan, A. (2019), 'Collection of biomeasures in a cross-national setting', in Johnson, T., Pennell, B.-E., Stoop, I. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 623–641.

West, B. and Blom, A. (2017), 'Explaining interviewer effects: a research synthesis', *Journal of Survey Statistics and Methodology*, Vol. 5, No 2, pp. 175–211.

West, B., Kreuter, F. and Jaenichen, U. (2013), '"Interviewer" effects in face-to-face surveys: a function of sampling, measurement error, or nonresponse?', *Journal of Official Statistics*, Vol. 29, No 2, pp. 277–297.

Wolf, C., Schneider, S., Behr, D. and Joye, D. (2016), 'Harmonizing survey questions between cultures and over time', in Wolf, C., Joye, D., Smith, T. W. and Fu, Y. (eds), *The SAGE Handbook of Survey Methodology*, SAGE Publications, London pp. 502–524.

25 The potential role of EU-SILC topics as part of an integrated social survey: the case of the United Kingdom

Ria Sanderson and Pete Betts ([139])

25.1. Introduction

In this chapter, we describe a programme that is transforming the way the UK Office for National Statistics (ONS) collects data through the integration of its social surveys and the applicability of this model to the topics relevant to European Union Statistics on Income and Living Conditions (EU-SILC).

Within the United Kingdom, data on the topics relevant to EU-SILC are compiled from the Household Finance Survey (HFS). At midnight on 31 December 2020, the United Kingdom's transition period with the EU ended, and the country entered into a new trade and cooperation agreement. The United Kingdom therefore left the European Statistical System and, with the exception of some transmissions required to support the Withdrawal Agreement and Northern Ireland Protocol, all obligations to provide data to Eurostat ceased. This includes EU-SILC. Hereafter, we therefore refer to the potential to collect information on the topics relevant to EU-SILC, rather than the compilation of these indicators themselves.

The HFS was established in 2017 to harmonise and integrate the collection of information on income and expenditure within the ONS. It brings together the longitudinal Survey on Living Conditions (SLC) and the cross-sectional Living Costs and Food Survey (LCF). In the former, the same households are contacted in multiple annual waves, whereas the latter is cross-sectional and collects information on both income and expenditure (which is used for national accounts purposes). The HFS introduced a joint sample design for its components and harmonised core questionnaire content; further details of the design are given in Section 25.2. Interviewers who visit sampled households administer the HFS through computer-assisted personal interviewing (CAPI). The survey therefore has cross-sectional and longitudinal elements that, until the end of the transition period with the EU, were used to meet all the requirements of EU-SILC. Prior to 2017, EU-SILC cross-sectional indicators were compiled from the Family Resources Survey, and longitudinal indicators were compiled from the Family Resources Survey and the SLC. For the purposes of this chapter, the reader should recognise any reference to the HFS as being the survey from which EU-SILC indicators were compiled prior to the end of 2020.

The HFS forms part of the suite of social surveys run by the ONS; these are surveys that collect information from households or individuals. They are usually carried out using face-to-face interviews ([140]), whereby an interviewer visits an address sampled from the population and seeks to encourage the eligible residents to participate in the survey and then administers the survey materials using a computer.

([139]) Ria Sanderson and Peter Betts work in the Methodology Division of the UK Office for National Statistics. The authors would like to thank their colleagues in Social Survey Transformation for their help in producing this chapter. The views expressed in this chapter are those of the authors. All errors are the authors' responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Ria Sanderson (ria.sanderson@ons.gov.uk) and Peter Betts (peter.betts@ons.gov.uk).

([140]) Some telephone interviewing is used as part of the Labour Force Survey and for follow-up surveys.

Having traditionally relied on this survey-based approach to collect information on both businesses and people, the ONS has begun to transform the way it collects data through the census and data collection transformation programme. This programme seeks to make greater use of non-survey data and to reduce reliance on surveys of businesses, individuals and households (ONS, 2020a). It is also focused on integrating and harmonising surveys and moving survey data collection online. For social surveys, this is being taken forward as a programme of social survey transformation. This aims to transform the existing statistical design of social surveys, in which surveys are largely designed independently, to a design that delivers an integrated approach. The key elements of this transformation are to rationalise existing survey content, maximise the use of non-survey data and collect data using the online mode first. In this model, residual surveys will be needed to fill gaps in the data, increase coverage of under-represented groups and explore topics not available from non-survey sources. The programme is also linked to a wider government initiative to move services online, known as 'digital by default' (Cabinet Office, Government Digital Service and The Rt Hon Lord Maude of Horsham, 2012), with the aims of increasing efficiency for service providers and improving service for users.

Moving data collection for social surveys online provides both challenges and opportunities. Although question design work is in general carried out with the needs of the respondent in mind, the lack of an interviewer in the online mode highlights the importance of designing for the needs of the respondent and the mode. Work carried out by the ONS on moving content from existing face-to-face surveys to the online mode showed that data quality and the respondent experience were both negatively affected if a simple translation process was used (Wilson, 2018). The ONS has therefore been taking a respondent-centred approach to designing online surveys, with the aims of maintaining data quality and providing a positive experience for the respondent. This re-focuses the data collection activity on the needs of the respondent, rather than the data user, and is especially pertinent for the online mode, in which there is no interviewer present to motivate and provide guidance to the respondent. The approach, which is described in more detail by Wilson (2018, 2020), follows the design principles developed by the Government Digital Service (GOV.UK, 2019). This transformative approach has so far been applied to questions about the labour market, including economic activity status (which is currently measured by the Labour Force Survey (LFS)), and tests have begun on the potential for collecting data on household finances online. Financial information is viewed as particularly complex and could be especially so in the online mode with no interviewer present, so these initial tests have focused on uptake of and engagement with such a survey, rather than testing the content in detail or the wider content of EU-SILC topics.

It is becoming clear that an online-only mode is 'unlikely to meet the quality needs for national statistics' (Couper, 2019), with a consensus view forming that it is highly likely that future data collection in the United Kingdom will need to have a mixed-mode design (Wilson and Maslovskaya, 2019). This presents further challenges, particularly in the form of mode effects. These can be related to questions and measurement error (where respondents' answers to the same concept may differ between modes, due to particular features of each mode), or they can relate more to observation (the coverage, selection and non-response biases arising because of the mode). In mixed-mode data collection, this can mean that estimates from different modes can differ (see de Leeuw (2018) for a general overview of mixed-mode data collection and the impact of mode effects). There is no standard approach to adjusting for mode effects (although there are a number of techniques available – see Kolenikov and Kennedy, 2014; Klausch et al., 2017; Olson et al., 2019), and the difference between modes can be large (see, for example, Williams, 2019).

In this chapter, we aim to present the experiences to date of transforming social surveys at the ONS and to indicate the potential role of this programme in the collection of EU-SILC topics such as income, poverty and material deprivation. In the following sections, we explain the aims of social survey transformation at the ONS (Section 25.3) and the design principles that have been adopted (Section 25.4); we report on the transformative

approach to designing online data collection for the labour market, which is used as a case study (Section 25.5); and we describe the tests that have been carried out so far on the collection of data on income and financial variables (Section 25.6). This last section considers the collection of complex financial information online, and the knowledge presented is therefore directly applicable to the some of the topics in EU-SILC. We go on to consider how online data collection could be extended to EU-SILC topics (Section 25.7) and consider the possible effects on national comparability (Section 25.7.1).

25.1.1. Note on the COVID-19 pandemic

This chapter was largely written before the impact of the COVID-19 pandemic emerged. In the spring of 2020, the ONS responded to the crisis, working to ensure that the United Kingdom had the best information about society and the economy to manage its response. The need to limit social contact to keep the public and interviewers safe resulted in all face-to-face data collection being stopped, with information collected by telephone or online instead. At the time of writing, as the pandemic continues, it is unclear when face-to-face interviewing will return. The remainder of this chapter is mostly written from the pre-pandemic perspective.

25.2. Data collection practices in the United Kingdom related to EU-SILC topics

ONS social surveys are carried out using address-based sampling. The sampling frame is the postcode address file, which is a list of all the delivery points in the United Kingdom, effectively letterboxes to which mail is delivered. The HFS sample design was introduced in 2017 to select a sample for the LCF and the SLC concurrently, which were brought together to form a single HFS at that point in time. A stratified clustered design is used, in which postcode sectors are selected as clusters (primary sampling units). The selection of these primary sampling units is explicitly stratified by the Nomenclature of Territorial Units for Statistics 2 region. It is further implicitly stratified by sector-level data on the National Statistics Socio-Economic Classification of household reference people and car ownership, obtained from the 2011 census. These stratification variables are known to be correlated with household income and expenditure. The implicit stratification is achieved by sorting the clusters on the sampling frame according to these variables prior to systematic selection. The selected clusters are then randomly assigned to either the LCF or the SLC. To ensure greater coverage of postcode sectors in the HFS as a whole, there is no overlap between clusters assigned to the LCF and clusters assigned to the SLC. A random sample of addresses is selected from each cluster to form the final selected sample.

Interviewers visit the selected addresses, identify whether they are eligible for the survey and seek the cooperation of the sampled households or individuals. CAPI is used to administer the survey. Indicators on the topics relevant to EU-SILC are derived from the harmonised question content in the HFS.

25.3. Social survey transformation

The social survey transformation programme was introduced in Section 25.1. In this section, we provide further details of its envisaged future state, which will have an impact on how social surveys, including those that provide data on EU-SILC topics, are designed and carried out by the ONS. The programme is taking an iterative approach to development. The future state envisaged for ONS social surveys is to have a large 'master wave' that is used as a sampling frame for social survey content. The master wave will consist of a large sample of households, with basic information on demographics and characteristics collected about each member of the household. To meet the aims of the census and data collection transformation

programme in evaluating the effectiveness of administrative data to produce population estimates (see, for example, ONS, 2020b), it is anticipated that this master wave will also allow an assessment of the population coverage of administrative data. The proposal is for the master wave to be compiled from a foundation of non-survey data (administrative and other sources) supported by a larger survey, which is currently referred to as the Integrated Population and Characteristics Survey (IPACS). This master wave will form the sampling frame for subsequent survey instruments covering topics such as the longitudinal measurement of the labour market, household finances, health, education and opinions. Data on income, poverty and material deprivation, for example, could be collected at the master wave or in one or more of the subsequent topic surveys. The model of selecting subsamples from a master wave is well described by the work of Karlberg et al. (2015).

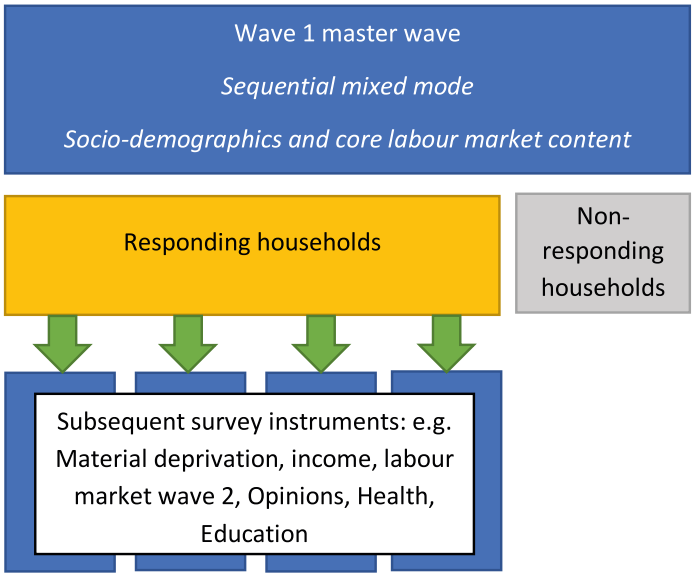
Alongside these changes to the model for data collection for social surveys, the mode of collection will become digital by default, meaning that the online mode will be offered first, with telephone or face-to-face modes used to follow up non-response, resulting in a sequential mixed-mode design.

A simplified version of this future model for social surveys is shown in Figure 25.1. It should be noted that this shows the survey component only; the expectation is that non-survey data will help to form the foundation of the master wave.

Some of the main benefits of such an integrated model are as follows.

- **Cost-efficiency.** It is expected that integrated data collection (making greater use of non-survey data, using online-first data collection and rationalising existing survey content) will reduce the overall cost of data collection.
- **Greater precision for labour market statistics (LMS).** To meet the requirements of population estimates, IPACS needs to be a large survey (approximately 500 000 households per annum); this means that labour market information will be available for a larger sample than is currently the case.

Figure 25.1: The future model envisaged for ONS social surveys



- **Reduction in respondent burden.** Both greater use of administrative data and the rationalisation of existing content provide an opportunity to reduce completion times for residual surveys by reducing questionnaire length.
- **More representative.** A more informative sampling frame, which contains some demographic characteristics, provides an opportunity for greater representativity in the sample designs or tailoring of survey content to specific groups. This does, however, rely on the master wave itself being representative of the target population.

However, the model is not without its challenges.

- **Questionnaire content needs to be transformed, not translated.** Early work on translating LFS content into the online mode found that some of the resulting questions were very difficult for respondents to understand (Wilson, 2018). This has led to the conclusion that simple translation of existing questionnaire material into the online mode will not be sufficient in this model, and a respondent-centred design approach will be needed to transform content. Without careful testing and consideration of the stimulus provided to respondents and their cognition, these changes could lead to mode effects being introduced in terms of the measurement of concepts. In addition, mode effects can arise from selection and coverage resulting from the choice or availability of modes.
- **Discontinuities.** Introducing a new model for data collection will inevitably lead to discontinuities in time series, which will need to be managed.
- **Harmonisation.** Questions and definitions will need to be harmonised across surveys, but any legislative reasons for differences will need to be respected. There are harmonised principles in place for some topics, which include definitions, survey questions, standards for administrative data, rules for presentation and guidance for users (GSS, 2020). Survey designs will also need to be considered in an integrated manner, which may require compromises for some data needs. In addition, concepts and definitions may differ between administrative or non-survey data sources and between these sources and surveys. A thorough evaluation of concepts would be required to ensure that like-for-like comparisons are made.
- **Non-response and attrition.** Like any social survey, the model is susceptible to non-response, and the two-stage nature of the model, whereby residual survey requirements are gathered by subsampling from the master wave, will also introduce attrition. Both non-response to the master wave and attrition could have a knock-on effect for the representativity of the sample. Knowledge about the respondents at the second stage will mean adjustments can be made to make the sample representative of the master wave; however, if the master wave itself is not representative of the population, this will introduce bias into estimates.
- **Interviewer-administered fieldwork.** It is possible that, in this model, fieldwork will become more challenging, as face-to-face and telephone interviewers will be asked to approach cases who are either not willing or not able to respond online. Such cases could be harder to reach on average.

25.4. Social surveys transformation: questionnaire design principles

Existing social survey questionnaires at the ONS were designed for interview modes, not for self-completion or mixed modes. Many questions have been in use for a number of years and were not always cognitively tested before being introduced. They can also rely on interviewers assisting the respondents.

Social survey questions may have been proposed by topic experts, policymakers or legislative requirements, and the content is usually driven by the needs of the data user. Questions that are proposed by stakeholders or users of the data can be cognitively burdensome and poorly understood by

the respondents (Nicolaas, 2019). This process can be improved by carrying out a respondent-centred design approach to survey development (Wilson, 2018), which involves:

- working with data users and stakeholders to understand and refine their needs;
- conducting exploratory work with the existing data set (such as frequency tables) to learn how respondents navigate through the questionnaire and identify potential bottlenecks in the flow;
- conducting exploratory work (such as focus groups, in-depth interviews and user research) to examine and understand the study population's mental models on the survey topics and concepts being addressed – sessions are also carried out with interviewers to learn from their experiences of working with the existing questionnaire;
- using these insights to develop and inform new questionnaire flow and wording, ensuring that they align with those mental models;
- conducting tests, including cognitive question testing, to explore comprehension and response processes of the redesigned questions;
- carrying out usability testing to explore the respondent's interaction with the online instrument (such as logging in, navigation, visual design, screen layouts and buttons).

This approach is a key element of social survey transformation and is described in more detail by Wilson (2018, 2020); however, it can be summarised in the following way.

- Importantly, content is not simply translated from existing interview materials. The approach taken has been to apply good design principles throughout, which has meant going back to basics and considering the question design and questionnaire flow from scratch. This requires an understanding of the data user needs, a discovery of the respondent user needs and the business requirements (for example the systems that are available or any legal obligations).
- The full end-to-end respondent journey has been transformed, developing new products that are respondent centred and apply the GOV.UK service standard ([141]). This design approach includes carrying out primary and secondary research and undertaking cognitive and usability testing combined in the same session ('cogability' testing, as termed by the researchers).
- The aim is to apply good design principles that apply regardless of mode – such as keeping questions as simple as possible – and that will provide an equivalent stimulus to respondents in each mode. The intention from the outset is not to standardise question design for all modes ([142]). Instead, a harmonised approach is taken in which the design is adapted to the mode (based on research evidence) in order to optimise the accuracy of the data collected and the respondent user experience. Through this approach, it is common to find that the design is often consistent between modes unless there is good reason to make adjustments required specifically for any mode. Cognitive testing is used in each mode; it is first used in the online mode to test cognition and usability, and then it is used in CAPI followed by CATI to confirm cognition and any design aspects specific to these modes. As the design process starts with the self-completion mode, efforts are made to ensure that the questions are understood without intervention. These then form the basis for the interviewer-administered mode and are adapted accordingly to optimise performance based on testing insights.

([141]) https://www.gov.uk/service-manual/service-standard

([142]) An example that shows that it is appropriate to optimise questioning for different modes is the count of household members. For online self-completion, a series of questions was designed to determine the eligibility of each person mentioned by the household reference person for the survey. This series methodically asks about a number of inclusion and exclusion criteria relating to temporary or longer-term periods of living at an address other than the sampled address, for example while studying at university. The online programme derives the eligibility of each person and an overall count. In testing, a single question with a set of guidance was found to be prone to error, as a result of being too complex and burdensome for respondents. In terms of interviewer administration, however, this approach did not match how interviewers work. A single question within the interview programme was found to be sufficient: interviewers were trained in the inclusion/exclusion criteria and developed experience in establishing household membership conversationally while setting up the interview and establishing rapport with respondents. A paper probing sheet was designed to assist the process, for those who were less experienced and/or for those dealing with more complex scenarios.

This approach therefore combines user research methodology with more standard data collection methodology techniques. Components of the approach have much in common with other models of best practice in questionnaire design (Geissen and Murphy, 2019; Willis, 2019).

de Leeuw (2018) and references therein summarise the 'unimode' approach to questionnaire design for mixed-mode studies, whereby the aim is to give an equivalent stimulus in each mode through the use of the same question structures and wording across modes. The approach presented above also aims to provide an equivalent stimulus in each mode. However, it recognises that there may need to be alternative wording or a different number of questions to meet the same data requirements in a different mode. It focuses on optimising for mode – presenting the same stimulus but recognising the potential need to tailor to the mode – which has been termed 'optimode' (Wilson, 2020). This approach should not introduce adverse mode effects because of the focus on consistency across modes and on obtaining a good understanding of both the data user needs and the cognition of the respondent.

In summary, it is not appropriate to copy questions from an existing mode, and it is not a straightforward exercise to translate questions from one mode to another. A transformative approach is required instead, and this needs to consider the wording of each question, the response categories, the response format (e.g. a radio button, a check box, a drop-down list, a coding frame, scales), how the question is administered, completion instructions, the guidance provided to respondents, accessibility requirements and the tool that is used to carry out the data collection (for example see Hox, de Leeuw and Klausch, 2017). This approach, which focuses on the needs of the respondent, requires a large amount of development and testing work; this may involve prioritising some areas for full transformation and perhaps taking a more light-touch approach for many questions (given the large number of questions in existing surveys). One of the additional challenges is informing data users and stakeholders of the need for transformation and the benefits it brings while highlighting that some compromises are needed in order for survey lengths to be reduced. As identified in Section 25.3, one of the key challenges with moving to a sequential mixed-mode design is the potential for differences between data collected by different modes (mode effects) and breaks in time series.

25.5. Case study: transforming labour market statistics

The social survey transformation programme began by considering how the collection of information on the labour market could be transformed. Official statistics on employment and unemployment are key economic indicators that are currently based on the LFS. This is the largest ONS social survey, with a sample comprising approximately 40 000 households across the United Kingdom every quarter (ONS, 2015). We focus on this case study, as it tested the implementation of the respondent-centred design approach in an online social survey questionnaire, and therefore starts to explore the potential of the integrated approach to social surveys described in Section 25.3, providing useful insights into how this model could be applied to the topics relevant to EU-SILC.

The aim of the transformation activity was to understand how a respondent-centred design approach could be used to collect labour market concepts in an online mode. An iterative approach to developing and testing online labour market content took place over a number of years, with quantitative tests carried out over the past 2 years. This involved a series of tests to establish the optimum design for future LMS. A prototype LMS questionnaire was developed, consisting of a core set of questions from the LFS that had been transformed using the approach described in Section 25.4. The ONS then commissioned a series of response rate tests to understand what uptake ([143]) and completion rates looked like for this transformed survey, which was designed to be optimised for the online mode.

([143]) Uptake is defined as the proportion of households that provide any number of data in the online collection tool, and thus includes partial completion as well as full completion. Ineligible households are excluded.

In the first test, a sample of 37 800 households across Great Britain were invited to take part in an online-only LMS questionnaire (Ipsos Mori, 2018a). Split-sample experiments were used to understand the effect of different survey materials and survey conditions (e.g. day of dispatch and timing of reminder letters). This test found that the most effective strategy was to send out an invite letter followed by two reminders issued more than 3 days apart (Ipsos Mori, 2018a). Unfortunately, this option is operationally challenging, so the next best approach identified by the test was implemented (this used a single reminder letter).

The second test focused on the effect of different incentive strategies on the uptake rate. A sample of 40 000 households was invited to take part in an online-only study. This test found that an unconditional GBP 5 voucher with a further GBP 10 voucher being given to households that complete the survey was the most effective strategy; however, alternative incentives such as a tote bag or an unconditional GBP 5 incentive were also found to be more effective than no incentive (Ipsos Mori, 2018b). On the basis of these results, a decision was made to use the non-monetary incentive of the tote bag, which proved to be effective in terms of uptake and cost.

A further test was carried out in 2018 to test an integrated labour market and population coverage survey (closer to the proposed IPACS than simply LMS alone). This was a mixed-mode test, in which the online mode was offered first followed by face-to-face follow-up. A sample of 14 419 households across Great Britain were invited to take part, with the aim of testing the mixed-mode uptake rate, the characteristics of the responding sample and the potential for non-response bias.

Data analysis compared statistics from the LMS test with the LFS at the national level. The results (ONS, 2020c ([144])) showed that there were no statistically significant differences for any of the main headline labour market estimates (employment, unemployment and economic inactivity). However, comparisons revealed a statistically significant difference in estimates for people aged 16–24 years in full-time education. These single-point-in-time differences from a relatively small test need further investigation in the context of the changes made to the prototype LMS questionnaire. Specifically, the questionnaire design and the approach to sampling will be further investigated.

([144]) Disclaimer: These research outputs are not official statistics relating to the UK labour market. Rather, they are published as outputs from research into an alternative prototype survey instrument (the Labour Market Survey) to that currently used in the production of LMS (the LFS). It is important that the information and research presented here is read alongside the accompanying technical report to aid interpretation and to avoid misunderstanding.

25.6. Application to the Household Finance Survey (including EU-SILC topics)

As described in Section 25.1, a single HFS was introduced in 2017 in order to start to harmonise the collection of information on household income and expenditure. As well as considering the potential further harmonisation of the questionnaire, the scope for the use of administrative data will be reviewed, to see whether questions can be replaced should the same information be available from an administrative data source. This will need careful evaluation.

In parallel with this wider harmonisation, consideration is also being given to how to transform the collection of financial information from individuals by moving this data collection online or into a sequential mixed-mode design. This longer-term goal relies on first testing the feasibility of collecting more complex concepts in an online mode. This work is at an early stage, but one of the early research questions identified is whether respondents would engage with an online survey on finances. The complex nature of financial concepts, the potential need to refer to financial documents and interest in the subject matter are all possible influencing factors for an individual choosing whether or not to respond. This is perhaps some of the most complex information collected within the EU-SILC framework, and testing individuals' willingness to respond online to this topic is crucial to understanding how applicable such an approach is in an environment in which an interviewer is not present to provide guidance and reassurance.

To test whether respondents will engage with a survey on finances, an 'uptake' test was designed with two research questions in mind.

- What proportion of households will participate in an online survey on the subject of household finances?
- What is the impact of telling respondents in advance that they will need to refer to financial documents?

The uptake test was therefore designed to test overall engagement with a survey asking for financial information, but also how response rates are affected by making sampled households aware in advance that they might need to consult financial documents.

The uptake test was carried out in 2019 and ran for a period of 4 weeks. A total sample of 16 320 addresses was used to carry out a split-sample experiment, in which 50 % of the addresses received advance materials that were generic in nature, and 50 % received tailored advance materials specific to the survey (mentioning some topics and the potential benefit to the respondent of having information from financial documents / online banking to hand). In line with the findings from the case study on LMS described in Section 25.5, a tote bag was offered as an unconditional incentive. The data collection tool was developed and hosted by Ipsos Mori.

Given that the focus of the test was on measuring uptake, the content of the questionnaire was not fully tested. If questions could be used from the LMS work presented above, they were. Other questions underwent expert review, which resulted in some questions being used without any changes being made, and some questions being broken down into subquestions or reworded. The aim here was to measure engagement with the survey, rather than the success of the survey materials themselves. The test asked one person in the household to complete information about who lived in the household (this was the first person who logged on to the online survey using the unique household access code). Individuals in the household could then select their name and answer questions relating to themselves, which were not visible to other members of the household (ONS, 2020d). A reminder letter was sent to the household a week after the survey went live.

The overall engagement rate (those who started the survey) was 20.7 %, with a 16.5 % completion rate, suggesting that households will engage with a request to take part in an online household finances survey (ONS, 2020d). Rates differed by treatment, with the group receiving generic materials having higher rates (engagement: 24.5 %; completion: 19 %) than the group receiving tailored materials (16.9 %; 14 %). Although the response rate of this survey was lower than the potential LMS response rate of almost 30 %, the LMS survey had been through more extensive research and development. The survey report also includes an analysis of some data quality metrics and completion times.

- Respondents who were given advance notice of the topics and documentation needed to complete the survey provided better quality data; for example, more of these respondents were able to answer the financial questions, and they provided more precise figures.
- Median response completion times were very similar, at approximately 25 minutes per household (7 minutes per individual questionnaire).
- Analysis of rates of proxy completion showed no difference by treatment group and an overall rate (18 %) similar to rates in comparable surveys.
- Regarding item non-response, the group receiving generic materials was less likely to provide complete responses to the survey sections than the group receiving tailored materials.
- The unit non-response rate (when an individual within a responding household refuses or fails to complete any part of the individual survey) was lower among the group receiving tailored materials than among the group receiving generic materials (7.2 % compared with 10.7 %).
- The group receiving tailored materials was more likely to provide permission to be recontacted (77.7 % compared with 72.9 %) and to provide a phone number and/or email address.

In addition, in its report on the survey, the ONS (2020d) found that the profile of the responding

households was in line with those in similar surveys, such as the LMS survey and the LCF.

- The survey had similar gender profiles for respondents to those in other similar surveys and population profiles.
- The age profile of those in the responding households was similar to the age profiles in similar surveys, such as the LMS questionnaire. However, under-representation of those aged 16–24 years and over-representation of those aged 65 years and over were greater in the HFS test.
- Consistent with other ONS surveys, non-white respondents were slightly under-represented.

The report recommends further research to look at the feasibility of collecting extensive financial data online and the trade-off between response rate and data quality. This general issue will be important for the move to the online mode. There are potentially two concerns here: (i) the representativity of the responding sample (whether certain groups are over- or under-represented, for example) and (ii) whether the quality of individual responses is comparable to the quality of individual responses in alternative modes. In this context, comparisons between the data collected in the online mode and the data collected using other sources (potentially at both the micro and the macro level) would be valuable in understanding whether lower response rates have an impact on data quality. This could be achieved by comparing micro-level survey responses and aggregated estimates with alternative sources (registers, administrative data, other surveys and census records, for example).

25.7. How online data collection could be extended to EU-SILC topics

Comparability of data between EU Member States is a key aim of the EU-SILC framework (European Commission, 2003), and as a result there is commonality in definitions, concepts and guidelines for the variables that must be collected and supplied to Eurostat. However, Member States may choose the most efficient implementation at national level to meet the data requirements (Eurostat, 2020a). This means that, in practice, the data collection activities within Member States will vary, but they are all designed to meet the common conceptual framework for the calculation of the EU indicators. For example, the mode of collection differs across Member States (Eurostat, 2020b; see also Chapter 24 of this book).

The UK experience of transforming social surveys has provided useful insights into the use of the online mode, including uptake, incentivisation and engagement in a complex subject matter. However, challenges remain around the transition to the online mode. Nicolaas (2019) characterises these as:

- the need to reduce questionnaire length;
- the need to redesign questions that are currently administered by interviewers;
- managing of the risk to the time series.

These are perhaps the key challenges that need to be overcome in order to adopt online data collection for EU indicators, whether this is in a single country or across all Member States. In the United Kingdom, the respondent-centred design approach (Wilson, 2018) to transforming survey content on the labour market that has been used in the social survey transformation programme has directly addressed these challenges, and we anticipate that this could be a useful approach more generally for Member States.

The growing consensus seems to be that mixed-mode collection is the way forward for official statistics (e.g. Couper, 2019), as the low uptake rates of online-only surveys risk introducing bias, even with a large sample size. If collection were to be carried out using a mixed-mode design, there would be a further challenge from the possibility of mode effects being introduced in a country's data, and mode effects being introduced between countries if different countries changed mode at different times. These mode effects can arise from differences in the responses given by respondents depending on the mode (measurement effects, the main concern for mixed-mode surveys) and in whether respondents are able to, or willing to, complete surveys online (coverage and selection effects). A mixed-mode approach is potentially advantageous

in addressing coverage and selection effects, as it offers alternative modes to the respondent. The ONS 'optimode' approach to designing questions for mixed-mode surveys will potentially minimise mode-related measurement effects. In contrast, single-mode surveys will not exhibit measurement effects arising from a choice of mode (although other measurement effects may still exist). Comparisons of countries could be confounded by mode effects if countries use different modes or different mixes of modes. The willingness to respond to an online survey (whether it is a single-mode survey or a mixed-mode survey) may well differ between countries because of different cultures and social norms.

The possibility of EU-SILC moving to an online-first mixed-mode approach also presents opportunities. For example, for a survey organisation, moving away from face-to-face data collection could potentially reduce fieldwork costs and has the potential to make data collection more efficient. For data users, the approach has the potential to improve data quality, both for observation and for measurement. However, there is a financial overhead in developing an online survey, and the benefits would need to be weighed against the costs. The UK situation is perhaps unique, as the move to a sequential mixed-mode design is embedded in a wider transformation of the model for producing statistics on society, meaning that wider benefits are also being realised. In addition, the wider government initiative to become digital by default in the administration of public services has meant that design principles (described in Sections 25.3 and 25.4) have been developed that give a robust framework for the use of the respondent-centred design approach.

Building on the knowledge acquired from the UK experience, we surmise that, to make any progress with an online data collection model for EU-SILC, the following needs to be achieved.

- **EU-SILC content needs to be designed for use in both online and interviewer-administered modes.** To capitalise on the digital by default approach, questions need to be designed to perform equally well in both online and interviewer-administered modes. The findings from the uptake test reported in Section 25.6 provide insights into how willing respondents are to engage with a survey on potentially complex financial topics in an online mode. Designing the online questionnaire would be a lengthy process given the complexity of surveys on household income. One of the challenges of moving such content online would be to design the content in a way that maximises the information that can be collected while minimising the burden placed on the respondent. The respondent-centred approach adopted by ONS would be helpful here; such a re-design would have the potential to overcome the challenges of questionnaire length and the re-design of questions, and would provide the opportunity to test different levels of prompting, devices, or the need for respondents to refer to financial information, for example. The mode on offer can also influence respondents' decision to participate in surveys (Collins and Mitchell, 2014), and the potential for this to impact on response rates and to introduce non-response bias also needs also to be considered.
- **A mixed-mode approach needs to be evaluated.** The gains in overall response rate for the LMS tests came once alternative modes of completion were offered to respondents who initially did not respond to the online survey (the sequential mixed-mode approach). Indeed, the consensus in the United Kingdom is that the future of survey data collection is the mixed-mode approach (Wilson and Maslovskaya, 2019); however, the effectiveness of mixed-mode designs for both cross-sectional surveys and the first wave of longitudinal surveys is not well tested (Couper, 2019). Research is also needed into how financial topics can be presented to respondents in an online mode and how accurate factual information can be collected for potentially complex topics. Consideration also needs to be given to the longitudinal elements of the collection required for EU-SILC topics. Much work was done during the LMS tests to develop the materials sent to sampled households in advance, but this will need to be considered again in order to see how to maximise response to a financial survey in which people

may feel less comfortable with the topic or less confident in providing their financial details online. Indeed, there some is evidence that respondents may be more open about sharing financial information in later waves of a longitudinal survey (Fisher, 2019). Mixed-mode approaches can also introduce mode effects, which can be unpredictable and not easy to explain (Couper, 2019). Aiming to present the same or equivalent stimulus for each mode and validating the effectiveness of this through qualitative and quantitative testing will help reduce some of this unpredictability.

- **The impact on data quality needs to be examined.** The design of the LMS tests enabled comparisons to be made between LFS estimates and LMS estimates, the main results of which are reported above. However, the focus of the HFS test was on uptake. Although some aspects of the quality of the data were measured, this would be an important avenue for further research. It is often face-to-face interviews that are considered the 'gold standard' (although even these are susceptible to quality issues), so comparing data quality with some benchmark, be it survey or administrative data, would be valuable. Thorough evaluation and understanding of the impact of changes are also necessary to allow discontinuities in time series to be managed.

25.7.1. Effects on national comparability

We have focused on presenting the work that has been done by the ONS to start transforming social surveys. We now consider the possible effects on data quality and national comparability as Member States transition to online-first modes of data collection.

The main impacts of the sorts of transformational changes we have discussed in this chapter are on existing time series. It is often desirable, from a user perspective, to have a comparable time series, but the introduction of transformational changes risks this comparability over time. Therefore, the treatment of discontinuities is a very important issue to consider in advance of any major changes being implemented (see, for example, van den Brakel et al., 2021). It is possible that, if countries move to the online mode or mixed-mode approaches at different times, this could lead to a series of discontinuities in the time series of EU-SILC. There is not an easy way to manage a discontinuity resulting from a change in mode; either the change is accepted or some form of parallel approach needs to take place (Nicolaas, 2019). The latter approach is one in which both data on the new basis and data on the old basis are collected, and these are used to form an adjustment to the time series (see, for example, van den Brakel et al., 2021).

Although the overall proportion of EU households with access to the internet has increased rapidly, reaching 89 % in 2018, there are still large differences between countries and within countries based on urban–rural classifications (Eurostat, 2019). This means that an approach that works well in one country may not work well in another, and the implementation of the online mode or mixed-mode approaches will inevitably reflect the environment within each Member State. If Member States move to online data collection at different times and in different ways, this could lead to further discontinuities of the type identified above and affect national comparability.

25.8. Conclusions

In this chapter, we have considered the approach used to collect EU-SILC variables in the United Kingdom prior to the United Kingdom leaving the European Statistical System (as described in the introduction); we have described the wider programme of social survey transformation and early results from this programme for the labour market and household finances; and we have discussed the applicability of these approaches to the collection of EU-SILC indicators across Member States. We summarise below the approach and findings, what we have learned from them, and the possible wider application in the compilation of EU-SILC indicators.

Within the United Kingdom, data on the topics relevant to EU-SILC are collected from the HFS. These data are collected using an address-based survey,

in which sampled households are approached by interviewers, who work to secure a response and administer the survey in respondents' homes.

In parallel, the ONS has implemented a census and data collection transformation programme, which seeks to make greater use of non-survey data and to reduce reliance on surveys of businesses, individuals and households (ONS, 2020a). This programme includes integrating and harmonising surveys and moving survey content online. For social surveys, this work is being taken forward as a programme of social survey transformation, which has developed a vision for the future state of ONS social surveys in which a large 'master wave' is used as a sampling frame for residual survey components that cannot be collected from non-survey data. The aim of this programme is also to make data collection online first, and a transformative approach has been taken so far that puts the respondent at the centre of the design of the data collection activity in an effort to overcome the challenges of moving to an online mode. This respondent-centred design (Wilson, 2018) aligns with a wider government initiative for public services to become digital by default and for designs to have the user of a service in mind. This approach has been applied to the labour market and has generated a number of useful insights into how to transform social surveys.

In terms of understanding the application of this model to the possible collection of data on the topics covered by EU-SILC, work to date has focused on household finances only, as this is a potentially complex area for respondents. One early concern with collecting such information online is whether respondents would engage with the content and be willing, and able, to provide the required factual information. The HFS uptake test examined how willing people are to participate in an online survey on household finances and whether the provision of different advance materials (relating to the need to refer to financial documents) has an impact on engagement. It provided promising results and useful insights into the ability to collect financial information online. Considering the applicability of such a model to collecting data on the topics relevant to EU-SILC, we have identified a number of challenges that still need to be addressed. These involve designing content for the online mode, evaluating a mixed-mode approach to data collection and considering the impact on data quality.

The main risk to implementing an online mode across EU-SILC is the potential for discontinuities in time series, both within countries when they adopt a change in mode and across countries if they move to different modes at different times. There would therefore need to be some consideration of the approach to managing discontinuities and a choice made as to whether these should simply be accepted or adjusted for in some way.

Finally, the COVID-19 pandemic has necessitated a change in how data are collected for surveys of society and the economy. In the spring of 2020, the ONS responded to the crisis, working to ensure that the United Kingdom had the best information about society and the economy to manage its response. All face-to-face data collection stopped, with information collected by telephone or online instead. At the same time, a need for information on the impact of the pandemic on society emerged, driving the development of surveys to understand infection rates and the social impact of COVID-19 (ONS, 2020e). The programme of social survey transformation emerged as an enabler here, with the LMS questionnaire being launched in 'beta' form and information on the social impact of COVID-19 being collected predominantly in the online mode. A subsampling approach also started to be used in much the same way as envisaged in Section 25.3.

The long-term implications of the COVID-19 pandemic for data collection for social surveys in the United Kingdom and across the world are unclear. Countries have needed to adapt quickly, meaning that there have been changes in collection modes, questionnaire content and the frequency and timing of data collection. Consequently, discontinuities may have been introduced in the data because of the necessary changes to methods, at a time when the concepts being measured were themselves also likely to change. This period of change and the lessons learned from making adaptations to existing approaches also potentially provide a unique opportunity to transform data collection designs in the longer term.

References

Cabinet Office, Government Digital Service and The Rt Hon Lord Maude of Horsham (2012), 'GOV.UK: making public service delivery digital by default' (https://www.gov.uk/government/news/launch-of-gov-uk-a-key-milestone-in-making-public-service-delivery-digital-by-default).

Collins, D. and Mitchell M. (2014), 'Role of mode in respondents' decisions to participate in IP5: findings from a qualitative follow-up study', *Understanding Society Working Paper Series*, No 2014-03 (https://www.understandingsociety.ac.uk/research/publications/522579).

Couper, M. (2019), 'Online surveys: opportunities and challenges', paper presented at the Conference on the Future of Online Data Collection in Social Surveys, University of Southampton (https://www.ncrm.ac.uk/research/datacollection/Couper%20SDAI%202019-06-20.pdf).

de Leeuw E. (2018), 'Mixed mode: past, present and future', *Survey Research Methods*, Vol. 12, No 2, pp. 75–89.

European Commission (2003), Regulation (EC) No 1177/2003 of the European Parliament and of the Council of 16 June 2003 concerning Community statistics on income and living conditions (EU-SILC), OJ L 165, 3.7.2003, p. 1 (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32003R1177&from=EN).

Eurostat (2019), 'Digital economy and society statistics – households and individuals' (https://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_households_and_individuals#Internet_access).

Eurostat (2020a), 'Income and living conditions – methodology' (https://ec.europa.eu/eurostat/web/income-and-living-conditions/methodology).

Eurostat (2020b), 'EU statistics on income and living conditions (EU-SILC) methodology' (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology).

Fisher P., (2019), 'Does repeated measurement improve income data quality?', *Oxford Bulletin of Economics and Statistics*, Vol. 81, No 5, pp. 989–1011.

Geissen, E. and Murphy, J. (2019), 'A compendium of web and mobile survey pretesting methods', in Beatty, P., Collins, D., Kaye, L., Padilla, J. L., Willis, G. and Wilmot, A. (eds), *Advances in Questionnaire Design, Development, Evaluation and Testing*, Wiley, Hoboken, NJ, pp. 287–384.

GOV.UK (2019), 'Government design principles' (https://www.gov.uk/guidance/government-design-principles).

GSS (Government Statistical Service) (2020), 'Harmonised principles by topic' (https://gss.civilservice.gov.uk/guidances/0-harmonised-principles).

Hox, J., de Leeuw, E. and Klausch, T. (2017) 'Mixed-mode research: issues in design and analysis', in Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E. et al. (eds), *Total Survey Error in Practice*, Wiley, Hoboken, NJ, pp. 511–530.

Ipsos Mori (2018a), *Labour Market Survey – Response rate experiments – Report for test 1: Materials experiment* (https://digitalblog.ons.gov.uk/wp-content/uploads/sites/9/2018/04/Test-1_Full-report_FINAL-for-publishing.docx).

Ipsos Mori (2018b), *Labour Market Survey – Response rate experiments – Report for test 2, tranche 1: Incentives experiment* (https://digitalblog.ons.gov.uk/wp-content/uploads/sites/9/2018/04/Test-2-Tranche-1-report-FINAL-for-publishing.docx).

Karlberg M., Reis F., Calizzani C. and Gras F. (2015), 'A toolbox for a modular design and pooled analysis of sample survey programmes', *Statistical Journal of the IAOS*, Vol. 31, pp. 447–462.

Klausch, T. Schouten, B., Buelens, B. and Van den Brakel, J. (2017), 'Adjusting measurement bias in sequential mixed-mode surveys using re-interview data', *Journal of Survey Statistics and Methodology*, Vol. 5, No 4, pp. 409–432.

Kolenikov, S. and Kennedy, C. (2014), 'Evaluating three approaches to statistically adjust for mode effects', *Journal of Survey Statistics and Methodology*, Vol. 2, No 2, pp. 126–158.

Nicolaas, G. (2019), 'Online data collection in social surveys: back to basics', paper presented at the Conference on the Future of Online Data Collection in Social Surveys, University of Southampton (https://www.ncrm.ac.uk/research/datacollection/Nicolaas_Soton%20ONS%20keynote%20June%202019%20FINAL.pdf).

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S. et al. (2019), *Transitions from telephone surveys to self-administered and mixed-mode surveys* (https://www.aapor.org/Education-Resources/Reports/Transitions-from-Telephone-Surveys-to-Self-Adminis.aspx).

ONS (Office for National Statistics) (2015), 'Labour Force Survey (LFS) QMI' (https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveylfsqmi).

ONS (2020a), 'Data collection transformation' (https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/datacollectiontransformationprogrammedctp).

ONS (2020b), 'Administrative data census project' (https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject).

ONS (2020c), *Labour Market Survey: Comparative estimates report* (https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourmarketsurveycomparativeestimatesreport).

ONS (2020d), *Moving Household Financial Surveys Online: Initial research findings* (https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/articles/onlineuptaketestofahouseholdfinancialsurvey/2019).

ONS (2020e), 'Ensuring the best possible information during COVID-19 through safe data collection' (https://www.ons.gov.uk/news/statementsandletters/ensuringthebestpossibleinformationduringcovid19throughsafedatacollection).

van den Brakel, J.A., Smith, P. A., Elliott, D., Krieg, S., Schmid, T. and Tzavidis, N. (2021), 'Assessing discontinuities and rotation group bias in rotating panel designs', in Lynn, P. (ed.), *Advances in Longitudinal Survey Methodology*, Wiley, Chichester, pp. 399-423.

Williams, J. (2019), 'Online data collection within a mixed mode design: learning from the Our Future longitudinal survey of young people', presented at 'The future of online data collection in social surveys', 20 June (https://www.ncrm.ac.uk/research/datacollection/Williams_Kantar%20presentation_final.pdf).

Willis, G. (2019), 'Questionnaire design, development, evaluation, and testing: where are we, and where are we headed?', in Beatty, P., Collins, D., Kaye, L., Padilla, J. L., Willis, G. and Wilmot A. (eds), *Advances in Questionnaire Design, Development, Evaluation and Testing*, Wiley, Hoboken, NJ, pp. 1–23.

Wilson, L. (2018), 'Using respondent centric design to transform social surveys at ONS', *Survey Methodology Bulletin,* Vol. 78, p. 45. (https://www.ons.gov.uk/file?uri=/methodology/methodologicalpublications/generalmethodology/surveymethodologybulletin/surveymethodologybulletinno.78january2018.pdf).

Wilson, L. (2020), 'User centred design approach to surveys', Government Statistical Service, 4 November (https://gss.civilservice.gov.uk/policy-store/a-user-centred-design-approach-to-surveys/).

Wilson, L. and Maslovskaya O. (2019), *Report: A summary of the agreed challenges, opportunities and best practice for online data collection in the UK in 2019 and the future* (https://www.ncrm.ac.uk/research/datacollection/ONS_Soton_Report_The%20future%20of%20online%20data%20colleciton%20conference.pdf).

26 A cost–benefit analysis of EU-SILC mode effect decomposition: a Dutch case study

Barry Schouten ([146])

26.1. Introduction

This chapter explores the utility of mode effect decomposition for European Union Statistics on Income and Living Conditions (EU-SILC) surveys that use a sequential mixed-mode design. It explores whether investment in a reinterview design to produce estimates of mode-specific selection and mode-specific measurement biases is beneficial from an accuracy point of view. More specifically, it evaluates whether a potential gain in bias may outweigh a potential loss in variance, because some of the survey budget is used for the reinterview.

Mode effects originate from different modes attracting different respondents and eliciting different answers from the same respondents. Mode effects can be viewed as the result of mode-specific selection and mode-specific measurement. As measurement follows selection, the two effects are confounded and cannot easily be separated. One may, furthermore, argue that the two types of effects are associated, as respondents may opt for the mode in which they feel most comfortable. This observation then leads to the notion of so-called potential outcomes: if a respondent had participated in another mode, would this respondent have given different answers? This question becomes very real when a survey migrates to a design with new modes. For EU-SILC, this question is also very relevant as some European Statistical System national statistical institutes employ multiple modes or are considering doing so.

The causes of such mode effects have been studied extensively and are discussed elsewhere (see Dillman, Smyth and Christian (2014) for a general overview). These causes are not elaborated here. This chapter takes the standpoint that mode effects may exist in EU-SILC and that one wishes to know the impact on comparability over time, comparability between relevant subpopulations, for example income classes, and comparability between countries. Table 26.1 presents respondent mean responses for three variables in the Dutch EU-SILC for web mode and telephone mode. Telephone mode is implemented as a sequential mode to web mode in the Dutch EU-SILC. The question that arises from this table is whether telephone mode attracts different respondents and/or whether telephone respondents give different answers.

Within the ESSnet Mixed Mode Designs for Social Surveys project, a cost–benefit analysis was performed for a reinterview design using the European Health Interview Survey and the Labour Force Survey as case studies. The results are reported in Schouten et al. (2019). They conclude that for the European Health Interview Survey a reinterview design may be profitable from an accuracy point of view, while it is not likely to be beneficial for the Labour Force Survey. In this chapter, we perform the same cost–benefit analysis, but for EU-SILC.

Reinterview designs correspond to repeated measurements on a subset of the respondents (see, for example, Biemer (2001), Schouten et al. (2013), Klausch et al. (2017)). Reinterview designs are different from full crossover designs in that only one or

([145]) Barry Schouten is employed by Statistics Netherlands and is Professor by Special Appointment at Utrecht University. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Barry Schouten (bstn@cbs.nl).

Table 26.1: Respondent mean responses for three key survey variables in the 2019 Dutch EU-SILC for web mode and telephone mode

Variable	Web (%)	Telephone (%)
Self-perceived health is (very) good	72.3	68.0
Make ends meets fairly easily to very easily	90.1	88.0
Socioeconomic status		
Employed	48.5	37.9
Unemployed	2.8	1.3
Volunteer	2.3	5.0
Disabled	4.8	5.0
Student	9.2	5.9
Household	3.7	12.1
Retired	23.1	30.0
Other status	5.7	2.7

a few of the possible mode sequences are implemented. In full crossover designs respondents are invited for all modes in a randomised order. Reinterview designs attempt to provide a non-experimental feel to respondents and avoid additional effects due to experimentation. Nonetheless, reinterview designs also make assumptions, which can be strong for certain surveys and/or settings. These assumptions will be made explicit in this chapter. A broader perspective of mode effect estimation is provided by Buelens, van den Brakel and Schouten (2019).

In Section 26.2 we describe the methodology for decomposing mode effects. In Section 26.3, we provide the results for the Dutch EU-SILC. Conclusions are provided in Section 26.4.

26.2. Mode effect decompositions using a reinterview design

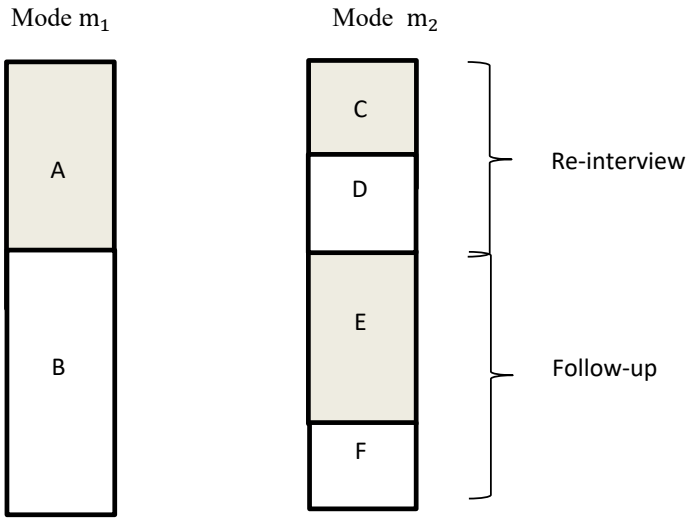
This section provides a basic description of reinterview design and estimation. It then explicitly mention the assumptions that are made. It ends by translating the methodology to a cost–benefit analysis from two viewpoints: design and adjustment. See Klausch et al. (2017) and Schouten et al. (2019) for a detailed description of the methodology adopted here.

26.2.1. Methodology

For simplicity it is assumed that a mixed-mode survey design has two modes, m_1 and m_2. Klausch et al. (2017) also describe how a reinterview design can be implemented for a design with three modes. We assume that the modes are sequential: non-respondents to m_1 are invited to participate in m_2. Concurrent mixed-mode survey designs, in which households can choose between two or more modes, are not considered. Concurrent designs often consist of web mode versus paper mode, or different online devices. For such designs it is harder to enforce a reinterview in a different mode without making it an experimental study. A mixed-mode design in which in the invitation letter for a web/paper questionnaire it is mentioned that an interviewer will contact the household if they do not want to participate through a web/paper questionnaire may be considered a hybrid form of a concurrent and sequential design. Such a design would still fit within the reinterview strategy.

Figure 26.1 shows the missing data pattern of a two-mode sequential mixed-mode design with reinterview. Respondents to mode m_1 (area A) are invited for a reinterview using mode m_2, which leads to a response (area C) and a non-response (area D). Mode m_1 non-respondents (area B) receive a default follow-up interview in mode m_2, leading to a response (area E). The non-respondents to both modes are outside the scope of this design, as they are not responsible for mode effects.

Figure 26.1: Reinterview design for a $m_1 \rightarrow m_2$ sequential survey design



NB: Grey areas represent m_1 response (A), reinterview m_2 response (C) and follow-up m_2 response (E). White areas represent m_1 non-response (B), reinterview m_2 non-response (D) and follow-up m_2 non-response (F).

Crucial choices in a reinterview design are the benchmarks for selection and for measurement. We assume that response to the sequential mixed-mode design (areas A + E) is the benchmark, that is, the representation for a single mode m_1 design is weaker than that for the sequential design. For the measurement benchmark, we can choose between m_1 and m_2. We leave the benchmark choice open and evaluate both options.

There are two strategies for separating and estimating mode-specific measurement and selection biases. One strategy is to directly model the difference between answers in areas A and C, that is, treat them as answers to the same question. Another strategy is to calibrate m_2 answers in area C to areas C + D using m_1 answers in area A as weighting variables, or calibrate m_1 answers in areas C + B to areas C + E using m_2 answers as weighting variables. The first strategy is adopted in Klausch et al. (2017) and the second in Schouten et al. (2013). We adopt the first strategy here.

26.2.2. Assumptions

It is important to stress that reinterview-based estimation of mode-specific measurement biases has associated assumptions. In order to be effective, reinterview designs require three assumptions.

1. Reinterview measurement behaviour is not affected by the first interview.
2. Reinterview non-response preserves the relative measurement errors between the modes.
3. True values of the survey variables of interest have not changed between the first interview and the reinterview.

Assumptions 1 and 2 may be combined into one assumption: The true relative measurement errors between the modes for the potential outcomes without reinterview hold for the reinterview sample. In more simple terms, it is assumed that the measurement error model that is posed is unaffected by the reinterview.

Assumptions often do not hold, in general, and the question is how far the assumptions are from true behaviour. Although the reinterview presentation and timing may be carefully designed and implemented, it is likely that at least some impact of the first interview will remain. This is not to say that a respondent is supposed to have lost all recall of the first interview – that the respondent must not be affected by it. It means that, for some causes

of measurement error, reinterviews may not work. Consider, for example, a question that requires cognitive effort, such as the number of hours that a respondent sits down for in an average day. The respondent might recall their answer after a month has passed and repeat it, although they might take little time to calculate it and/or read through the question introduction text. However, reinterviews may also work under recall. A respondent may have answered honestly to a sensitive question in self-reporting, but may opt not to do so to an interviewer.

The third assumption can and will be relaxed here. Time change will, in general, lower the association/correlation between the interview and reinterview measurements. This loss in reliability is accounted for in the analysis by introducing a predicted correlation between the two measurements as a function of time passed. The lower the anticipated reliability, the larger the reinterview sample needs to be to arrive at the same accuracy of bias estimation.

The assumptions we make should, however, also generally be compared against assumptions made by alternative mode effect estimation and adjustment techniques, in particular covariate-based adjustment such as weighting and regression estimation. These techniques assume that the available covariates explain the selection effect between modes. In our view, these data are often too weak to plausibly make these assumptions and the reinterview design is then a useful alternative.

For a description of how to time a reinterview and how to present a reinterview to respondents, see Schouten et al. (2013).

26.2.3. A cost–benefit analysis

In order to consider a reinterview as a means to inform the actual design or adjustment of mixed-mode surveys, more is needed than a reinterview estimation strategy.

First, the reinterview implies extra costs that need to be funded. For academic purposes, such a design may be funded from an external budget. For practical purposes, it typically needs to be funded from the survey budget itself. This means that the overall sample size needs to be reduced to free just enough of the budget to finance a reinterview. If T represents the length of time that the survey design is kept constant after the reinterview has been applied and B represents the survey budget per round, then TB budget is available. If the reinterview costs are B_{RE} and are spent in round 1, then $(TB - B_{RE})/(T-1)$ is available for the remaining rounds. In the EU-SILC study, we consider $T = 2, 5, 10$ years, that is, we do a reinterview in year 1 that needs to be funded from the budget for 2 years, 5 years or 10 years of the survey, respectively.

Second, a choice must be made between the design perspective and the adjustment perspective. The design perspective corresponds to choosing a future design that is most accurate given the available budget. In the Dutch EU-SILC case, the choice then is between the single mode m_1 and the sequential design with mode m_1 followed by mode m_2. The adjustment perspective means that the estimate of the mode-specific measurement bias is employed to adjust future estimates, that is, future EU-SILC estimates are adjusted towards the measurement benchmark mode. When mode m_1 is considered the measurement benchmark, then m_2 answers are adjusted. Similarly, when mode m_2 is viewed as the measurement benchmark, m_1 answers are adjusted. Under the adjustment perspective there are two more alternatives: the adjusted single mode m_1 design and the adjusted sequential mode design with m_1 and m_2. However, the adjusted single mode m_1 design is not evaluated here. When the adjusted single mode m_1 design is the measurement benchmark, then this design is the same as the unadjusted single mode m_1 design. When m_2 is the measurement benchmark, then the single mode m_1 design suffers from both a selection and a measurement bias. Unless they cancel each other exactly, this design has inferior accuracy. In the results, therefore, three estimators are considered: the unadjusted single mode m_1, the unadjusted sequential mode and the adjusted sequential mode. These are labelled as $\hat{Y}_{m_1}$, $\hat{Y}_{m_1 \to m_2}$ and $\hat{Y}^{\text{adj}}_{m_1 \to m_2}$. It is important to note that the adjusted sequential mode design will have superior bias properties, as it is adjusted towards the measurement benchmark, but that it will have larger variances. The properties of the unadjusted single mode m_1 and unadjusted sequential mode are evaluated as if there had not been a reinterview.

Third, it is clear that a reinterview is beneficial only under specific magnitudes of the mode-specific measurement bias. If mode-specific measurement bias is absent, then modes m_1 and m_2 provide the same answers and the sequential mode design is always superior, as it is the selection benchmark. If mode-specific measurement bias, on the contrary, is very large, then a reinterview is almost always beneficial, unless the measurement bias is completely cancelled out by mode-specific selection bias. The consequence is that one would not blindly carry out a reinterview without deriving first where the turning point is, that is, under which scenarios the reinterview can be beneficial. This is also the purpose of the cost–benefit analysis: vary mode-specific measurement bias levels over a plausible range and determine if and how the reinterview would beat both the unadjusted single mode and the unadjusted sequential mode design. In Schouten et al. (2019) a detailed description is provided of how such a plausible range can be constructed based on expert knowledge. For ease of display, we consider three levels in this chapter: the left and right extremes of the mode-specific measurement bias interval and the midpoint of the interval. For EU-SILC, intervals were derived by assuming that selection biases for dichotomous survey variables are smaller in absolute size than 5 percentage points.

Fourth, mode-specific measurement biases may change gradually over time. For example, respondents are likely to have become more familiar with web questionnaires over the past decade, but, simultaneously, the range of online devices is growing. Ideally, a reinterview should be conducted with a certain frequency. In Schouten et al. (2019) the adjustment perspective under time-varying measurement errors is also considered. Here, it is assumed that changes in mode-specific measurement bias are negligible over a period of T years.

Fifth, the m_1 respondents do not all have to be invited for a reinterview and the m_1 non-respondents do not all have to be followed up by mode m_2. The subsampling probabilities are denoted as π_{RE} and π_{FU}, respectively. Schouten et al. (2019) derive optimal values of both subsampling probabilities for different values of T.

26.3. Results for the Dutch EU-SILC

This section provides a brief description of the Dutch EU-SILC data. It then provides the outcomes of the cost–benefit analysis under the design perspective and the adjustment perspective.

26.3.1. EU-SILC data

The 2019 Dutch EU-SILC data were used to construct parameters for the cost–benefit analysis. Unfortunately, this means that the face-to-face survey mode, which is still a prominent mode in EU-SILC in other countries, is not considered in this analysis.

The 2019 EU-SILC had a sample size of 15 494 households, of which 4 214 responded in web mode and 1 551 responded by telephone. This amounts to a response rate of around 37 %, of which web respondents accounted for 73 % and telephone respondents accounted for 27 %.

Three EU-SILC variables are considered (presented in Table 26.1):

- household can make ends meet (HS120): very easily, easily, fairly easily, with some difficulty, with difficulty or with great difficulty,
- self-assessed health (PH010): very good, good, fair, bad or very bad, and missing values are set to a separate category,
- socioeconomic status (RB210): employed, unemployed, volunteer work, student, disabled, retired, household, other status.

Household income itself is not considered, as in the Netherlands this is a register variable that is linked to EU-SILC.

As some categories of socioeconomic status are relatively small or strongly related to age, such as students and retired people, only the categories of employed, unemployed, disabled and household are considered in the cost–benefit analysis. Each category is dichotomised; hence, there are six variables in the analysis for which the three estimators are evaluated.

26.3.2. Design perspective

Under the design perspective, the choice is between the single mode and the sequential mode design. Following the approach in Schouten et al. (2019), the root mean square error (RMSE) values are derived for the two estimators under both measurement benchmarks and for the three points in the mode-specific measurement bias interval. Table 26.2 show the results for the six dichotomised variables of EU-SILC presented in Section 26.3.1.

In Table 26.2 the estimators that have superior accuracy are highlighted. Only if the preferred estimator varies greatly, even when a measurement benchmark is fixed, can a reinterview be beneficial under the design perspective. This is because the reinterview acts as arbitrage between the two estimators. Looking at Table 26.2, it is clear that the preferred estimator varies little once the measurement benchmark is fixed. When mode m_1 is the benchmark, the single mode design is usually the best (15 out of 18 times in this illustration). When mode m_2 is the benchmark, then the sequential mode design is often the best (14 out of 18 times). The conclusion, thus, is that a reinterview design is not beneficial from a design perspective. Under the design perspective, the choice amounts to a choice of measurement benchmark.

26.3.3. Adjustment perspective

Under the adjustment perspective, the adjusted sequential mode design also comes into play and the comparison is between three estimators. Again, we follow the approach taken by Schouten et al. (2019). We compare the RMSE values for different measurement benchmarks and different mode-specific

Table 26.2: RMSE values for the single mode and sequential mode design for the two measurement benchmarks and for different mode-specific measurement bias levels

Variable (*Y*)	Estimator	Benchmark = m_1; measurement error bias level			Benchmark = m_2; measurement error bias level		
		Left	Mid	Right	Left	Mid	Right
Health	$\hat{Y}_{m_1}$	< 0.1	0.6	1.2	4.3	2.7	1.0
	$\hat{Y}_{m_1 \to m_2}$	1.3	0.8	0.6	3.2	1.6	0.6
Make ends meet	$\hat{Y}_{m_1}$	< 0.1	0.4	0.8	2.1	1.0	0.1
	$\hat{Y}_{m_1 \to m_2}$	0.7	0.4	0.5	1.6	0.6	0.8
Employed	$\hat{Y}_{m_1}$	0.1	0.7	1.4	10.6	8.8	7.0
	$\hat{Y}_{m_1 \to m_2}$	2.9	2.3	1.7	7.8	5.9	4.1
Not employed	$\hat{Y}_{m_1}$	< 0.1	0.2	0.4	1.5	0.9	0.3
	$\hat{Y}_{m_1 \to m_2}$	0.5	0.3	0.2	1.1	0.5	0.2
Disabled	$\hat{Y}_{m_1}$	< 0.1	0.3	0.6	0.2	0.6	1.4
	$\hat{Y}_{m_1 \to m_2}$	0.3	0.4	0.6	0.3	0.7	1.4
Household	$\hat{Y}_{m_1}$	< 0.1	0.3	0.5	8.4	7.7	7.0
	$\hat{Y}_{m_1 \to m_2}$	2.3	2.0	1.8	6.1	5.4	4.8

NB: For each of the six estimates, the estimator that has the lowest RMSE per benchmark and level is shaded in blue. For example, for the variable 'household', the single mode estimator has lower RMSE values at all three bias levels if m_1 is the benchmark, but the sequential mode estimator has lower RMSE values if m_2 is the benchmark.

measurement bias levels. We also evaluate the performance of the adjusted sequential mode design for the three time periods of 2 years, 5 years and 10 years.

Table 26.3 shows the RMSE values for the different benchmarks, bias levels and time horizons. The results indicate that when mode m_1 is the measurement benchmark the single mode design is preferred under many scenarios. This means that a reinterview would not be beneficial. However, when mode m_2 is the measurement benchmark,

Table 26.3: RMSE values for the three estimators under the two measurement benchmarks, three measurement bias levels and three survey time horizons

Variable (Y)	Estimator	Time (years)	Benchmark = m_1; measurement error bias level			Benchmark = m_2; measurement error bias level		
			Left	Mid	Right	Left	Mid	Right
Health	$\hat{Y}^{adj}_{m_1 \to m_2}$	2	0.9	0.9	0.9	0.9	0.9	0.9
		5	0.8	0.8	0.8	0.8	0.8	0.8
		10	0.7	0.7	0.7	0.7	0.7	0.7
	$\hat{Y}_{m_1}$		< 0.1	0.6	1.2	4.3	2.7	1.0
	$\hat{Y}_{m_1 \to m_2}$		1.3	0.8	0.6	3.2	1.6	0.6
Make ends meet	$\hat{Y}^{adj}_{m_1 \to m_2}$	2	0.6	0.6	0.6	0.6	0.6	0.6
		5	0.5	0.5	0.5	0.6	0.6	0.6
		10	0.5	0.5	0.5	0.5	0.5	0.5
	$\hat{Y}_{m_1}$		< 0.1	0.4	0.8	2.1	1.0	0.1
	$\hat{Y}_{m_1 \to m_2}$		0.7	0.4	0.5	1.6	0.6	0.8
Employed	$\hat{Y}^{adj}_{m_1 \to m_2}$	2	0.9	0.9	0.9	1.0	1.0	1.0
		5	0.8	0.8	0.8	0.9	0.9	0.9
		10	0.8	0.8	0.8	0.9	0.9	0.9
	$\hat{Y}_{m_1}$		0.1	0.7	1.4	10.6	8.8	7.0
	$\hat{Y}_{m_1 \to m_2}$		2.9	2.3	1.7	7.8	5.9	4.1
Not employed	$\hat{Y}^{adj}_{m_1 \to m_2}$	2	0.3	0.3	0.3	0.4	0.4	0.4
		5	0.3	0.3	0.3	0.3	0.3	0.3
		10	0.3	0.3	0.3	0.3	0.3	0.3
	$\hat{Y}_{m_1}$		< 0.1	0.2	0.4	1.5	0.9	0.3
	$\hat{Y}_{m_1 \to m_2}$		0.5	0.3	0.2	1.1	0.5	0.2
Disabled	$\hat{Y}^{adj}_{m_1 \to m_2}$	2	0.4	0.4	0.4	0.4	0.4	0.4
		5	0.4	0.4	0.4	0.4	0.4	0.4
		10	0.4	0.4	0.4	0.4	0.4	0.4
	$\hat{Y}_{m_1}$		< 0.1	0.3	0.6	0.2	0.6	1.4
	$\hat{Y}_{m_1 \to m_2}$		0.3	0.4	0.6	0.3	0.7	1.4
Household	$\hat{Y}^{adj}_{m_1 \to m_2}$	2	0.4	0.4	0.4	0.5	0.5	0.5
		5	0.4	0.4	0.4	0.4	0.4	0.4
		10	0.4	0.4	0.4	0.4	0.4	0.4
	$\hat{Y}_{m_1}$		< 0.1	0.3	0.5	8.4	7.7	7.0
	$\hat{Y}_{m_1 \to m_2}$		2.3	2.0	1.8	6.1	5.4	4.8

NB: For each of the six estimates, the estimator that has the lowest RMSE per benchmark and bias level is shaded in blue.

the picture is quite different and the adjusted sequential design is often superior to the other two estimators. This means that the gain in bias from a reinterview outweighs the loss in variance. Remarkably, the length of the time horizon T is not very influential. Even for relatively small time horizons of 2 years the reinterview can be worthwhile. This means that the bias is the dominant term in the RMSE.

26.4. Conclusions

The cost–benefit analysis for the Dutch EU-SILC demonstrated that reinterview designs to decompose mode effects on key variables can be beneficial from an accuracy perspective. In other words, the loss in variance that results from freeing some budget to pay for the reinterview is smaller than the gain in bias resulting after adjustment. This conclusion was true only under the adjustment perspective, where measurement biases are corrected, and not under the design perspective, where a reinterview is merely used to choose between designs.

It should be noted that reinterview designs go with assumptions. These have been discussed. We believe that EU-SILC statistics are relatively stable over time, so change over time will not be too influential when using reinterview answers. We also believe that a reinterview may be organised such that it is acceptable to respondents that they are asked for another interview. This, however, implies that a new questionnaire will be needed that includes some modules/topics that were not part of the original questionnaire. In order to avoid context effects, the first part of the questionnaire should remain the same. Consequently, it will not be possible to decompose mode effects for key EU-SILC variables that appear late in the questionnaire. Furthermore, EU-SILC is a panel study and the situation in which the reinterview leads to intermediate attrition on top of the normal attrition should be avoided.

This study has two clear limitations. The first is that data for only one country, the Netherlands, have been investigated. It would be relatively straightforward to replicate the findings for other countries that have employed a sequential mixed-mode design. R code is available on request for that purpose. The second limitation, related to the first, is that we considered only web mode and telephone mode as survey modes. As the Dutch EU-SILC does not employ face-to-face interviews, this mode was out of scope. Clearly, face-to-face mode is an important mode. Klausch et al. (2017) show how a reinterview can be carried out for designs with three survey modes.

Future studies should replicate the findings for other countries and designs that include face-to-face and possibly paper modes.

References

Biemer, P. (2001), 'Nonresponse bias and measurement bias in a comparison of face-to-face and telephone interviewing', *Journal of Official Statistics*, Vol. 17, No 2, pp. 295–320.

Buelens, B., van den Brakel, J. and Schouten, B. (2019), 'Detecting and adjusting for mode effects – a literature review', ESSnet MIMOD, Statistics Netherlands, Heerlen, the Netherlands (https://www.istat.it/en/research-activity/international-research-activity/essnet-and-grants).

Dillman, D. A., Smyth, J. D. and Christian, L. M. (2014), *Internet, Phone, Mail and Mixed-mode Surveys – The tailored design method*, 4th edition, Wiley, New York.

Klausch, L. T., Schouten, B., Buelens, B. and van den Brakel, J. (2017), 'Adjusting measurement bias in sequential mixed-mode surveys using re-interview data', *Journal of Survey Statistics and Methodology*, Vol. 5, No 4, pp. 409–432.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J. and Klausch, L. T. (2013), 'Disentangling mode-specific selection and measurement bias in social surveys', *Social Science Research*, Vol. 42, pp. 1555–1570.

Schouten, B., Klausch, T., Buelens, B. and van den Brakel, J. (2019), 'A cost–benefit analysis of re-interview designs for mode-specific measurement bias', ESSnet MIMOD, Statistics Netherlands, Den Haag, The Netherlands (https://www.istat.it/en/research-activity/international-research-activity/essnet-and-grants).

27 Mode and web panel experiments in the European Social Survey – lessons for EU-SILC

Rory Fitzgerald and Eva Aizpurua ([147])

27.1. Introduction

Using different modes of data collection within or between countries in a cross-national social survey has the potential to introduce methodological artefacts into the data (Martin, 2011). This means that analysts may appear to find no differences in the data or find differences that reflect the mix of data collection modes used rather than the real-world situation. The literature has shown that mode effects are likely to vary by topic, question type and country context, and are therefore hard to predict (Martin and Lynn, 2011). At the same time, methods to measure and control for mode effects are difficult and costly to implement and, in many respects, still in their infancy in terms of development (Olson et al., 2020). This means that, when modes are mixed, great care should be taken in the design and analysis stages to take account of the impact that different modes can have on the data collected.

The best way to eliminate the risk of mode effects is to ask all respondents to complete the survey using the same mode or to use a combination of modes in which all respondents answer each question using the same mode (e.g. self-administering sensitive questions in the context of an interviewer-administered survey). If a combination of modes is to be used, efforts should be made to minimise mode effects (e.g. by ensuring that visual information is provided, rather than sometimes providing information visually and on other occasions providing it aurally). The European Social Survey (ESS) therefore decided in 2001, when it was established, to use face-to-face interviewing in all countries and has done so ever since. This decision was made because no other mode could be effectively used for interviewing in all countries and for all respondents, especially considering the length (approximately 60 minutes in English) and complexity of the survey, as well as the differences in penetration of internet and telephone technology across countries. However, it was acknowledged at the time that other modes might become more feasible in the future as a single-mode alternative and that it might become essential to combine modes for other reasons, such as increased costs of face-to-face data collection. In order to gather information to make an informed decision, the ESS established a mixed-mode methodological research programme composed of a series of six experiments. Following the conclusion of that work, it was decided on the advice of its Methods Advisory Board not to switch to mixed-mode data collection. Instead, the ESS experimented with recruiting a web panel off the back of its face-to-face survey (the Cross-national Online Survey (CRONOS)). This experimental work is of relevance to EU-SILC, in which modes are routinely mixed.

This chapter starts by providing an overview of the ESS before moving on to discuss the challenges

([146]) Rory Fitzgerald is the director of the European Social Survey (ESS) European Research Infrastructure Consortium and professor of practice in survey research at City, University of London, United Kingdom. Eva Aizpurua is a research fellow at the ESS headquarters (City, University of London, United Kingdom). The authors would like to thank Peter Lynn and Lars Lyberg for their feedback. All errors and opinions are the authors' responsibility. This work was supported by Net-SILC3, funded by Eurostat and coordinated by the Luxembourg Institute of Socio-Economic Research. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the authors. Correspondence should be addressed to Rory Fitzgerald (r.fitzgerald@city.ac.uk).

Table 27.1: Overview of the ESS

Time span	2002 to present
Frequency	Every 2 years
Management	ESS ERIC
Design	Repeated cross-sectional
Central topics	Attitudes, beliefs, values, perceptions and behaviour patterns
Target population	Residents aged 15 and over
Sampling	Probabilistic
Sample size	1 500 in each country (effective) 800 in countries with a population of less than 2 million (effective)
Survey mode	Face-to-face interviewing (CAPI)
Source questionnaire language	English
Translation	Languages spoken by 5 % of the population and more
Interview duration	60 minutes (English questionnaire)
Data access	Free of charge for non-commercial use

Source: ESS website (https://www.europeansocialsurvey.org/).

faced by cross-national surveys in respect of data collection, focusing on increased costs, decreasing response rates and a contraction in interviewer capacity. We then introduce different mixed-mode designs and summarise the experiments conducted to assess the feasibility and the impact of mixed-mode data collection in the ESS and more recent experiments conducted by the European Values Study (EVS). We continue with a description of the first cross-national, input-harmonised, probability-based web panel, CRONOS, which has been implemented in three countries. In the last section of the chapter, we discuss some of the lessons learned and introduce CRONOS-2, a 12-country web panel currently under construction by the ESS. The chapter concludes by considering the possible implications of this experimental work on data collection mode conducted by the ESS for EU-SILC.

27.2. European Social Survey

The ESS is an academically led cross-national survey that has been conducted across large parts of Europe since its establishment in 2001. From 2002 to 2019, 39 countries participated in one or more rounds of the ESS ([147]). The survey is currently conducted face-to-face using computer-assisted personal interviewing (CAPI) and consists of a core module, which remains largely stable from round to round, and rotating modules, which may be new or repeated from previous rounds (e.g. personal and social well-being, timing of life, welfare attitudes). Table 27.1 summarises the main characteristics of the ESS.

The aim of the ESS is to measure attitudes, beliefs, values and behaviour patterns, providing comparative data across countries and time. The ESS uses probability sampling with the aim of covering residents aged 15 and over, regardless of their nationality, citizenship or language. An effective sample size ([148]) of 1 500 is aimed for in each participating country (800 for countries with fewer than 2 million inhabitants). The ESS was awarded European Research Infrastructure Consortium (ERIC) status in 2013. The data are available free of charge for non-profit purposes and are widely used (as at July 2020, the ESS had 162 730 registered users from over 240 countries ([149])). The number of English-language publications and presentations exceeded

([147]) Further information about the survey is available on the ESS website (https://www.europeansocialsurvey.org/).

([148]) Effective sample size refers to the actual sample size (i.e. the number of observations) divided by the design effect. This is the size of a simple random sample that would have produced the same precision.

([149]) Monthly statistics on ESS data usage can be found on the ESS website (https://www.europeansocialsurvey.org/about/user_statistics.html).

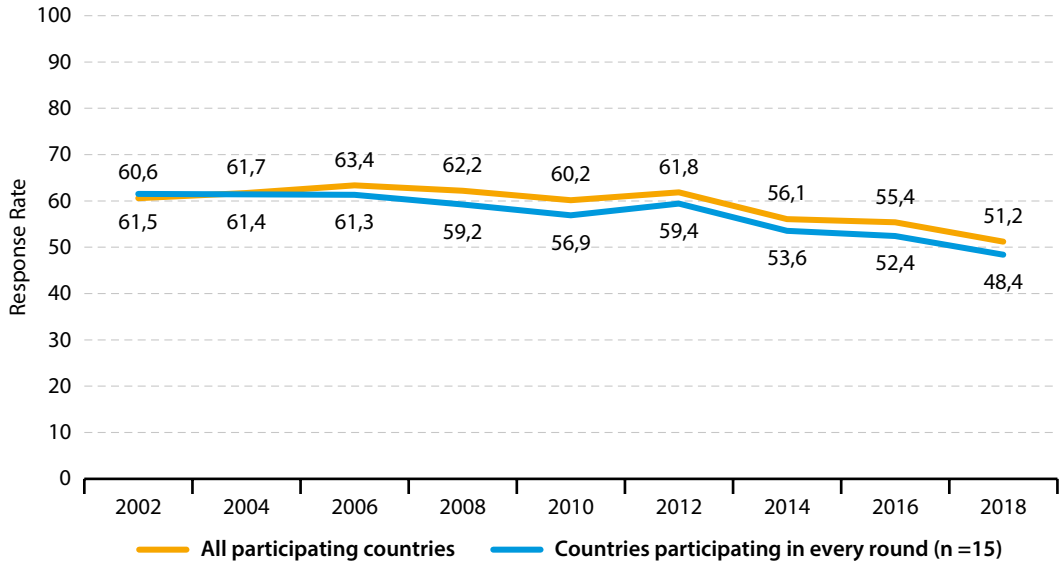
4 400 in 2019 (Malnar, 2020). The data also have an extensive impact beyond academia in policy and third-sector work (see Technopolis Group, 2017). The survey further aims to increase cross-national comparability by using an input-harmonised approach where possible and functionally equivalent approaches where total harmonisation is not possible (Fitzgerald and Jowell, 2010).

27.3. Challenges faced by cross-national general social surveys in terms of data collection and mode

Contemporary survey research faces important challenges, mainly related to declining response rates and increasing data collection costs (Leeper, 2019; Luiten, Hox and de Leeuw, 2020). Despite having relatively high response rates and increased fieldwork efforts to maximise these rates, the ESS has not been immune to this trend, particularly in recent years. As shown in Figure 27.1, there has been a substantial decline in response rates, which has been attributed more to refusals than to non-contacts (Beullens et al., 2018).

Although this trend is observable at the aggregate level, large differences exist across countries, in both the direction and the magnitude of the change in response rates over the years, because of differences in survey climate and tradition (ESS survey documentation reports, rounds 1–9 ([150])). Figure 27.2 displays the changes in response rates across the rounds for three selected countries with distinct trends. Although Finland and Switzerland both achieved the same response rate (51.8 %) in round 9 (2018) of the ESS, the former exhibits a downward trend (from 73.2 % in 2002), whereas Switzerland's response rates increased by 19.9 percentage points from 2002 to 2010 and stabilised afterwards. Belgium, however, remained relatively stable over the years, with only small fluctuations in response rates from round to round.

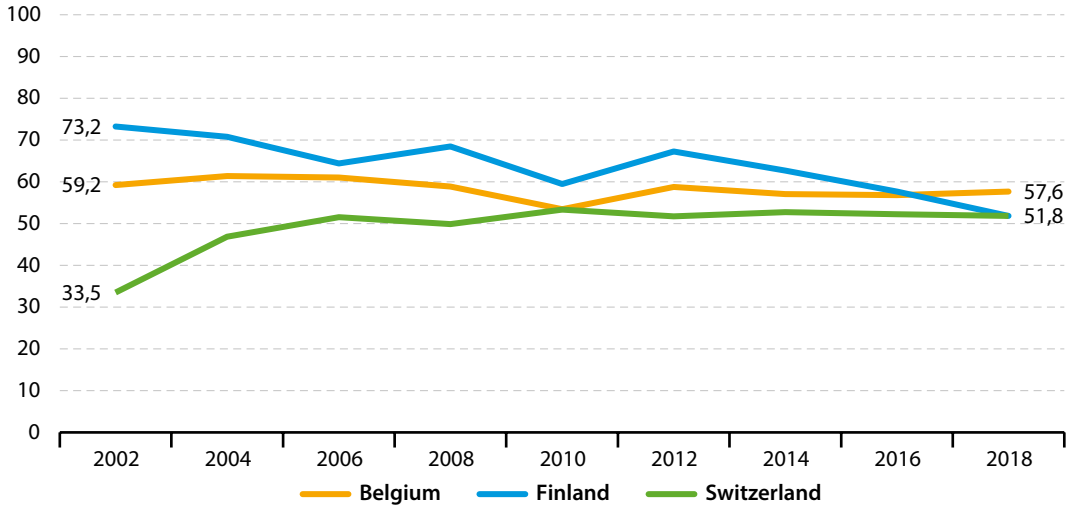
Figure 27.1: Average ESS response rates in rounds 1–9, 2002–2018



Source: ESS survey documentation reports, rounds 1–9.

([150]) https://www.europeansocialsurvey.org/data/round-index.html

Figure 27.2: ESS response rates in rounds 1–9 in selected countries, 2002–2018

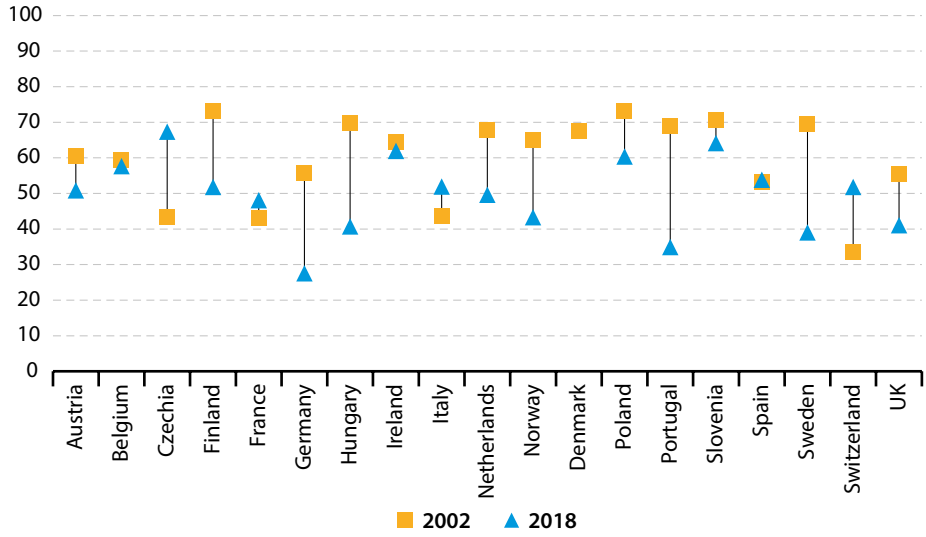


Source: ESS survey documentation reports, rounds 1–9.

Figure 27.3 shows response rates achieved in 2002 and 2018 in the 18 countries participating in rounds 1 and 9 of the ESS. Despite a general decline, which is observable for 11 countries, the differences are heterogeneous, with large decreases in Germany (from 55.7 % to 27.6 %), Hungary (from 69.9 % to 40.7 %) and Sweden (from 69.5 % to 39.0 %), and more modest decreases in others countries, including Austria (from 60.4 % to 50.9 %) and Slovenia (from 70.5 % to 64.1 %). A few countries obtained very similar response rates in both rounds (Belgium, Ireland and Spain), whereas others improved theirs. In this group are Czechia and Switzerland with large increases (24.0 and 18.4 percentage points, respectively) and France and Italy with smaller gains (5.0 and 8.2 percentage points, respectively).

In a recent assessment of response rates across surveys in Europe from 1998 to 2015, it was found that, although there were differences in the rate of decline, all types of surveys showed a downward trend in response rates (Luiten, Hox and de Leeuw, 2020). The decline in response rates has prompted survey organisations worldwide to consider alternative modes of data collection, including combinations of survey modes. In addition, as more research has moved away from using face-to-face data collection, fewer agencies now offer this to the standard required for high-quality surveys such as the ESS and the Survey of Health, Ageing and Retirement in Europe (Sommer, 2019). In cross-national surveys, however, combining survey modes brings additional challenges associated with differences in technology penetration and disparate sampling frames (de Leeuw, Suzer-Gurtekin and Hox, 2019). The sampling frames available in some countries require contact to be made in person, although using mail or telephone is sometimes also possible. In general, if an interviewer is not to be used, then a frame of individuals is needed, so that letters can be sent to the target respondents, or alternatively a relatively complete frame of telephone numbers. These, however, are not always available. When population registers are not available, in-person contact is the best way to appropriately sample individuals within households. Unlike interviewer-administered selection procedures, which are well established, within-household selection in self-administered modes is more difficult to accomplish, with studies finding between 20 % and 30 % of selections in mail and web surveys to be inaccurate (Olson and Smyth, 2014, 2017). In addition, there are no acceptable general population frames of email addresses and no acceptable ways of drawing probability samples from them

Figure 27.3: ESS response rates in rounds 1 (2002) and 9 (2018)



Source: ESS survey documentation reports, rounds 1 and 9.

(Dillman, 2017). In the case of online surveys, it is therefore necessary to mix modes of contact (e.g. sending advance letters and invitations to participate in web surveys by post), which increases the costs and complexities of data collection.

Logistical demands are also amplified in cross-national surveys, as the number of actors increases along with diverging contexts. In addition, within research infrastructures levels of experience and expertise vary widely, even across European countries. Although recruiting qualified interviewers may be relatively easy in some countries, other countries – particularly those in which the chosen mode is infrequently used – may experience difficulties in hiring seasoned interviewers (De Jong, 2016). This has important implications, because inexperienced interviewers tend to produce lower response rates (West and Blom, 2017; Wuyts and Loosveldt, 2020). At the same time, the scarcity of interviewers often results in increased workloads, which lead to larger interviewer effects (West and Blom, 2017; see also Chapter 28 of this book). Furthermore, the reliability of postal systems is rather uneven across European countries, making contact by post difficult to implement consistently. This contributes to cross-national differences in the outcomes of push-to-web or postal self-completion approaches, threatening the comparability of the data. A push-to-web (online only) cross-national design implemented by the European Union Agency for Fundamental Rights in 2017 achieved an average response rate of 18 % in countries using named person samples (Denmark, Estonia, Italy, Luxembourg, Hungary, Malta, the Netherlands, Poland, Slovenia, Finland and Sweden). However, response rates were much lower in countries not using individual registers. Specifically, address-based samples (in Austria, Belgium, Bulgaria, Croatia, France, Ireland, Latvia, Lithuania, Spain and the United Kingdom) had an average response rate of 7 %, whereas this percentage dropped to 3 % in enumeration countries (Czechia, Germany, Greece, Italy, Cyprus, Portugal, Romania and Slovakia; Smith, 2018). The ESS has recently (2021) conducted a push-to-web mixed-mode study (web and postal) in three European countries – Austria, Hungary and Serbia – with the goal of testing this approach in a cross-national setting. This three-country study has provided encouraging results, with response rates of around 40 % in all countries, two of which used address-based samples (Austria and Serbia) and only one of which used a named person sample (Hungary). It will also provide insights into the representativeness of the achieved samples. These findings should help inform survey designs as push-to-web

data collection methods gain popularity (for a review of push-to-web surveys, see Dillman, 2017).

Equivalent measurement is one of the main challenges of cross-national surveys that are comparative by design, such as the ESS. Questionnaires not only are required to be culturally relevant within and across countries but also need to accommodate multiple languages and provide invariant measures (Pennell et al., 2017). When comparing groups, functional equivalence is necessary to ensure that observed differences represent actual differences and are not the result of other factors, such as differences in data collection modes or non-equivalent translations. However, removing all sources of error from surveys is not possible. In single-country surveys, the challenge is the optimal allocation of resources to minimise total survey error. In the context of cross-national research, however, the goal is to minimise error and to make error components similar in magnitude and direction across countries (Smith, 2011). This can be promoted through planning, coordination and adoption of comparable protocols of data collection. In cross-national surveys, strong infrastructures that support and monitor the design and implementation of the survey are particularly important. The ESS infrastructure, for example, is led by a Core Scientific Team that ensures careful joined-up planning, provides support to national teams and monitors activities at the national level to maximise compliance at every stage. The Scientific Advisory Board and Methods Advisory Board ensure that the approaches used remain 'state of the art'.

27.4. Mixed-mode survey designs

Mixed-mode designs are those in which respondents answer the same questions using different modes (e.g. some respondents are interviewed face-to-face while others complete the survey online). Sometimes respondents are offered a choice between multiple modes (e.g. web, telephone, face-to-face) in what is called a *concurrent mixed-mode design*. At other times potential respondents are assigned to different modes depending on the information that is available (e.g. telephone surveys for sample members with a telephone number and face-to-face interviews for those with addresses only). On other occasions, respondents are invited to participate using a certain mode first (usually the most cost-effective mode) and offered additional modes if they are unable or unwilling to respond (*sequential mixed-mode design*; for a review of mixed-mode survey designs, see de Leeuw, 2018).

In the context of cross-national research, countries might use different modes of data collection, resulting in what has been called *across-country mixed-mode designs*. An example of this design is found in the International Social Survey Programme, which in recent years has allowed countries to choose between face-to-face interviewing, self-administered surveys and telephone surveys ([151]). When one or more countries combine modes of data collection, using a concurrent or a sequential approach, the design becomes a *within-country mixed-mode design*. The third variation in cross-national time-series surveys is the *across-time mixed-mode design*, which occurs when countries transition from a single mode or combination of modes to a different mode (Martin, 2011). These three designs are not exclusive and, as shown in Chapter 24 of this book, they coexist in the context of EU-SILC.

Mixed-mode designs have increased in recent years due to their potential to lower financial costs and reduce coverage and unit non-response errors. However, mixing modes of data collection is not without drawbacks, as this practice may threaten the comparability of the data between groups and, in the case of time-series surveys, across time (de Leeuw, Hox and Scherpenzeel, 2019). In this context, measurement differences attributed to survey mode (e.g. social desirability bias) may be confounded with substantive differences, undermining the comparisons made. Mixed-mode designs also increase the logistic complexity of the survey, requiring additional work at the design (e.g. adapting the questionnaires to different modes), implementation (e.g. following up in different modes, additional coordination) and analysis (e.g. data cleaning, adjustment and harmonisation) stages (Martin, 2011).

([151]) Information on modes of data collection is available on the website of the Leibniz Institute for the Social Sciences (https://zacat.gesis.org/webview/index/en/ZACAT/ZACAT.c.ZACAT/ISSP.d.58/by-Year/fCatalog/Catalog69).

Table 27.2: Summary of the mixed-mode experiments conducted as part of the ESS methodological programme, 2003–2012

Study	Year (round)	Country	Mode	Sampling	Research design
Study 1: Measurement differences across four modes	2003 (R1)	Hungary	Face-to-face, telephone, PAPI and web surveys	Convenience	Within-subjects reinterview design
Study 2: Causes of measurement differences between face-to-face surveys and telephone surveys	2005 (R2)	Hungary (Budapest) and Portugal (Lisbon)	Face-to-face (with and without showcards) and telephone surveys	Probabilistic	Between-subjects design
Study 3: Measurement differences between face-to-face surveys and web surveys	2010–2011 (R5)	United Kingdom	Face-to-face and web surveys	Probabilistic	Within-subjects reinterview design
Study 4: Feasibility of conducting the ESS using telephone surveys	2006 (R3)	Cyprus, Germany, Hungary, Poland and Switzerland	Face-to-face and telephone surveys	Probabilistic	Between-subjects design
Study 5: Comparing concurrent and sequential mixed-mode designs	2008 (R4)	Netherlands	Face-to-face, telephone and online surveys	Probabilistic	Between-subjects design
Study 6: Feasibility of mixed-mode designs across countries	2012 (R6)	Estonia, Sweden and the United Kingdom	Face-to-face, telephone and web surveys	Probabilistic	Between-subjects design

NB: PAPI, paper and pencil interviewing.

Source: Adapted from Villar and Fitzgerald (2017).

27.5. Mode experiments in the European Social Survey

To assess the feasibility and implications of transitioning from the face-to-face mode to a different mode, or a combination of modes, the ESS implemented a methodological programme to assess the impact of mixing data collection modes on the quality of survey estimates (for a detailed review of the experiments and findings, see Villar and Fitzgerald, 2017). This programme included six experiments conducted in 10 European countries between 2003 and 2012. A summary of the experiments can be found in Table 27.2. The goal of the programme was to inform the implementation of future rounds of the ESS, providing evidence to support or discard the adoption of a mixed-mode design. The first three experiments focused on measurement equivalence, whereas the following three examined the feasibility of conducting the ESS using a different mode or a combination of modes. The remainder of this section summarises the findings of this methodological programme with regard to survey participation (response rates and sample composition) and measurement effects. This is supplemented by the results of mixed-mode experiments conducted by the EVS in six countries (Denmark, Finland, Germany, Iceland, the Netherlands and Switzerland) during wave 5 (2017).

27.5.1. Survey participation

One of the premises of mixed-mode designs is that each survey mode may attract different types of respondents. As a result, combining data collection modes has the potential to reduce coverage and non-response errors (de Leeuw, 2018). For this reason, selection effects (i.e. different types of respondents selecting different modes) are desired, although later adjustments may be needed, particularly in across-country mixed-mode designs

in which selection effects may differ between the countries. The findings from the ESS methodological programme did not show improvements in response rates compared with the face-to-face main survey; they pointed, instead, to a deterioration that varied in magnitude depending on the country and the mode(s). For example, in the study conducted in the Netherlands comparing face-to-face interviewing with a concurrent mixed-mode design and a sequential mixed-mode design (online, telephone and face-to-face interviews), response rates were lower for the mixed-mode designs (46 % in the concurrent design and 45 % in the sequential design, compared with 52 % in the face-to-face single-mode design). In an earlier study comparing telephone and face-to-face surveys in five European countries, it was found that, when administering the full ESS (which takes approximately 60 minutes), response rates were consistently lower in the telephone mode. Differences varied widely across the countries, ranging from a relatively small difference in Switzerland (38 % versus 46 %) to a very large difference in Hungary (18 % versus 66 %; Villar and Fitzgerald, 2017). The lack of improvement in response rates was expected, given that face-to-face surveys tend to have the highest coverage and response rates. The more recent experiments conducted during the last wave of the 2017 EVS (Luijkx et al., 2020) show that, in three countries, response rates were lower in the self-administered modes than for face-to-face interviewing (15 % lower in Switzerland, 17 % lower in Denmark and 29 % lower in Finland), whereas in Germany and Iceland response rates were actually lower in the face-to-face mode (28 % versus 35 % and 41 % versus 45 %, respectively) ([152]) (Christmann et al., 2019).

Another dimension explored in the ESS experiments was the demographic composition of the samples achieved. Different modes are linked to different levels of non-coverage and non-response errors, and they are likely to attract different groups of respondents. This is, in fact, one of the benefits of mixed-mode designs: because of selection effects, respondents who would not or could not participate in a single-mode design may participate if multiple modes are offered. The major problem is that selection effects and measurement effects (e.g. social desirability, acquiescence) are often confounded, making it difficult to ascertain the extent to which differences (or similarities) between the modes are the result of differences in respondents or differences in measurement error (Vannieuwenhuyze, Loosveldt and Molenberghs, 2010). The findings of the ESS programme pointed to small differences in sample composition between the single-mode design and the mixed-mode designs. When differences were found, the composition of the face-to-face survey sample was generally closer to the population estimates than the composition of mixed-mode surveys (Villar and Fitzgerald, 2017). For example, two of the studies comparing telephone surveys with the standard face-to-face mode indicated that telephone interviews tended to over-represent those with higher educational levels. However, the results were not consistent, revealing cross-national differences in how mode affected sample composition.

Results from the 2017 EVS experiments are consistent with these findings, suggesting that samples obtained through face-to-face interviewing tend to be more similar to the overall population, although differences were generally small (Christmann et al., 2019). In the case of Germany, where the face-to-face mode was compared with two self-administered mixed-mode (computer-assisted web interviewing and paper self-completion) designs – one a matrix design ([153]) and one with a full-length questionnaire – differences were found in some variables (e.g. age, nationality, household size) but not in others (e.g. gender), and the size of the differences fluctuated. For instance, the samples achieved under-represented foreigners in all cases, but differences were larger in the two mixed-mode designs. Although nearly one in eight individuals in the population was foreign, this fraction dropped to one in nine for the face-to-face interviews and to 1 in 15 for the mixed-mode designs (Christmann et al., 2019). In terms of education, the samples achieved over-represented the group with the highest educational level, with the largest difference being between the mixed-mode

([152]) As part of the EVS experiments, full-length and matrix questionnaires were used. The figures reported here refer to the comparisons between the full length (approximately 60 minutes) face-to-face and self-administered questionnaires.

([153]) In a matrix design, the questionnaire is split into shorter versions to which respondents are randomly assigned. For a full description of the experiments, please see Luijkx et al. (2020).

designs and the population (40.3 % for the matrix design and 39.0 % for the design with the full questionnaire, compared with 24.3 % among the population) and a smallert difference between the face-to-face single-mode design (34.8 %) and the population (Christmann et al., 2019).

27.5.2. Measurement effects

Mode effects are the result of both mode self-selection effects, which are produced when assignment to the modes is not randomised, and measurement effects, which are attributed to mode differences (e.g. interviewer effects, questionnaire design). Unlike mode selection effects, mode measurement effects represent a source of measurement error and, in a mixed-mode survey, constitute undesired effects. They arise when respondents' answers depend on the mode of data collection (e.g. more honest responses to the same question when it is self-administered). The results from the ESS experiments revealed differential mode measurement effects that threatened the equivalence and the comparability of the data. For example, the study conducted in Hungary and Portugal indicated that telephone respondents were more likely to provide socially desirable responses across a range of indicators than face-to-face respondents (Jäckle, Roberts and Lynn, 2006). In general, attitudinal questions, which are dominant in the ESS, showed a lower level of consistency across modes than behavioural questions. Among these attitudinal questions, the largest differences were found for estimates of personal well-being, political attitudes and participation, and attitudes towards immigrants. In addition, large differences were found for self-reported income, a variable that, in general surveys, tends to yield low-quality data in terms of high item non-response and inconsistencies with administrative data (Moore, Stinson and Welniak, 2000) – which is particularly important for EU-SILC. One of the experiments used a reinterview design, randomly assigning respondents to complete the survey in a different mode, finding that answers to the income question were different in 48 % of the cases (Villar and Fitzgerald, 2017). A later study also revealed that telephone respondents were less likely to report lower household income than those interviewed face-to-face (Jäckle, Roberts and Lynn, 2006).

Although some measurement effects are inherent in the mode, others can be avoided or minimised. Changes in questionnaire design between modes can exacerbate the unwanted effects, threatening the comparability of the data (de Leeuw, Hox and Scherpenzeel, 2019). In mixed-mode surveys, designing and implementing questionnaires that are equivalent is particularly important to prevent avoidable mode effects. For instance, a study comparing face-to-face and web responses to the 2008 Dutch EVS found that the responses to 64 % of the items differed between the modes (Bennink, Moors and Gelissen, 2013). The differences were attributed to changes in question wording (e.g. definitions provided by interviewers in face-to-face interviewing that had to be included as part of the question in the online survey) and the ways in which non-substantive responses (refusals, 'don't know') were presented (visible versus non-visible) and navigated (possibility of leaving a question unanswered).

Major challenges associated with differential measurement error include the existence of heterogeneous effects across variables and the lack of a single method that could be used to adjust for these differences in all types of analyses (Martin and Lynn, 2011). Based on the findings from the ESS programme, the Core Scientific Team decided, on the advice of the Methods Advisory Board, not to adopt a mixed-mode strategy, continuing instead with the face-to-face mode. It was agreed, however, that implementing a cross-national probability-based web panel, to be recruited off the back of the ESS, would be trialled. In a sense, this still leads to a mixed-mode design, as analysts can combine answers from the face-to-face survey with those from web follow-ups at the individual level, but the main ESS remains, at least for now, in face-to-face mode ([154]).

27.6. Cross-national Online Survey

Internet use continues to increase, with 85 % of Europeans using the internet at least once a week

([154]) The Core Scientific Team of the ESS is currently reviewing whether to recommend a change to the mode of data collection in the future, including the possibility of using a combination of modes.

Table 27.3: Characteristics of CRONOS

Participating countries	Estonia Great Britain Slovenia
Data collection years	2016–2018
Recruitment approach	Piggyback sampling (ESS round 8)
Population	All ESS respondents aged 18 and over
Incentives	Unconditional (GBP 5 / EUR 5 with each survey invitation)
Number of waves	Six waves plus a welcome survey
Periodicity of waves	Bimonthly
Survey duration	20 minutes
Data access	Free of charge for non-commercial use (CRONOS data can be linked to ESS round 8 data)

Source: Adapted from Jessop et al. (2019).

in 2019 (DESI, 2020). This, along with the reduced costs and fieldwork times associated with online surveys, has resulted in a very rapid increase in this mode of data collection. In addition, online surveys are associated with reduced social desirability biases, one form of measurement error that occurs when respondents provide inaccurate responses to comply with social norms (Krumpal, 2013). Because interviewers are absent, acquiescent or agreeable responses, in which individuals tend to agree or provide affirmative answers to questions, are also reduced (Liu, Conrad and Lee, 2017). Despite this, online surveys have important shortcomings, including low response rates and self-selection biases. A recent meta-analysis, for example, revealed that web surveys still yield lower response rates than other modes. Daikeler, Bošnjak and Manfreda (2020) found that response rates were 12 percentage points lower for online surveys than for other modes. Online surveys have also been found to be less representative than other single-mode surveys (Cornesse and Bošnjak, 2018). Although the digital divide has lessened, access to the internet still varies widely across and within countries, with large differences in the percentage of people who regularly use it. In the United Kingdom, for instance, 99 % of adults aged 16–44 were recent internet users in 2019, whereas this percentage dropped to 47 % in the case of those aged 75 and older (ONS, 2019). As at June 2020, just 67 % of adults in Bulgaria had access to the internet, compared with 96 % of those in the Netherlands ([155]). Therefore, under-coverage and non-response are still serious threats to the validity of online surveys, especially at the cross-national level.

CRONOS was implemented during round 8 of the ESS ([156]). The objective of CRONOS was to assess the feasibility of establishing a cross-national probability-based panel following a harmonised approach. This assessment was used to create a blueprint intended to guide the development of such a panel in the future (Jessop et al., 2019). As shown in Table 27.3, CRONOS was piloted in three countries – Estonia, Great Britain and Slovenia – acting as proof of concept for the viability of a European online panel. Because CRONOS used a 'piggy-back' recruiting approach, in which all ESS adult respondents were invited – at the end of their ESS interview – to join the panel, fieldwork costs were significantly reduced (e.g. there was no need to source a new sampling frame or to hire additional interviewers). CRONOS followed a centralised management approach, with a high level of standardisation of procedures across countries while allowing adaptations if needed. A panel design such as that used in CRONOS provides important advantages, making it possible to capture individual-level variation across time. For this reason, CRONOS was seen as a very valuable complement to the main ESS, although the sample size at country level remained rather small, largely due to the orig-

([155]) https://www.internetworldstats.com/stats9.htm

([156]) The CRONOS panel work was developed under the Synergies for Europe's Research Infrastructures in the Social Sciences project, which was funded by the EU's Horizon 2020 research and innovation programme under grant agreement No 654221. The CRONOS initiative was also supported by the 2015–2017 and 2017–2019 ESS ERIC work programmes.

inal sample size in the face-to-face study and the limited cooperation in joining the panel.

The results of CRONOS highlighted the feasibility of implementing a cross-national online panel. Participation rates ([157]) were reasonable, ranging from 56 % in Great Britain to 78 % in Estonia. However, comparisons between the sample composition of CRONOS and the target population revealed multiple discrepancies. For example, CRONOS over-represented females, citizens and married individuals while under-representing older and the least educated groups (Bottoni and Fitzgerald, 2021). Although internet-enabled tablets were provided to potential respondents who had no internet access, the propensity to join the panel increased with the frequency of internet use. In addition, when comparing individuals who participated in the panel with those who did not, some differences emerged in attitudinal and behavioural indicators. For instance, it was found that CRONOS respondents had higher levels of social and institutional trust, greater life satisfaction and more tolerant attitudes towards the lesbian, gay, bisexual and transgender community than non-participants. They also reported higher levels of political participation and better perceived health.

Differences in measurement quality between questions included in the main ESS (round 8) and questions included in CRONOS have been found to be small. Non-differentiation (i.e. variance in the respondents' answers to a given topic) was equivalent across modes, whereas item non-response, although generally low, was higher in the online panel. Primacy effects (i.e. tendency to select the first answer categories) were larger in CRONOS, whereas recency effects (i.e. tendency to select the last answer categories) were generally comparable. There was also evidence of metric equivalence, providing support to the comparison of unstandardised relationships across the ESS and CRONOS. The results for scalar equivalence were less robust, suggesting some caution when comparing means (Cernat and Revilla, 2020).

CRONOS was the first attempt to establish a cross-national probability-based panel under an input harmonisation framework, in which panel design and maintenance followed the same principles in all participating countries. The results of this experience showed the feasibility of developing such a panel in terms of costs, response rates and data quality. Web panels such as CRONOS are a viable complement to ongoing cross-national surveys, providing the opportunity to further explore certain topics and evaluate individual-level differences. However, they present important challenges associated with non-response bias that require further attention. Following the successful implementation of CRONOS and taking into consideration the challenges encountered during the project, a blueprint for a comparative probability-based online survey was developed (Jessop et al., 2019). The main recommendations included in the blueprint, grouped by stage of the survey cycle, are summarised in Table 27.4.

Building on the knowledge acquired from CRONOS, the ESS is working on the implementation of a larger-scale probability-based panel. During round 10 of the ESS, adult participants will be recruited at the end of the interviews. CRONOS-2 will cover 12 European countries and will comprise six waves, allowing the study of individual- and country-level differences. It is anticipated that CRONOS-2 will help to build expertise and infrastructure so that the field is prepared for a large-scale switch to the online mode in the future. This includes the development of a sample management system for cross-national surveys ([158]) that is linked to the Qualtrics survey platform. Procedures for management of translation and for centralised communication with the panel are also being trialled. Most notably, CRONOS-2 aims to introduce web-based interviewing in a comparative format to new countries where probability-based national panels have not been established. Although piloting and capacity building are the focus now, the longer-term 'dream' is pan-European coverage, with online interviews being dominant, and a complementary mode designed to include those without internet accessgradually being phased out over time.

([157]) Calculated by dividing the number of actual participants by the number of people invited to participate in the panel.

([158]) This tool is being developed under the Horizon 2020 Social Sciences & Humanities Open Cloud project under grant agreement No 823782.

Table 27.4: Recommendations for the design and implementation of a cross-national panel

Sampling and sample management
• The parent survey recruits participants using probability sampling. • Equivalent sampling approaches are used in all participating countries, using the best random sampling practice in each case. • The sample size achieved is sufficiently large for the effective statistical analysis of country-level data. • Participants' contact details are updated throughout the duration of the panel. • Targeted approaches to fieldwork and data collection are used based on available data.
Recruitment
• All eligible people who complete the parent survey are invited to participate in the panel, including those who do not have internet access. • The recruitment approach is standardised across countries. • Panel members recruited early in the parent survey's fieldwork receive a 'welcome mailing' and a 'welcome survey' to prevent disengagement. • Incentives are used in recognition of the time and effort of panel members.
Questionnaire development
• Questionnaire content is carefully developed taking into consideration comparability issues. • The questionnaire is translated and pretested. At a minimum, a cross-national expert review and advance translation efforts are used before the questions are fielded. • Questionnaires are adapted to be displayed on multiple devices (smartphones, tablets, PCs). • Questionnaire length ranges between 15 and 20 minutes, to prevent data-quality issues and break-offs.
Fieldwork
• Between 4 and 12 waves of data are collected per year. • For each wave, fieldwork periods of around 4 weeks are recommended. • Fieldwork protocols (e.g. incentives used) are adapted to the countries to optimise response rates and sample representativeness. • Panel members are sent multiple communications to keep them engaged and informed. • The primary mode of communication is email, supplemented by other modes (e.g. postal, text messaging). • Reminders are sent at different times and on different days, not exceeding more than one in any given week. • Between-wave mailing is used to maintain the engagement of panel members.
Management and data security
• A centralised survey management approach is used to achieve high input harmonisation. • Data reduction is practised to minimise the risk of harm. • Only those who need it, and are trained, have access to identifiable information. • All data outputs are reviewed for disclosure risk.

Source: Adapted from Jessop et al. (2019).

27.7. Conclusions and lessons from the European Social Survey for EU-SILC

Cross-national general social surveys have important challenges ahead, resulting from differences in technology penetration, survey tradition and climate across countries, as well as declining response rates, rising costs and decreasing face-to-face capacity. In this environment, mixed-mode and online surveys have become increasingly popular. Although mixed-mode surveys have the potential to reduce coverage and non-response error, the results obtained under the ESS programme show that these designs do not always lead to smaller errors or better data quality and can actually introduce other forms of error (Villar and Fitzgerald, 2017). Isolating and correcting for mode-specific measurement error is still a complex task for which no universally accepted procedure exists (Martin and Lynn, 2011). This has implications for comparative surveys such as the ESS and EU-SILC, in which achieving measurement invariance is essential to study differences across countries and over time. In addition, the impact of mixed-mode approaches on the planning and management of the fieldwork requires consideration, as well as the adjustments needed after data collection, to prevent changes in

data collection modes from threatening the comparability of the estimates.

The use of online surveys has also grown considerably in the past decade, allowing for rapid data collection. Despite this, response rates remain lower than in other modes (Daikeler, Bošnjak and Manfreda, 2020), and the absence of general sampling frames of internet users requires the use of alternatives modes of contact, which increases fieldwork costs. Other challenges of online surveys, such as variations in internet penetration and differences in technology access and use, are likely to lessen over time. All things considered, the evidence today suggests that face-to-face interviewing is still needed in the short term, although the use of online surveys and mixed-mode approaches is on the rise and is likely to continue to increase (Schober, 2017).

The experience of the ESS with CRONOS indicates that building a cross-national probability-based panel off the back of an established survey (or perhaps recruiting directly) is feasible and provides important opportunities for the research community. CRONOS was successfully implemented in three European countries and recruited participants off the back of the main ESS (round 8). The characteristics of CRONOS panellists were not very different from those of the target population, although older respondents and those who used the internet less often were under-represented (Bottoni and Fitzgerald, 2021). Building on the pioneering experience of CRONOS, the ESS is currently planning CRONOS-2 to test the implementation of an online panel across a larger and more diverse set of countries. Its results will contribute to further developing the methodology for a cross-national web panel and will provide further open access data for researchers and the general public.

For EU-SILC, these experiments underline the importance of designing questionnaires across modes to minimise mode effects. Although mode-inherent factors, such as interviewers being present or absent, cannot always be avoided, mode measurement effects can be reduced by the design of the questionnaire. For this, the adoption of a unified mode design, in which equivalent questionnaires (e.g. question structures, wording) are developed for each mode, is recommended (Dillman, 2017; de Leeuw, Suzer-Gurtekin and Hox, 2019). This approach precludes design differences across modes (e.g. the use of grids in online/paper questionnaires versus sequential questions in face-to-face/telephone interviews) that may lead to unintended mode differences and, ultimately, threaten the validity of comparisons across groups. In addition to adopting a unified mode design, mixing modes that are most similar will restrict mode-specific errors (de Leeuw, 2018). Two characteristics are often considered when comparing modes: the degree of interviewer involvement (e.g. self-administered surveys versus interviewer-administered surveys) and the channel of communication used to present questions and provide answers (e.g. aural communication versus visual communication).

Because questionnaire design cannot reduce mode-inherent errors (e.g. how people answer sensitive questions), estimating and adjusting for unwanted mode effects is necessary (de Leeuw, 2018). In addition, treating differences found in data with some caution would be advisable, especially for more subjective measures and sensitive topics. In the longer run, if quality is to be improved, efforts to reduce the variety of modes used within and between countries should be a priority, especially with a shift towards greater use of online interviewing.

References

Bennink, M., Moors, G. and Gelissen, J. (2013), 'Exploring response differences between face-to-face and web surveys: a qualitative comparative analysis of the Dutch European Values Survey 2008', *Field Methods*, Vol. 25, No 4, pp. 319–338.

Beullens, K., Loosveldt, G., Vandenplas, C. and Stoop, I. (2018), 'Response rates in the European Social Survey: increasing, decreasing, or a matter of fieldwork efforts?', *Survey Methods: Insights from the Field* (https://surveyinsights.org/?p=9673).

Bottoni, G. and Fitzgerald, R. (2021), 'Establishing a baseline: bringing innovation to the evaluation of cross-national probability-based online panels', *Survey Research Methods*, Vol. 15, No 2, pp. 115–133.

Cernat, A. and Revilla, M. (2020), 'Moving from face-to-face to a web panel: impacts on measurement quality', *Journal of Survey Statistics and Methodology*, Vol. 9, No 4, pp. 745–763.

Christmann, P., Gummer, T., Hähnel, S. and Wolf, C. (2019), 'Does the mode matter? An experimental comparison of survey responses between face-to-face and mixed-mode surveys', paper presented at the 8th Conference of the European Survey Research Association, 15–19 July, Zagreb, Croatia.

Cornesse, C. and Bošnjak, M. (2018), 'Is there an association between survey characteristics and representativeness? A meta-analysis', *Survey Research Methods*, Vol. 12, No 1, pp. 1–13.

Daikeler, J., Bošnjak, M. and Manfreda, K. L. (2020), 'Web versus other survey modes: an updated and extended meta-analysis comparing response rates', *Journal of Survey Statistics and Methodology*, Vol. 8, No 3, pp. 513–539.

De Jong, J. (2016), 'Data collection: face-to-face surveys', in *Guidelines for Best Practice in Cross-cultural Surveys*, Survey Research Centre, Institute for Social Research, University of Michigan, Ann Arbor, MI (https://ccsg.isr.umich.edu/index.php/chapters/data-collection-chapter/face-to-face-surveys).

de Leeuw, E. D. (2018), 'Mixed-mode: past, present, and future', *Survey Research Methods,* Vol. 12, No 2, pp. 75–89.

de Leeuw, E. D., Hox, J. and Scherpenzeel, A. (2019), 'Mode effects versus question format effects: an experimental investigation of measurement error implemented in a probability-based online panel', in Lavrakas, P., Traugott, M., Kennedy, C., Holbrook, A., de Leeuw E. and West, B. (eds), *Experimental Methods in Survey Research: Techniques that combine random sampling with random assignment*, Wiley, Hoboken, NJ, pp. 151–165.

de Leeuw, E. D., Suzer-Gurtekin, Z. T. and Hox, J. J. (2019), 'The design and implementation of mixed-mode surveys', in Johnson, T. P., Pennell, B. E., Stoop, I. A. L. and Dorer, B. (eds), *Advances in Comparative Survey Methods*, Wiley, Hoboken, NJ, pp. 387–408.

DESI (Digital Economy and Society Index) (2020), *Use of Internet Services*, European Commission, Brussels (https://digital-strategy.ec.europa.eu/en/policies/desi-use-internet).

Dillman, D. A. (2017), 'The promise and challenge of pushing respondents to the web in mixed-mode surveys', *Survey Methodology*, Vol. 43, No 1, pp. 3–30.

Fitzgerald, R. and Jowell, R. (2010), 'Measurement equivalence in comparative surveys: the European Social Survey – from design to implementation and beyond', in Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P. et al. (eds), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, Wiley, Hoboken, NJ, pp. 485–495.

Jäckle, A., Roberts, C. and Lynn, P. (2006), 'Telephone versus face-to-face interviewing: mode effects on data quality and likely causes – report on Phase II of the ESS-Gallup Mixed Mode Methodology Project', *Institute for Social and Economic Research Working Papers*, No 2006-41, Institute for Social and Economic Research, University of Essex, Colchester.

Jessop, C., Bottoni G., Sommer E., Sibley E. and Fitzgerald R. (2019), 'Blueprint for comparative web panel', deliverable 7.7 of the Synergies for Europe's Research Infrastructures in the Social Sciences project funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221 (www.seriss.eu/resources/deliverables).

Krumpal, I. (2013), 'Determinants of social desirability bias in sensitive surveys: a literature review', *Quality and Quantity*, Vol. 47, pp. 2025–2047.

Leeper, T. J. (2019), 'Where have the respondents gone? Perhaps we ate them all', *Public Opinion Quarterly*, Vol. 83, pp. 280–288.

Liu, M., Conrad, F. G. and Lee, S. (2017), 'Comparing acquiescent and extreme response styles in face-to-face and web surveys', *Quality and Quantity*, Vol. 51, pp. 941–958.

Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Frederiksen, M., Ketola, K. et al. (2020), 'The European Values Study 2017: on the way to the future using mixed-modes', *European Sociological Review*, Vol. 37, No 2, pp. 330–347.

Luiten, A, Hox J. and de Leeuw, E. (2020), 'Nonresponse trends and fieldwork effort in the 21st century: results of an international study across coun-

tries and surveys', *Journal of Official Statistics*, Vol. 36, No 3, pp. 469–487.

Malnar, B. (2020), *European Social Survey Academic Impact Monitoring – Annual report 2019*, University of Ljubljana, Ljubljana.

Martin, P. (2011), 'A good mix? Mixed mode data collection and cross-national surveys', *Research & Methods*, Vol. 20, No 1, pp. 5–26.

Martin, P. and Lynn, P. (2011), 'The effect of mixed mode survey designs on simple and complex analyses', *Centre for Comparative Social Surveys Working Paper Series*, No 4, Centre for Comparative Social Surveys, City, University of London, London.

Moore, J. C., Stinson, L. and Welniak, E. J. (2000), 'Income measurement error in surveys: a review', *Journal of Official Statistics*, Vol. 16, No 4, pp. 331–361.

Olson, K. and Smyth, J. (2014), 'Accuracy of within-household selection in web and mail surveys of the general population', *Field Methods*, Vol. 25, No 1, pp. 56–69.

Olson, K. and Smyth, J. (2017), 'Within-household selection in mail surveys: explicit questions are better than cover letter instructions', *Public Opinion Quarterly*, Vol. 81, No 3, pp. 688–713.

Olson, K., Smyth, J., Horwitz, R., Keeter, S., Lesser, V., Marken, S. et al. (2020), 'Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report', *Journal of Survey Statistics and Methodology*, pp. 1–31.

ONS (Office for National Statistics) (2019), 'Internet users in the UK: 2019' (https://www.ons.gov.uk/releases/internetusersintheuk2019).

Pennell, B. E., Cibelli Hibben, K., Lyberg, L. E., Mohler, P. and Worku, G. (2017), 'A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts', in Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E. et al. (eds), *Total Survey Error in Practice*, Wiley, Hoboken, NJ, pp. 179–201.

Schober, M. F. (2017), 'The future of face-to-face interviewing', *Quality Assurance in Education*, Vol. 26, No 2, pp. 290–302.

Smith, P. (2018), 'The feasibility of a cross-European push to web survey', in Cleary, A., Stannard, J., Smith, P., de Luis Iglesias, R., Horton, S., Clemens, S. et al., *Ipsos Research Methods Centre Update: November 2018*, pp. 7–10 (https://www.ipsos.com/sites/default/files/ct/publication/documents/2018-11/ipsos-research-methods-newsletter-november-2018.pdf).

Smith, T. (2011), 'Refining the total survey error perspective', *International Journal of Public Opinion Research*, Vol. 29, No 4, pp. 464–484.

Sommer, E. (2019), 'Survey experts network meeting 4: survey fieldwork cost', deliverable 5.12 of the Synergies for Europe's Research Infrastructures in the Social Sciences project funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No 65422 (https://seriss.eu/wp-content/uploads/2020/08/SERISS-Deliverable-5.12-fieldwork-cost-workshop-report_final.pdf).

Technopolis Group (2017), *Comparative impact study of the European Social Survey (ESS) ERIC – Final report* (https://www.europeansocialsurvey.org/docs/findings/ESS-Impact-study-Final-report.pdf).

Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2010), 'A method for evaluating mode effects in mixed-mode surveys', *Public Opinion Quarterly*, Vol. 74, No 5, pp. 1027–1045.

Villar, A. and Fitzgerald, R. (2017), 'Using mixed modes in survey research: evidence from six experiments in the ESS', in Breen, M. J. (ed.), *Values and Identities in Europe – Evidence from the European Social Survey*, Routledge, New York, pp. 259–293.

West, B. and Blom, A. G. (2017), 'Explaining interviewer effects: a research synthesis', *Journal of Survey Statistics and Methodology*, Vol. 5, No 2, pp. 175–211.

Wuyts, C. and Loosveldt, G. (2020), 'Measurement of interviewer workload within the survey and an exploration of workload effects on interviewers' field efforts and performance', *Journal of Official Statistics*, Vol. 36, No 3, pp. 561–588.

28 Interviewers and their impact on survey quality: lessons for EU-SILC from the European Social Survey

Geert Loosveldt ([160])

28.1. Introduction

The interaction in a face-to-face interview is often assumed to be a 'paradigmatic' (Schaeffer and Maynard, 1996, p. 66) or 'straightforward' (Sykes and Morton Williams, 1987, p. 200) question and answer sequence: the interviewer asks a question as worded in the questionnaire and the respondent immediately reacts with an appropriate answer. With regard to closed questions, the respondent selects the response category that best describes his or her situation. In such a simple interaction, one could assume that the impact of interviewer behaviour on the respondent's answer is negligible. However, results from systematic analysis of the interaction between interviewers and respondents show that deviations from this paradigmatic sequence frequently occur (Ongena, 2005). For example, the respondent may not understand a question and therefore ask for clarification. The interviewer can then appropriately clarify the question, but it is also possible that the interviewer can solve the respondent's problem by suggesting an answer. This is a typical example of an interviewer–respondent interaction in which respondent behaviour provokes an inappropriate reaction from the interviewer. It also illustrates that interviewers can actively influence the respondents' answers. One can consider this an undesirable interviewer effect, in which the interviewer plays an active role (an active interviewer effect). Systematic interaction analysis can document this type of interviewer effect. Interviewers can also have an effect on the respondents' answers in a passive way. Certain interviewer characteristics (race, age, etc.) can provoke socially desirable answers (Anderson, Silver and Abramson, 1988).

In general, interviewer effects can be defined as the undesirable active or passive influence of an interviewer on the answers obtained. Both active and passive effects can be variable or systematic. 'Variable effects' refers to situations in which the effects differ within and between interviewers. However, although these effects are variable, it does not imply that they are harmless and negligible, as variable effects are not without consequences. They create additional noise in the data and they can also contribute to differences between the observed and the correct answers. The additional noise makes it more difficult to determine the associations between variables. The observed associations will be weaker than the true associations. Systematic coding of the interaction between interviewers and respondents can be used to assess the impact of variable interviewer effects.

Two types of systematic effects can be distinguished. When all interviewers have the same systematic effect on the answers, it results in 'pure' bias (Loosveldt and Wuyts, 2020, p. 312). This means that there are no differences between the interviewers in the way they influence the respondents' answers. For example, pure bias can occur when all the interviewers change the reference period mentioned in a particular question in the same way. However, it is also possible that there are differences between interviewers in their systematic effects. This means

([159]) Geert Loosveldt is professor emeritus at the Centre for Sociological Research at the Katholieke Universiteit Leuven. This work was supported by Net-SILC3, funded by Eurostat and coordinated by LISER. The European Commission bears no responsibility for the analyses and conclusions, which are solely those of the author. Correspondence should be addressed to Geert Loosveldt (Geert.Loosveldt@Kuleuven.be).

that the effects of each interviewer are systematic in the same direction, but that these systematic effects are not the same for all interviewers. For example, they can change the reference period in a question in a systematic way for all the respondents they interview, but these changes differ between interviewers. This creates additional variability in the answers, which can be explained by the systematic differences between interviewers rather than by differences between respondents. When the differences between interviewers in their systematic effects do not neutralise each other, they can also contribute to the overall bias.

The proportion of variance due to differences between interviewers in their systematic effects is termed interviewer variance and can be considered one form of interviewer effect that contributes to measurement error. It is common practice to evaluate these interviewer effects by means of calculating the interviewer variance. It should be noted that, with this approach, only one type of interviewer effect is documented. For example, pure bias has no influence on interviewer variance and thus cannot be captured by analysing interviewer variance.

In the following sections, we accordingly limit ourselves to the evaluation of interviewer effects that can be measured by means of interviewer variance. The basic model for the evaluation of these types of interviewer effects is presented first. This is followed by a discussion of several applications of this model – or a somewhat modified one – that can be used to assess diverse aspects of data quality. Data from several rounds of the European Social Survey (ESS) are used for our purposes. Based on the results of the assessment of interviewer effects in the ESS, a number of suggestions and recommendations are made in the conclusion for the analysis of interviewer effects in EU-SILC

28.2. The basic model for the assessment of interviewer variance

A two-level, random intercept model can be used to calculate the proportion of variance in a dependent variable that can be explained by differences between interviewers. This model takes into account the fact that the respondents are nested within interviewers. The nesting results in a two-level hierarchical data set, with the respondents at the first level and the interviewers at the second level. Given this data structure, the intercept of a simple linear model can vary across the interviewers. This variability expresses the differences between the interviewers, and the idea can be formally shown in the following model:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij};\ \beta_{0j} = \gamma_{00} + \mu_{0j}$$

We can integrate the two expressions into one equation:

$$Y_{ij} = \gamma_{00} + \mu_{0j} + \varepsilon_{ij}$$

In this model, Y_{ij} is the value of variable Y for the i-th respondent ($i = 1 \ldots N$) interviewed by the j-th interviewer ($j = 1 \ldots J$), β_{0j} is the intercept for interviewer j and ε_{ij} is the residual error term for respondent i interviewed by interviewer j. This intercept for interviewer j can be separated into a fixed (overall) intercept γ_{00} and an interviewer-specific residual error term μ_{0j}. The residual error term at interviewer level is the random intercept. A normal distribution for the residual term at respondent level and interviewer level is assumed. In these distributions, the means are zero, and the variance at respondent level and the interviewer-related error terms are equal to σ_e^2 and σ_u^2, respectively. There are significant differences between interviewers when σ_u^2 differs significantly from zero. The interviewer effect for each variable Y is estimated by means of the intra-interviewer correlation (IIC):

$$\rho_{int} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

The IIC (ρ_{int}) expresses the degree of homogeneity of the responses obtained by the same interviewer. The assumption is that the homogeneity of the responses observed for an interviewer is due to the interviewer's systematic effect on these responses. The IIC can be interpreted as the proportion of variance in Y explained by the interviewers, and this proportion can thus be considered the numerical expression of the interviewer effect (Loosveldt and Wuyts, 2020).

Just like a multiple regression analysis, one can elaborate the basic model with a number ($r = 1 \ldots R$) of respondent characteristics (X) as independent variables to explain the variance in the dependent variable Y:

$$Y_{ij} = \gamma_{00} + \sum_{r=1}^{R} \gamma_{rj} X_{rij} + \mu_{0j} + e_{ij}$$

Including the respondent characteristics in the model should improve the comparability of the respondent groups when evaluating the interviewer effects. With the variance components of the elaborated model, it is possible to calculate the IICs in the same way as in the basic model. The interviewer effects are evaluated after controlling for these respondent characteristics.

One important reason to take interviewer effects seriously is the need to estimate appropriately the standard errors of the estimates. After all, interviewer effects can be considered cluster effects with an impact on the variance estimates. This impact is expressed by the interviewer design effect, which is the increase in the variance of an estimate under the sample design assumption (e.g. simple random sample) due to interviewer effects (Biemer and Lyberg, 2003). In combination with the average number of completed interviews per interviewer ($\bar{m}$), we obtain:

$$deff_{intv} = 1 + (\bar{m} - 1)\rho_{intv}$$

In turn, the interviewer design effect determines the effective sample size (n_{eff}):

$$n_{eff} = n / deff_{int}$$

Because of clustering of the sample elements within interviewers, the effective sample size is smaller than the initial sample size (n). This expression makes it clear that a high IIC combined with a high average workload can have a serious impact on the effective sample size.

To be able to interpret the differences between the interviewers as interviewer effects, it is a prerequisite that they have interviewed comparable groups of respondents. This is the 'comparable respondent groups' assumption in the basic model used to evaluate interviewer effects. The most obvious method to realise comparable groups is random allocation. Although it is possible, for example in telephone surveys, random assignment can rarely be applied when organising face-to-face fieldwork for interviewers. One of the problems is that interviewers work in a particular area, and it is not easy to distinguish between interviewer effects and area effects. Two strategies can be applied to tackle this problem: control by design and control by model. Control by design means the use of design features when evaluating interviewer and area effects. The allocation of respondents to the interviewers and the fact that interviewers are active in certain areas are design characteristics. These design characteristics can be taken into account to evaluate interviewer effects and area effects simultaneously using a cross-classified multilevel model. In the control by model strategy, respondent characteristics are added to the models in order to control for the assumed non-comparable respondent groups within interviewers and areas. This means that interviewer effects and area effects are evaluated by elaborating the basic model with a set of relevant respondent characteristics. With this model, it is possible to calculate IICs after controlling for the respondent characteristics (Peeters, Wuyts and Loosveldt, 2019).

Most existing research concerning the relative impact of interviewer and area effects suggests that interviewer effects are more important than area effects with regard to both participation rates and responses (O'Muircheartaigh and Campanelli, 1999; Schnell and Kreuter, 2005; Peeters, Wuyts and Loosveldt, 2019).

28.3. Different types of analysis of interviewer variance

The dependent variable Y used in the basic model (presented in the previous section) is typically a numerical and substantial respondent characteristic, because an evaluation of the impact of the interviewers on the answers obtained is a key element of data quality assessment. However, the dependent variables in the analysis of interviewer variance

do not have to be limited to these substantial variables. Based on the type of dependent variable, it is possible to make a distinction between a respondent-oriented and an interviewer-oriented analysis of interviewer effects (Loosveldt, 2018).

In a respondent-oriented evaluation of interviewer effects, substantive variables can be used as the dependent variable in the basic model. It is nevertheless also possible to use response behaviour characteristics, such as willingness to participate in an interview, item non-response, response differentiation, straight-lining and other indicators of response styles. The results of this type of analysis show that response behaviour is not only a matter of how respondents perform their tasks. Interviewer effects on response behaviour characteristics illustrate how interviewers differ in the extent to which they are able to optimise the response process. In this approach, interviewers can be seen as co-responsible for response styles (Loosveldt and Beullens, 2017).

A typical example of an interviewer-oriented method is the evaluation of interviewer effects on the outcomes of the contact procedure (e.g. response rates, refusal rates and contact rates). With this type of analysis, it is possible to evaluate the differences between interviewers in terms of how successful they are in contacting sample units and persuading them to cooperate. Other task-related variables that can be used in an interviewer-oriented evaluation of interviewer effects are the speed of interviewing and other interaction characteristics (e.g. probing, and the number of appropriate and inappropriate reactions to inadequate respondent behaviour). The results of an interviewer-oriented evaluation of interviewer effects provide valuable information to assess differences in the way interviewers perform their tasks.

28.4. Interviewer effects on substantive variables

The evaluation of interviewer effects on substantive variables can be considered the most typical application of the basic model, and should be standard practice in the assessment of data quality. In this section, we start with an evaluation of interviewer effects on separate variables.

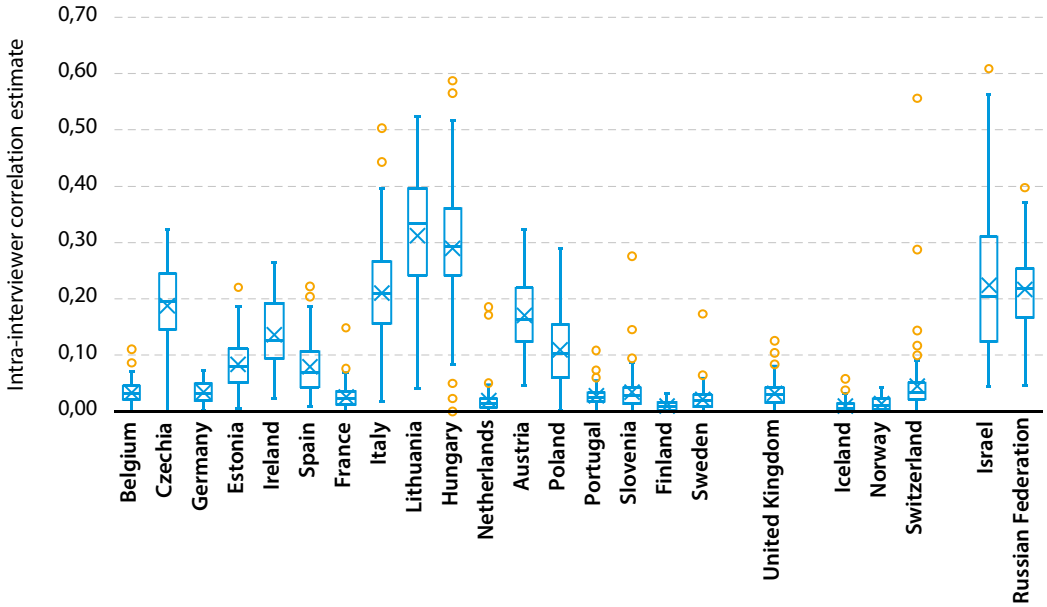
28.4.1. Separate variables

In the overall fieldwork and data quality report of ESS round 8 (Wuyts and Loosveldt, 2019), the IICs for 114 numerical and ordinal items measured on at least a 4-point scale in the main questionnaire are presented for each participating country. A random intercept model with the interviewer at the second level was used to estimate the variance components and the IICs. To control for area effects, the geographical region and self-reported degree of urbanisation of respondents' domicile were included in the models. It is clear that, with the inclusion of these two regional characteristics, only a partial control of the area effects is possible. This implies that it is still possible that differences between interviewers are related to differences between the areas in which they work. Therefore, the estimates of the interviewer effects that are presented are probably slight overestimates. Estimates for items administered by fewer than 30 interviewers or from fewer than five respondents for each interviewer were suppressed.

Using a horizontal box plot, Figure 28.1 illustrates the distribution of IICs ($n = 114$) for each participating country in ESS round 8. This clearly shows that the distribution of the IICs differs widely between countries. The average IIC ranges from 0.009 (Iceland) to 0.314 (Lithuania), with 13 countries in the 0.023–0.203 range. We observe an average IIC of 0.045 for the median country. In 11 countries (Belgium, Finland, France, Germany, Iceland, the Netherlands, Norway, Portugal, Slovenia, Sweden and the United Kingdom), none or almost none of the IICs exceed 0.10. More than half of the IICs exceed this value in nine countries (Austria, Czechia, Hungary, Ireland, Israel, Italy, Lithuania, Poland and Russia).

In the same quality report (Wuyts and Loosveldt, 2019, p. 131), the results of the analysis of interviewer variance in round 8 are compared with a similar analysis of the ESS round 7 data. It is notable that the five countries with the highest mean IICs in round 8 are also the five countries with the highest mean IICs in round 7. The order itself is also the

Figure 28.1: Distribution of the IIC, ESS round 8 (2016–2017)



NB: The IIC is estimated from a linear model with an interviewer-level random effect, controlling for geographical region and self-reported degree of urbanisation of respondents' domicile, for all numerical and ordinal items measured on at least a 4-point scale in the round 8 main questionnaire (n = 115). Countries are grouped in four categories: EU Member States, the United Kingdom, European Free Trade Association countries and other countries.

Source: Wuyts and Loosveldt (2019).

same: from highest to lowest, Lithuania, Hungary, Israel, Czechia and Austria. It should be noted that, in an earlier analysis of interviewer effects of the first six ESS rounds, Czechia, Hungary and Israel were also on the list of countries with average IICs of approximately 10 or above (Beullens and Loosveldt, 2016).

Similar large differences between countries were observed in an analysis of interviewer effects in the Programme for the International Assessment of Adult Competencies (PIAAC) survey. In this regard, it is notable that the pattern in the PIAAC results is similar to the results in the ESS analysis of interviewer effects (Keslair, 2018, p. 27). The regularities across ESS rounds and across different surveys indicate that country-specific survey practices and capacity are important factors to take into account in the analysis of this type of interviewer effects. The results also clearly indicate that, in cross-national surveys, it is necessary to control for interviewer effects; otherwise, there is a risk that comparisons between countries will be adversely affected by differences in how interviewers perform their tasks.

28.4.2. Question characteristics

The box plots in Figure 28.1 show not only large differences in IICs between countries, but also large differences between the IICs for the survey questions within a country. In some countries the variability of the IICs is fairly large (e.g. Hungary). The set of 114 items contains some from throughout the ESS round 8 questionnaire. The question with the highest median IIC across countries (= 0.15) is E8, which has a format that is typical of a survey questionnaire. It concerns having sufficient childcare services for working parents, with an 11-point response scale:

> *People have different views on what the responsibilities of governments should or should not be. For each of the tasks I read out, please tell me on*

> *a score of 0–10 how much responsibility you think governments should have. 0 means it should not be governments' responsibility at all and 10 means it should be entirely governments' responsibility, (E8) firstly, to ensure sufficient childcare services for working parents.*

The median IIC of the two items just preceding question E8 on the same list equals 0.06 (E6, standard of living for the old; E7, standard of living for the unemployed). It is not evident why the item about sufficient childcare would be more sensitive to interviewer effects than other questions.

Of the 114 items, 25 questions have a median IIC across countries that is higher than 0.08. It is striking that 7 of the 25 questions (28 %) with the highest median IICs are part of a list of 21 items used for the measurement of human values (core module H). These questions have a format familiar to respondents:

> *Now I will briefly describe some people. Please listen to each description and tell me how much each person is or is not like you. Use this card for your answer. (A) Thinking up new ideas and being creative is important to him. He likes to do things in his own original way … (I) He seeks every chance he can to have fun. It is important to him to do things that give him pleasure.*

The answer categories are 'very much like me', 'like me', 'somewhat like me', 'a little like me', 'not like me' and 'not like me at all'.

It seems that such a long list of items at the end of the questionnaire stimulates respondents' and interviewers' satisficing (Krosnick, Narayan and Smith, 1996) through the intention to obtain answers quickly. In this context, the large interviewer effects observed for a considerable proportion of the items could be an expression of interviewers' satisficing. However, the interaction between interview speed, respondents' satisficing and interviewers' satisficing needs further investigation. It was also found that the IICs for almost all 25 items are high (> 0.2) in countries with high IICs overall.

Most of the 25 questions on the list with the highest median IICs in ESS round 8 are attitudinal questions. However, this does not mean that high interviewer effects occur with only this type of question. This can be illustrated by factual questions about alcohol consumption that were part of the ESS round 7 questionnaire. Based on these questions, four alcohol consumption measurements were created: 1-year abstinence, the drinking frequency in the 12 months before the survey, the amount of alcohol consumed during weekdays and the amount of alcohol consumed during weekends. The questions on alcohol consumption are relatively difficult and sensitive, and therefore one can expect that they are prone to social desirability bias and unintended interviewer interventions, with a high risk of related interviewer effects. Social desirability bias refers to the tendency to present a favourable image of oneself that maximises social conformity and minimises negative judgement by others. The results of the interviewer variance analysis support this expectation. The median IICs over countries are 0.15 for 1-year abstinence, 0.05 for drinking frequency in the last 12 months, 0.09 for the amount of alcohol consumed during weekdays and 0.10 for the amount of alcohol consumed during weekends. In nine countries, the IIC exceeds 0.10 for at least three of the four measurements (Austria, Czechia, Estonia, Hungary, Ireland, Israel, Lithuania, Poland and Portugal). Only in Finland are the IICs close to negligible (between zero and 0.02) (Loosveldt, Wuyts and Beullens, 2018). It can be noted again that the aforementioned five countries with high IICs in ESS round 8 are also on this list.

28.4.3. Differences between countries versus differences between questions

Previous results clearly show differences in IICs between countries and questions. The question to ask is whether there is more variability in IICs between countries than between questions. To answer this, we can use the results presented in a paper about interviewer effects among older respondents (Beullens, Vandenplas and Loosveldt, 2018). In this paper – based on ESS round 7 data – IICs for 72 questions in 13 countries were modelled using a cross-classified multilevel model taking into account the differences between countries and between questions. In this model, country and question were included as random effects. This means

that one variance parameter was obtained to express the variability in the IICs between countries, and one parameter to express the variability in the IICs between questions. The results show that the variance in random intercepts (80) for countries is much larger than that for questions (12.5), indicating that the differences in IICs are much larger between countries than between survey questions. These findings were supported in research into the link between interview speed and interviewer effects, in which the same modelling approach for the IICs was applied (Vandenplas et al., 2019).

28.4.4. Relationship between variables

As already mentioned, the interviewer variance analysis captures the systematic differences between interviewers. However, the systematic influence of interviewers on the responses obtained need not be limited to the answers to individual questions. It can be assumed that the influence of an interviewer on a question is not completely independent of his or her influence on a related question. This implies that interviewer effects can also influence the relationship between items. An analysis of the interviewer effects in the first six ESS rounds (Beullens and Loosveldt, 2016) also evaluates the interviewers' impact on the relationship between 50 pairs of variables. For each pair of variables a simple regression model is specified in which one of the variables is used as the dependent variable and the other as the independent variable. For each pair the regression coefficient and the standard error are calculated twice. One analysis is based on the total covariance matrix (interviewer effects are not taken into account) and the other analysis uses the within-interviewer covariance structure. The ratio of the estimates for the two analyses (estimate based on within-interviewer covariance matrix / estimate based on the total covariance matrix) can be considered an expression of the interviewer effect. The results show that the average regression coefficients decrease when interviewer effects are taken into account, implying that the regression coefficients are overestimated when the interviewer effects are ignored. In some countries, there is a decrease in the average effect sizes of 20 % or more. The differences in effect sizes between countries are also smaller when the interviewer effects are not ignored. Not only are the regression effects sensitive to interviewer effects; the standard errors of the regression coefficients also change. Some 86 % of all the estimated regression coefficients show increased standard errors when interviewer effects are taken into account.

28.4.5. Latent constructs

The analysis of interviewer effects on the relationship between two variables can be expanded to the analysis of interviewer effects on latent variables measured by a measurement model. A measurement model can be used to measure theoretical (latent) constructs that are not directly observable but can be inferred from multiple observed indicators. The basic assumption is that relationships between the empirical indicators are caused by the latent constructs. In a measurement model the researcher specifies the relationships between latent constructs and indicators. The covariances (or correlations) between the empirical indicators are used as input to test a measurement model (hypothesised relationships between latent construct and indicators) (Kline, 2011). In general, one assumes that the impact of measurement errors on the results of measurement models for latent constructs is smaller. This also implies that the measurement of latent constructs is less sensitive to interviewer effects on the latent construct. To test this assumption, Beullens and Loosveldt (2014) used nine items from eight countries participating in round 5 of the ESS. These items represent three latent constructs (social trust, political trust and perceived threat from immigrants). The results indicate that items of the same construct are correlated at interviewer level, and that in some countries the factor loadings are smaller after removing these interviewer effects. In addition, the standard errors of the estimates in the measurement models are somewhat larger when only the within-interviewer covariance matrix is used and, as a consequence, the clustering due to interviewer effects is taken into account. The authors concluded that, although interviewer effects on correlations between variables are present in the data, the impact of these effects on the measurement models is relatively modest. This confirms the idea that the results of a meas-

urement model are less affected by measurement error. However, countries showing considerable levels of inter-interviewer variance in the univariate sense also seem to have greater interviewer effects with regard to relationships between survey variables. Therefore, interviewer variance in single items may be considered an indication of interviewer effects on the results of structural equation models or measurement models. Countries with high IICs for the separate variables will also be characterised by higher interviewer effects on the latent constructs.

28.5. The relationship between respondent characteristics and interviewer effects

Similarly to the issue of whether some questions are more sensitive to interviewer effects than others, one can question whether interviewer effects are higher for particular respondent groups. To answer the first matter, one can apply the basic interviewer variance model, presented here, to several substantive variables and compare the IICs. The solution to answering the second question is less obvious. Adding respondent characteristics as independent variables in the basic model is not an appropriate approach. In such an elaborated basic model with respondent characteristics as independent variables, the variance components are obtained after the respondent characteristics have explained part of the variance in the dependent variable. This means that the effect of the respondent characteristics on a substantive dependent variable is modelled. Therefore, with the elaborated model it is possible to evaluate the effect of these respondent characteristics on the substantive dependent variable and the interviewer effects after controlling for the respondent characteristics. In fact, the relationship between respondent characteristics and interviewer effects is not specified in this model. Therefore, the model is not capable of answering the question of whether there is a relationship between a respondent characteristic and the interviewer effects as expressed by the IICs. To answer this question, we need a model with the IICs as the dependent variable and a respondent characteristic as the independent variable.

To obtain such a model we can apply a two-step procedure (Beullens, Vandenplas and Loosveldt, 2018; Loosveldt and Wuyts, 2020). In the first step, the IICs are calculated for a number of questions (q) within the categories of a respondent characteristic (c) that we want to investigate in terms of the direct effect on the IICs. The basic model described earlier is used to calculate these conditional IICs. This results in a data set with $q \times c$ IICs. Each record in this data set contains at least the value of the IIC, the identification of the question and the category of the respondent characteristic. Other question characteristics – such as the type of question, its position in the questionnaire and the number of response categories – can be added to each record. This data set is the output of the first step of the procedure and it has a hierarchical structure: for each question, c IICs are calculated, so the IICs are nested within the questions. In the second step, a multilevel model with the IICs as the dependent variable is specified. This multilevel model takes the hierarchical structure into account. To evaluate the direct effect of the respondent characteristic, the categories of the characteristic used to calculate the conditional IICs can be added as dummy variables, with for example the first category as the reference category. The estimated regression parameter for each dummy can be interpreted as the direct effect on the IICs of belonging to that particular category. A significant parameter means that there is a significant difference between the mean IIC in the reference category and the category to which the parameter is linked. This is exactly what we need in order to answer the question of whether interviewer effects are higher for particular groups.

Using data from ESS round 7, this two-step procedure was employed in order to answer the question of whether interviewer effects are higher for older respondents (Beullens, Vandenplas and Loosveldt, 2018). It was hypothesised that interviewers play a more active role when interviewing older people, which is why increased interviewer variance could be expected among older respondents. The results support the two aspects of the hypothesis. Based

on the information in the interviewer's report that contains some assessments of response behaviour (asking for clarification, reluctance to answer, ability to answer, understanding of the questions), one can conclude that the interviewer–respondent interaction is more intensive with older respondents. There is also a significant difference in the expected direction between the mean IIC in the oldest age group (70 years and above) and that in the reference group (14–25), indicating that IICs are indeed higher for older respondents. This pattern is more pronounced in countries with lower or more moderate levels of interviewer effects, which seems to indicate that, in the high-IIC countries, the interviewer effects are generated more by a mechanism of general interviewer satisficing regardless of the response behaviour. In the countries with lower and more moderate IICs, interviewer behaviour seems influenced more by response behaviour, and interviewer effects are probably more a result of an intensified and more complex respondent–interviewer interaction.

Like the relationship between age and IICs, one can also examine the relationship between education and interviewer effects. It can be assumed that more highly educated people have more cognitive abilities and seem better equipped and trained to answer different types of questions in a survey interview. As a result, the interviewer–respondent interaction during an interview with a more highly educated respondent can be expected to run more smoothly. This means that the interviewer has to intervene less during the interaction, which in turn means that there are fewer possibilities to influence the answers. Data from 21 countries that participated in round 8 of the ESS were used to test the hypothesis that interviewer effects are smaller in the group of more highly educated respondents. The results of a preliminary analysis of information from the interviewers' report show that less-educated respondents more frequently ask for clarification, and experience greater comprehension problems, indicating the expected more complex interaction with less-educated respondents. The two-step procedure was subsequently applied to evaluate the effect of educational level on the IICs. The estimated mean IIC for the less-educated respondent group was quite high (14 %). As expected, there was a significant decrease for higher education levels. For the moderately and highly educated respondent groups, the mean IICs were, respectively, about 2 and 3 percentage points smaller (Loosveldt and Wuyts, 2020).

28.6. Conclusion and recommendations

The results of the interviewer variance-based analysis of interviewer effects using data from several rounds of the ESS clearly demonstrate that, in many participating countries, interviewers have systematic effects on the answers. These systematic effects differ from interviewer to interviewer, meaning that part of the variance in the answers obtained can be explained by the interviewers. The differences between the countries are noteworthy, as is the observation that in some countries the interviewer effects are high and persistent through different rounds. These results demonstrate that the impact of country-specific survey practices should not be underestimated and neglected in the analysis and explanation of interviewer effects in a cross-national survey. The analysis of interviewer effects in PIAAC confirms these differences between countries. The results in both cross-national surveys strongly suggest that this will also be the case in EU-SILC data sets.

Recommendation 1. As EU-SILC is an important source of comparative socioeconomic information, it seems advisable and appropriate to pay more attention to the evaluation of interviewer effects in it. It should be noted that applying the models presented to measure interviewer effects in a survey with a panel design poses additional challenges (e.g. different interviewers in different waves for the same respondent) but also offers additional possibilities (e.g. evaluation of the trend in interviewer effects for particular variables). It is necessary not only to assess and document the interviewer effects, but also to try to mitigate them. The challenge is to connect the interviewer effects with the concrete survey practices in general and specific interviewer behaviour in particular. This could include an assessment of the content and organisation of the interviewer training and the evaluation of interviewers' behaviour. The latter is highly rec-

ommended in countries always characterised by high IICs without any explanation.

Recommendation 2. An analysis of interviewer variance for the key variables of the EU-SILC questionnaire about health (e.g. self-perceived general health, access to healthcare), labour (e.g. available for work, number of hours usually worked per week in main job) and income must be part of the assessment of measurement quality. One can repeat this analysis for several survey years and panel waves and check whether there are certain questions with high intra-interviewer correlations. These analyses need not be limited to the substantive answers, but one can also check whether there are systematic differences between interviewers with regard to specific characteristics of response behaviour (e.g. item non-response, non-differentiation).

The impact of interviewer effects is not restricted to single items, but can also have an effect on the relationship between variables. However, it can be noted that the results of measurement models are less influenced by interviewer effects. Nevertheless, high IICs for single variables are a good indicator of interviewer effects on the results of measurement models.

Recommendation 3. Inspired by the analysis of interviewer effects on measurement models, one can evaluate interviewer effects on composite variables or indicators using multiple EU-SILC variables, for example indicators of persistent poverty.

It was also argued that, although the basic model for the analysis of interviewer variance is not suitable to analyse the relationship between IICs and respondent characteristics, it can be used in a two-step procedure to answer the relevant question of whether we find higher interviewer effects on particular groups of respondents. The results of the analysis of the relationship of IICs with age and education show higher IICs for older and less-educated respondents. There is empirical support that this is related to a less smooth interaction between interviewers and respondents.

Recommendation 4. One can check whether interviewer effects differ between groups of analytical interest (e.g. low income versus high income). When this is the case, it may be an indication that the interviewer–respondent interaction is less smooth in particular groups.

What are the consequences for users and producers of data sets in a cross-national survey project? Users must be aware that single items are sensitive to interviewer effects and that measurement models can be used to mitigate these effects. However, the use of measurement models requires multiple indicators, and these are not always available.

Recommendation 5. When measurement models are not possible or are unsuitable, users need to take into account in their substantive analyses the possible additional clustering within countries due to interviewers. Otherwise, when comparing countries, they risk confusing substantive differences between countries with differences in the way interviewers perform their tasks in particular countries.

Recommendation 6. For the producers of data sets, especially in a cross-national survey, the results make it clear that the evaluation of interviewer effects is an essential part of data quality assessment. Information about interviewer effects should form part of the basic metadata documentation of a survey, in addition to information about sampling design, response rates, refusal rates, etc. Unfortunately, this is rarely the case. EU-SILC could and should be an exception.

References

Anderson, B., Silver, B. and Abramson, P. (1988), 'The effects of the race of the interviewer on race-related attitudes of black respondents in SRC/CPS national election studies', *Public Opinion Quarterly*, Vol. 52, No 3, pp. 289–324.

Beullens, K. and Loosveldt, G. (2014), 'Interviewer effects on latent constructs in survey research', *Journal of Survey Statistics and Methodology,* Vol. 2, No 2, pp. 433–458.

Beullens, K. and Loosveldt, G. (2016), 'Interviewer effects in the European Social Survey', *Survey Research Methods,* Vol. 10, No 2, pp. 103–118, doi:10.18148/srm/2016.v1012.6261.

Beullens, K., Vandenplas, C. and Loosveldt, G. (2018), 'Interviewer effects among older respondents in the European Social Survey', *International Journal of Public Opinion Research,* Vol. 31, No 4, pp. 609–625, doi:10.1093/ijpor/edy031.

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality*, Wiley, New York.

Keslair, F. (2018), 'Interviewers, test-taking conditions and the quality of the PIAAC assessment', *OECD Education Working Papers*, No 191 (https://www.oecd-ilibrary.org/education/interviewers-test-taking-conditions-and-the-quality-of-the-piaac-assessment_5babb087-en).

Kline, R. (2011), *Principles and Practice of Structural Equation Modeling*, 3rd edition, Guilford Publications, New York.

Krosnick, J., Narayan, S. and Smith, W. (1996). 'Satisficing in surveys: initial evidence', in Braverman, M. T. and Slater, J. K. (eds), *Advances in Survey Research*, Jossey-Bass, San Francisco, pp. 29–44.

Loosveldt, G. (2018), 'Ask the experts: what are interviewer effects on measurement error?', *The Survey Statistician,* No 78, pp. 14–21.

Loosveldt, G. and Beullens, K. (2017), 'Interviewer effects on non-differentiation and straightlining in the European Social Survey', *Journal of Official Statistics*, Vol. 33, No 2, pp. 409–426, doi:10.1515/jos-2017-0020.

Loosveldt, G. and Wuyts, C. (2020), 'A comparison of different approaches to examining whether interviewer effects tend to vary across different subgroups of respondents', in Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F. and West, B. T. (eds), *Interviewer Effects from a Total Survey Error Perspective*, Chapman & Hall / CRC Press, Boca Raton, FL, pp. 311–322.

Loosveldt, G., Wuyts, C. and Beullens, K. (2018), 'Interviewer variance and its effects on estimates', *Quality Assurance in Education*, Vol. 26, No 2, pp. 227–242.

O'Muircheartaigh, C. and Campanelli, P. (1999), 'A multilevel exploration of the role of interviewers in survey nonresponse', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 162, No 3, pp. 437–446.

Ongena, Y. (2005), *Interviewer and Respondent Interaction in Survey Interviews*, Vrije Universiteit, Amsterdam.

Peeters, L., Wuyts, C. and Loosveldt, G. (2019), *Evaluation of interviewer and area effects in the ESS round 8*, technical report, Centre for Sociological Research, Katholieke Universiteit Leuven.

Schaeffer, N. and Maynard, D. (1996), 'From paradigm to prototype and back again: interactive aspects of cognitive processing in standardized interviews', in Schwarz, N. and Sudman, S. (eds), *Answering Questions: Methodology for determining cognitive and communicative processes in survey research*, Jossey-Bass, San Francisco, pp. 65–88.

Schnell, R. and Kreuter, F. (2005), 'Separating interviewer and sampling-point effects', *Journal of Official Statistics,* Vol. 21, No 3, pp. 389–410.

Sykes, W. and Morton-Williams, J. (1987), 'Evaluating survey questions', *Journal of Official Statistics,* Vol. 2, No 2, pp. 191–207.

Vandenplas, C., Buellens, K. and Loosveldt, G. (2019), 'Linking interview speed and interviewer effects on target variables in face-to-face surveys', *Survey Research Methods*, Vol. 13, No. 3, pp. 249–265.

Wuyts, C. and Loosveldt, G. (2019), *Quality matrix for the European Social Survey, round 8: Overall fieldwork and data quality report*, European Social Survey European Research Infrastructure Consortium, London (http://www.europeansocialsurvey.org/docs/round8/methods/ESS8_quality_matrix.pdf).

29 The importance of occupation coding quality: lessons for EU-SILC from SHARE and other international surveys

Kea G. Tijdens ([161])

29.1. Introduction

Occupation is a key variable in socioeconomic research because occupation is equally as important for an individual's identity as for their working life, earnings capacity, social life, friendships and social status. In many surveys, including EU-SILC, respondents are asked for their occupation, and occupations are classified according to the worldwide International Standard Classification of Occupations (ISCO). Although this classification is used widely and is the standard in the EU, the academic literature about occupational coding is predominantly country specific. Cross-national reliability of coding is important when drafting conclusions about occupational careers, occupational earnings, occupational entry levels, occupational certification, occupational boundaries and other labour market features in a European or wider context (e.g. Meng et al., 2020).

This chapter will discuss the implications of national findings for multicountry surveys, with the aim of outlining lessons for EU-SILC regarding the measurement of occupations in multicountry surveys. The chapter first sets the scene by describing the dynamics of job titles in labour markets and their implications for the measurement of occupations (Section 29.2). The chapter then reviews the details and logic of occupational classifications (Section 29.3). This is followed by an overview of occupational measurement by means of open versus closed survey questions (Section 29.4). Occupational coding is discussed in Section 29.5, including *ex post* coding and coding during the interview. Occupational coding is error prone, and therefore reliability is discussed, as are issues specific to occupational coding in multicountry surveys. The chapter ends with lessons for EU-SILC (Section 29.6).

([160]) Kea Tijdens is a research coordinator for WageIndicator Foundation. Previously she worked for the University of Amsterdam. Correspondence should be addressed to Kea Tijdens (k.g.tijdens@uva.nl).

29.2. Setting the scene: job titles and labour markets

The occupational distributions in national labour markets have features that are relevant for the measurement of occupations. First, in any labour market the stock of job titles is large and can easily include tens of thousands of entries, and the distribution of job holders over job titles is extremely skewed with a very long tail, as there tend to be many nurses and only a few C+ programmers. Hence, in surveys with relatively small sample sizes, the chance of including occupational titles from the long tail is limited.

Second, the stock of job titles is unstructured, because only vague boundaries exist between job titles, especially for licensed occupations. Job titles are shaped in an organisational context, and the division of work between job holders is different across large and small companies, with the former using a fine-grained division of labour within its workforce and the latter tending to use all-encompassing job titles. The division of work is mostly driven by skill level, as skilled workers are typically paid higher salaries than low-skilled workers, and this is commonly formalised in job classification schemes or pay scales. In multicountry surveys the

distribution of the labour force by company size varies and will therefore influence the details provided by survey respondents about their job titles. For example, in Poland 27 % of enterprises employ one to four employees and in Spain 34 % of enterprises employ one to four employees ([161]).

Third, the stock of job titles is unlimited, as there is no fixed list and there are many entries and exits over time. The factchecker did not exist until the early 2010s, whereas the milk vendor disappeared towards the end of the 20th century. For surveys this implies that measuring current occupations is different from measuring parents' occupations or measuring an occupational career during a respondent's lifetime. For newly emerging occupations, coding efforts should be backed by an authority that classifies these new entries. To the author's knowledge, the Office for National Statistics (ONS) in the United Kingdom is the only national statistical institute (NSI) in Europe with a permanent task group that does this, classifying new entries into its Standard Occupational Classification (SOC) system. Neither Eurostat nor the International Labour Organization (ILO) has a similar group.

Fourth, the stock of job titles is related to the skill level in the labour force. In countries with a relatively low-skilled labour force, the low-skilled occupations are classified to a high degree of detail, whereas the opposite holds for countries with a high-skilled labour force. In these countries the division of work for low-skilled occupations is limited, whereas it tends to be very detailed for high-skilled occupations.

Fifth, any labour market includes occupational segregation by gender, by ethnic group, by geography and by age. Women work in different occupations from men, ethnic groups may specialise in certain occupations, and the chance that a 25-year-old individual is working in an emerging occupation is much higher than the chance that a 65-year-old is doing so. Any occupational classification should to a certain extent allow such groups in the labour force to be classified to the same degree of detail.

These labour market dynamics require regular updates of national occupational classifications, but more so for international classifications, as the dynamics may vary across countries. Measuring occupations reliably in multicountry surveys is challenging and deserves substantial effort if it is to be done well.

29.3. Occupational classifications

Given the almost unlimited number of job titles, the responses to the survey question 'What is your job title?' need to be coded into a limited set of aggregated occupational titles, that is, an occupational classification system. For this purpose, the statistical agencies of 150 countries that are members of the ILO, a UN affiliate, have adopted ISCO to harmonise the measurement of occupations across countries, dating back to 1958, with revisions carried out in 1968, 1988 and 2008 ([162]). The ILO has a long-term record of contributing to discussions about multicountry occupational measurement (see Hoffmann et al., 1995, with references to ILO publications in the 1970s and 1980s).

ISCO is a hierarchical classification. ISCO-08, the current version, distinguishes nine major groups plus all armed forces occupations at the highest level of aggregation, breaking these down stepwise into 433 occupational units at the classification's lowest four-digit level. As was the case for its predecessors, ISCO-08 defines a job as a set of work tasks and duties performed by one person. Jobs with the same set of main tasks and duties are aggregated into the four-digit occupational units. On the basis of similarity in the tasks and duties performed, the units are grouped into three- and two-digit groups, which in turn, on the basis of the skill level, are grouped into one-digit groups (ILO, 2012).

Unfortunately, survey respondents are rarely familiar with the ISCO-08 classification and are therefore unable to provide the interviewer with a job title that directly fits into a four-digit occupational unit. Most survey respondents will provide a job title that is far more detailed, usually referred to as the five-digit classification. Five-digit occupations are

([161]) Eurostat data, Business demography by size class, 2017 [bd_9bd_sz_cl_r2].

([162]) https://www.ilo.org/public/english/bureau/stat/isco/

Table 29.1: Details and logic of ISCO-08 and stylised numbers of occupations

Detail	Logic	Number of entries (stylised)
ISCO-08 one-digit level	Skill level	10
ISCO-08 two-digit level	Similarity of tasks and duties	42
ISCO-08 three-digit level	Similarity of tasks and duties	131
ISCO-08 four-digit level	Occupational unit (similarity)	433
Occupational title (five-digit level)	Beyond the workplace (coding indexes)	1 000+
Job title	Workplace (job classifications)	10 000+
Tasks and duties	Clustered into jobs	100 000+

Source: Tijdens (2014, p. 15).

job titles that are aggregated beyond the organisational context. Table 29.1 presents the details and logic of this hierarchy and stylised numbers of entries. Although the measurement of occupations is at the five-digit level, most surveys release their data with two- or three-digit codes and rarely four-digit codes due to the risk of identifying individual respondents.

In the 1990s, the ILO undertook substantial efforts to implement ISCO-08 widely (Hoffmann et al., 1995). At that time countries used their own national occupational classifications. These classifications tend to differ cross-nationally with respect to the level of detail and the specific occupational titles included in the classifications, and in their logic (Pignatti Morano, 2014). Attempts to harmonise national occupational classifications (UN and ILO, 2014))were hampered by, among other things, the fact that ISCO does not allow skill levels of occupations to vary across different national contexts (Elias, 1997; ILO, 2012). However, countries that held their first Labour Force Survey or census in the late 1980s or in the 1990s mostly adopted ISCO or related classifications as their occupational classification, and in the early 2000s ISCO had become the standard classification in many countries. The European Commission (2009) adopted ISCO-08 as its occupational classification, and the European statistical agency Eurostat has made efforts to support European countries in developing coding indexes for their occupational data collected in the Labour Force Survey and similar surveys. Many international surveys such as the European Social Survey (ESS), European Working Conditions Survey (EWCS), European Values Survey, International Social Survey Programme and Programme for the International Assessment of Adult Competencies used ISCO-08 in recent waves. The UN (2015, pp. 112–113) considered occupation a core topic in the 2020 censuses of population and housing and recommended that countries use ISCO-08 to facilitate international comparisons, and ask respondents for their job title and a brief description of their main tasks and duties. Countries using national classifications should establish correspondence with ISCO either through double coding or through 'mapping' from the detailed groups of their national classifications to ISCO.

The ISCO-08 manual (ILO, 2012) includes extensive descriptions of tasks and duties for the four-digit occupational units. The number of tasks varies between 2 and 14 per unit, resulting in 3 264 tasks, which is an average of 7.6 tasks per unit (Tijdens, 2019a). The manual details which five-digit titles should or should not be classified into a unit, with approximately 3 500 unique five-digit titles listed (Tijdens, 2019b). Unfortunately, on its website the ILO has neither made the manual available as a coding index in a spreadsheet or other format nor provided translations into other languages, as was the case for previous ISCO versions. The European Commission has provided translations of the classification into the languages of the EU ([163]), resulting in the European multilingual classification of Skills, Competences, Qualifications and Occupations ([164]). This classification, however, is not particularly suited for coding the verbatim responses on occupation from survey data. The WageIndicator Founda-

([163]) http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2009.292.01.0031.01.ENG
([164]) https://ec.europa.eu/esco/portal/howtouse/21da6a9a-02d1-4533-8057-dea0a824a17a

tion ([165]) has made efforts to compile a so-called World Database of ISCO Occupations (WISCO) for five-digit occupations, all coded to ISCO-08, for more than 100 locales ([166]). A locale is a combination of language and country, for example en_US or es_US. More than 40 languages are available and the database is freely available online ([167]).

Finally, the reader should be aware that ISCO is a theoretical framework, based solely on desktop research in a handful of countries, with assumptions about the boundaries between occupations, the required skill levels, and the tasks and duties within occupational units. It is not supported by large-scale, multicountry survey data. Being a worldwide classification, an empirical study about the boundaries, required skill levels and tasks and duties of the 433 occupational units, and the need to keep these updated, would be beyond the financial and organisational capacities of many NSIs. However, it would be a step forward if multicountry surveys systematically asked about tasks and duties. This variable could not only help coders with coding titles that are hard to code but also be used to provide an empirical underpinning for the classification.

([165]) The WageIndicator Foundation (https://wageindicator.org/) is a non-governmental organisation that operates websites providing labour market information and conducts related projects. In 2001 a website including work-related content and a web survey about work and wages was launched in the Netherlands. In 2003, this activity was transformed into a foundation, established under Dutch law. The University of Amsterdam chairs its supervisory board. The foundation's mission is to advocate for labour market transparency for the benefit of all employers, employees and workers worldwide by sharing and comparing information on wages, labour law and careers. The foundation has gradually expanded its operations, and in 2020 it was managing national websites in 140 countries. All websites are in the national language(s). The survey has developed into a multilingual, multicountry, continuous web survey and is posted on all websites, with tools that allow face-to-face interviews to be conducted offline using tablets. It uses a multilingual occupational database to allow respondents to self-select their occupation in the web survey and when using the Salary Check tool.

([166]) The WISCO occupational database has been developed gradually since 2000, starting with a coded occupation list for the Netherlands, used for self-identification in the WageIndicator web survey on work and wages. From 2004 onwards, countries, languages, features, and coding of the database have been improved as a result of the following projects: WOrk LIfe Web (WOLIWEB) (EU-FP6, No 506590, 2004–2006), EurOccupations (EU-FP6, No 028987, 2006–2009), GLOBAL – Towards a global WageIndicator (FNV Mondiaal-Netherlands, 2008–2010), Decisions for life in non-European countries (MDG 3, Ministry of Foreign Affairs, the Netherlands, 2008–2012), WISUTIL for occupations in the energy sector (EU social dialogue programme, No VS/2010/0382, 2010–2011), WICARE for occupations in the care sector (EU social dialogue programme, No VS/2013/0404, 2013–2014), Inclusive Growth Research Infrastructure Diffusion (InGRID) (EU-FP7, No 312691, 2013–2017), EDUWORKS (Marie Curie Initial Training Network, No 608311, 2013–2017), Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) (EU Horizon 2020, No 654221, 2015–2019) and Social Sciences & Humanities Open Cloud (SSHOC) (EU Horizon 2020, No 823782, 2019–2022).

([167]) https://www.surveycodings.org/

29.4. Survey questions and answers for the measurement of occupations

29.4.1. Open format survey questions

Can survey respondents report their occupational titles? Yes, they can. For the large majority of respondents, their occupation is so much a part of their identity that they will give a valid job title. Moreover, they like to provide details about their job content. Most respondents will also understand how to communicate their job title beyond the organisational context. 'Don't know' or 'refuse to answer' responses are rare for the question on occupation.

In many surveys the occupation variable is collected through questions such as 'What is your occupation?' or 'What kind of work do you do?' (Hoffmann et al., 1995). An inventory of the occupation question in 33 international and national questionnaires revealed that an open text format was used predominantly for the occupation question (in 25 of the 33 questionnaires), that the phrasing of these open questions was different across almost all surveys, and that the words 'job title' and 'occupation' were used interchangeably, in some instances even within one question (Tijdens, 2014) ([168]). In terms of interview time efficiency, the question 'What is your job title?' from the Programme for the International Assessment of Adult Competencies seems to be

([168]) The selection criteria in this study were, first, that the questionnaire was freely available on the internet and, second, that it was available in a language understood by the author (Dutch, English, French, German). The inventory included 33 surveys carried out in Europe and the United States, including international surveys such as the ESS and the EWCS, national Labour Force Surveys carried out by NSIs, and other national surveys such as the German Socio-Economic Panel.

the optimal question. Almost all 25 surveys with an open text question included interviewer instructions such as 'Avoid vague occupational titles such as manager, clerk or farmer', 'Write in full details' and 'Describe fully, using two words or more (do not use initials or abbreviations), e.g. primary school teacher, state registered nurse, car mechanic, benefits assistant. If you are a civil servant or local government officer, please give your job title, not your grade or pay band'. The eight surveys with a closed format question provided a list or showcard, with six providing the 10 one-digit codes of ISCO and two providing a mixture of employment status, occupational titles, skill levels and supervisory position.

Only 14 of the 25 surveys with an open text question included an additional question asking for a job description, with the question wording varying across the surveys. Such a question takes a substantial amount of interview time and its purpose appears to be solely to assist the coder in case a job title cannot be coded directly, which is usually the case in less than 10 % of the records. This raises the issue of whether the same goal could be achieved in a more efficient way.

In the open response format questions, respondents choose how to report their job titles, resulting in responses at various levels of aggregation. Interviewers can control these responses by asking for additional details if needed, but in self-administered surveys such as web surveys this is not the case. Respondents tend to report their job title based on their employment contract, job classification scheme, collective labour agreement, job advertisement or a common understanding in the workplace. Detailed job titles may result in coding errors in cases of rare titles, for example lithographic stone grinder, or in uncodable titles in cases of firm-specific job titles, for example 'Appls prog I'. In contrast, some respondents may report a crude or highly aggregated title that cannot be coded at the desired level of detail, for example clerk or teacher, or that is uncodable, for example 'employee of department X', 'senior supervisor' or 'dogsbody'. Reviewing the occupational coding in two German surveys by different survey agencies, Massing et al. (2019) found that 1.8–4.9 % of the occupational titles were uncodable, but what was classified as 'uncodable' varied across coding agencies. The authors showed that between 0 % and 15 % of occupational titles were coded only at the one- to three-digit level, whereas the four-digit level was the target. For open format questions, survey organisations usually have manuals to guide interviewers. In the manual for the US Current Population Survey, for example, interviewers are instructed that single-word responses such as 'clerk', 'engineer', 'manager', 'nurse' or 'teacher' are usually too general to be coded accurately and that they should probe to obtain more specific responses (US Census Bureau, 2013).

It should be noted that the level of occupational detail provided differs depending on whose job title is being asked about. A question about a respondent's current occupation elicits a more detailed response than a question about past occupations, a partner's occupation or parental occupations, because the respondent will have less information about these occupations. Such responses mostly do not need to be coded at a great level of detail. Although the ISCO-08 classification requires farming occupations to be broken down into crop farmers, livestock farmers and mixed-crop farmers for a classification at three-digit level, interviewers for the Survey of Health, Ageing and Retirement in Europe (SHARE) indicated the need for a two-digit occupation of farmer for its question about parental occupations because respondents were not able to identify whether their parents had a crop farm, livestock farm or mixed-mode farm (Brugiavini et al., 2017).

29.4.2. Closed format survey questions

For closed format questions, a tick list provides respondents with a choice of occupational titles or occupational categories. This self-identification method can be used in all survey modes, but the size of the choice set varies widely across the modes. Telephone surveys typically ask respondents to select one of at most five highly aggregated occupational categories; otherwise, respondents would not be able to remember all the different options. Paper-based or face-to-face surveys allow a choice set of at most 50 categories when using showcards. In round 7 of the ESS in 2014, a showcard with nine categories consisting of a combina-

tion of ISCO-like categories and skill-level categories was used ([169]). A limited choice set may result in lower data quality because it is difficult to ensure consistency in how respondents fit their job titles into the highly aggregated categories, thereby introducing aggregation bias (De Vries and Ganzeboom, 2008). Web surveys allow self-identification of occupation from a choice set of thousands of titles, using text string matching. Such look-up tables do not require respondents to classify their job title in an aggregated category, thereby facilitating detailed measurement while solving the problem of aggregation bias. Increasingly, look-up tables are also used in face-to-face or telephone surveys, whereby the interviewer enters the job title, a text matching script shows the most likely titles and the interviewer selects the correct job title instantly or after asking the respondent for confirmation.

As described in the previous section, for almost two decades the author has made efforts to develop the WISCO database of occupations used in the WageIndicator web survey on work and wages and its related Salary Check tool in 140 countries. The database has been gradually expanded to 1 700 ISCO-08-coded occupational titles and more than 40 languages. The WageIndicator Foundation has compiled the WISCO database of five-digit occupations, all coded to ISCO-08 and translated for more than 100 locales, including more than 40 languages in total. With tens of thousands of users per year, the WageIndicator Foundation receives less than one email a month with a 'my job is not in your list' complaint, indicating that the look-up tables function well. Based on the SERISS project, the WISCO database could be expanded to 4 233 titles in 47 languages, all coded at ISCO-08 four-digit level (for details see Tijdens, 2019b).

This WISCO database was used in SHARE wave 6 (Brugiavini et al., 2017; see Section 29.4.3). A few SHARE countries proposed using the NSIs coding index as the choice set. However, it should be noted that coding indexes are designed for alphabetic searches and not for self-identification. Most coding indexes should be reviewed to make them user-friendly for self-identification.

([169]) See card 74, question F55 (https://www.europeansocialsurvey.org/docs/round7/fieldwork/source/ESS7_source_showcards_main_questionnaire.pdf).

29.4.3. Conclusion

For four reasons a closed format question is advantageous over an open format question. First, if designed well, the choice set will consist only of occupations at the same level of aggregation, so the data set will not include data at different ISCO levels. Second, the choice set will not include unidentifiable occupational titles. Third, costly field or office coding is not needed, and the data can be delivered in a timely fashion. Finally, in cross-country data collections and survey operations, if a consistently translated multilingual database is used, the occupations will be comparable across countries.

For four reasons a closed format question is however disadvantageous. First, for respondents or interviewers it is cognitively demanding to search for a job title, particularly when using a search tree instead of a text string matching tool. With Google Search and other search engines so widespread, however, text string searching and selecting the relevant match has become a familiar activity for a growing proportion of respondents and interviewers, although illiterate respondents will be unlikely to be able to self-identify their occupation. Second, the choice set is by definition incomplete and therefore some respondents may not find their job title or may be unable to aggregate it into an occupational title present in the choice set. A 'my occupation is not in your list' response option is advised, followed by a text field and subsequent office coding. Third, it may be time-consuming for respondents or interviewers to search for the accurate job title, although Schierholz et al. (2017) report that it took on average less than a minute to do so. Finally, in mixed-mode surveys with a combination of open and closed questions in the modes, bias effects may occur.

29.5. Occupational coding

29.5.1. *Ex post* coding of verbatim answers

For decades the verbatim answers from the occupation question in the Labour Force Survey and other surveys required office coding. For this pur-

pose, NSIs and survey agencies developed coding indexes or dictionaries. Lyberg (1982) provided a detailed description of the requirements for a coding index. Coding methods based on coding indexes are also referred to as 'dictionary approaches'. Most NSIs publish their coding indexes, but survey agencies for competitive reasons usually do not. A recent inventory of coding indexes in 99 countries (Tijdens and Kaandorp, 2018) revealed that 34 had been published, of which five used a different classification from ISCO-08, four had no coding index beyond the four digits of ISCO-08 and two referred to the coding index of another country (Germany referred to that of Austria, and Montenegro referred to that of Serbia). In a further four countries technical difficulties arose: an incomplete index in Greece, difficulties in the automatic translation of the right-to-left script in Hebrew (Israel), and technical errors in two other indexes that meant that they had to be dropped. NSIs in only 19 of the 99 countries had published a coding index beyond the ISCO-08 four-digit level. In these 19 indexes the number of five-digit entries varied from 103 in Finland to 13 314 in Austria. In Europe, two countries stand out for their large national coding indexes. The German Institute for Employment Research maintains the German Classification of Occupations; this includes approximately 24 000 job titles, including links to ISCO-08 (Paulus and Matthes, 2013). In the United Kingdom the ONS maintains its own SOC2010 dictionary with more than 28 000 entries (ONS, 2016). To keep up to date with new job titles, SOC2010 users are invited to forward information to help in the compilation of the job title index and feed into the work for the next update.

Well-known software programs for the coding of batches of occupational titles are Cascot and Cascot International ([170]). This coding is based on an approach using (semi-)automatic thresholds. Each job title in the batch is compared with coded job titles in a coding index, resulting in a certainty score for the matches. The user can set a threshold above which the input file is processed automatically and below which the user is prompted for a decision. Cascot is used by the ONS and by survey agencies in the United Kingdom. Statistics Netherlands, also using Cascot, applies a four-step coding process whereby the first step is based on job titles only. If insufficient, in a second step the job description is included for coding. If still insufficient, in a third step coding also considers industry of employment and – for managers – answers to the closed questions about managerial tasks. Here, codes are assigned according to rules that are specified beforehand. If still inconclusive, in a fourth step the job title is manually coded (Westerman and Offermans, 2014). Other surveys also use auxiliary variables for the coding. The American Community Survey uses education, age and geographical location (Cheeseman Day, 2014). For the EWCS of the European Foundation for the Improvement of Living and Working Conditions (Eurofound), the variables education, economic sector, number of co-workers, age when full-time education was completed, employment status and number of people under the supervision of respondent are used (Gallup Europe, 2010). For the validation of coding of parental occupations in the ESS, the coding quality between coders from different countries was examined using variables such as respondents' education, income and other occupations (Ganzeboom, 2014).

The scientific literature about the measurement of occupations predominantly focuses on the reliability of occupational coding, with a history dating back to the 1970s (e.g. Kalton and Stowell, 1979). Elias (1997), after reviewing UK studies evaluating the quality of occupational data through recoding, concluded that agreement rates increased with higher levels of aggregation, that is at one- or two-digit levels. At the three-digit level, agreement rates in excess of 75 % were hard to obtain. Summarising several studies, Mannetje and Kromhout (2003) found agreement rates of 44–89 % at the four-digit level and 75–97 % at the one-digit level. Schierholz et al. (2017) found a 72 % agreement rate for data from a German telephone survey. Belloni et al. (2016), recoding the SHARE data for the Netherlands, found that the incidence of miscoding was high even when comparison was performed at the one-digit level – at 28 % for the last job and 30 % for the current job. The authors found significant effects of being male, being more highly educated (only for last job) and being self-employed on coding disagreement. Conrad, Couper and Sakshaug

([170]) https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/

(2016) analysed double-coded descriptions in the Current Population Survey in the United States to identify which features are a factor in intercoder reliability. One factor was strongly related, namely the length of the occupation description: longer descriptions were less reliably coded than shorter ones. This negative relationship between answer length and coding reliability was confirmed by Massing et al. (2019). However, one-word responses, for example 'clerk' or 'manager', are usually far too general to be coded accurately, which challenged the US Census Bureau (2013, p. C4-40) to suggest useful follow-up questions.

Many NSIs and survey agencies have developed coding indexes. As in Cascot, these dictionary approaches facilitate rule-based coding schemes for batch coding, including changing all letters to lower case, removing duplicate blank spaces, correcting for misspellings, controlling for word order, removing stop words, identifying equivalents, and reducing words to their grammatical root (stemming). Thresholds can indicate which job titles are coded automatically, with the remainder to be coded manually. The American Community Survey uses a so-called occupation autocoder, which is a set of logistic regression models, data dictionaries and consistency edits ('hardcodes'), developed from around 2 million manually coded records. The autocoder assigns an occupation code if the quality score, based on agreement with clerk-coded records, is sufficiently high (Cheeseman Day, 2014). In Germany, the Institute for Employment Research applied machine-learning algorithms to 300 000+ verbatim answers, which were manually coded with high quality ([171]). Using this large-scale training data, batches of job titles were successfully coded. As Bethmann et al. (2014) phrase it: 'From a total survey error perspective this would free resources formerly spent on the reduction of processing error and offer the opportunity of employing those resources to reduce other error sources.' This 'machine learning' approach, also called statistical learning, can substantially improve the speed and accuracy of the coding, but it requires huge training sets of high-quality coded job titles. Gweon et al. (2017) explored the accuracy of three methods using data from the German General Social Survey, namely identifying duplicates from not-exact text string matching, coding of titles in a hierarchical classification structure, and a combination of these two methods. The first method was the most accurate. Although countries such as Australia, Germany and the United States have developed advanced autocoders for batch coding, based on machine learning, these are rare for other countries and other languages.

29.5.2. Coding during the interview

In 2015, for its wave 6, SHARE aimed to increase coding quality and to reduce the costs and efforts related to *ex post* coding of occupations. It introduced a 'coding-during-the-interview' approach for the face-to-face interviews. SHARE uses Blaise CAPI software, but the full choice set of the WISCO database could not be implemented in Blaise. Therefore, SHARE partner Centerdata[172] developed an external plug-in called Job Coder that could be called from Blaise (Brugiavini et al., 2017). In the computer-assisted personal interviewing (CAPI) mode, the interviewer asks for the respondent's job title and fills in the answer in open text format. A pop-up window then appears in which the Job Coder shows the matches and asks the interviewer to select the correct match or to skip when no match is found. Brugiavini et al. (2017) concluded that, except for Denmark, where technical problems were encountered, the overall performance of the Job Coder was good: 'Portugal and Sweden were the countries where the application worked better (it could code 90 % of the answers in the EP module). Luxembourg was the country where the Job Coder was less effective still coding about 70 of the cases in the EP module' (p. 69).

Similar approaches are reported by Statistics Netherlands, which from the early 2010s has used a tool for coding during interviews or for self-identification by respondents. Based on text similarity and statistical approaches, a list of best-matching titles pops up during the interview. Statistics Netherlands also developed an external plug-in for Blaise (Hacking and Willenborg, 2012). Schierholz et al.

([171]) http://fdz.iab.de/339/section.aspx/Projektdetails/k140424305

([172]) http://www.centerdata.nl/

(2017) tested coding during interviews in a computer-assisted telephone survey in Germany using input from both the training data and from the job title database. Of the 1 064 respondents, 72.4 % could find a job title directly from the database. The titles of another 13.6 % could be identified after selecting 'other occupation' – additional lists were suggested in a hierarchical setting, with a list of detailed descriptions being presented once a higher aggregated title had been chosen. For the remaining 10.0 % the algorithm did not suggest a single job category. These had to be coded manually.

29.5.3. Occupational coding in multicountry surveys

Data from multicountry surveys are typically merged data from multiple national survey agencies ([173]). In these cases, the survey operations, the question formulations and the coding procedures are probably not fully harmonised, affecting the comparability of the resulting statistics. The coding instructions are the only guidelines available for ensuring that the same job titles are coded similarly across countries. The central organisation cannot exhibit control over the coding process, particularly in cases of language discrepancies. For these reasons, occupational coding in multicountry surveys may seem like a black box, with it being impossible to know whether NSIs have classified the same occupational titles into the same category across countries.

Two approaches can be distinguished for the classification of five-digit occupations into an ISCO-08 four-digit code. The first one argues that the ILO manual and descriptions are sufficiently detailed and hence it is assumed that national coding of five-digit occupations leads to similar results across countries. This method is applied in many multicountry surveys, in which the field organisations code the occupations for their respective countries. Hence, pooling national coding indexes would be sufficient to classify occupations at ISCO-08 four-digit level. Tijdens and Kaandorp (2018) pooled the coding indexes of 19 countries, of which 18 were not in English, resulting in a database with more than 70 000 entries. Using online dictionaries and Google Translate these entries were translated into English (4.2 % could not be translated). The codes of the national indexes were checked to see whether they existed in the ISCO-08 index (10.3 % of the entries had non-existent codes). The remaining database had 60 559 records, of which 32 % had at least one duplicate title in another national coding index (19 044 records). These duplicate records could be aggregated into 5 350 occupational titles. Only 64 % of these titles had the same ISCO-08 four-digit code, 70 % had the same three-digit code, 74 % had the same two-digit code, and 80 % had the same one-digit code. In conclusion, when classifying occupations based on national coding indexes for cross-national surveys, it cannot be assumed that the same occupational titles will be consistently coded to the same occupational codes in all countries.

The second approach states that only English occupational titles should be coded, because the ISCO-08 manual is in English. Therefore, national job titles should first be translated and then coded according to their English title. For three countries (Albania, Kosovo ([174]) and Montenegro) this method was followed for Eurofound's 2010 EWCS. The verbatim responses were translated into English to facilitate central quality control (Gallup Europe, 2010). In retrospect, Ganzeboom (2014), after applying the first approach in his effort to code parental occupations in the ESS, acknowledges that it would have been much better to ask the coders to translate the occupational titles into English and then code these. Ganzeboom (2014) states that Google Translate is a big help in this respect. The WISCO occupational database also follows the second approach, departing from an English source list, consisting of the ISCO-08 coding index plus additional job titles, and using translations into other languages. The design requirements for this occupational database aimed at self-selection are detailed in Tijdens (2019b).

([173]) Note that Eurostat does not have a centralised coding system for occupations for the European Labour Force Survey. The European Labour Force Survey is merged from national Labour Force Survey data sets, which NSIs deliver to Eurostat in a defined format.

([174]) This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo declaration of independence.

Both approaches are associated with costs. In the first approach the costs are related to the coding by national survey agencies and, in addition, no multicountry quality control can be applied. In the second approach translations might be costly but central coding of the English occupations for the entire multicountry data collection is relatively cheap, especially when using a coding tool such as Cascot. In addition, once the translated database has been established it can be reused at very little cost in multiple surveys, provided Cascot is able to archive the coded verbatim responses.

29.6. Lessons for EU-SILC

29.6.1. Open text question and coding the verbatim response

EU-SILC is a multicountry survey that aims for cross-country coding validity of respondents' occupations. This section describes some approaches that could improve cross-country comparability. If EU-SILC decides to continue with its open text format question for job titles, the most efficient way forward in terms of coding would be to liaise with Cascot International or any other institute that can provide a similar coding tool for automatic batch coding. To serve all languages used in the EU-SILC surveys, the number of languages served in Cascot should be extended, which could be achieved using the WISCO database and by asking national contacts to add a set of coding rules. After the fieldwork, the national survey agencies should be urged to use this coding tool and to manually code all titles above a defined threshold. For future use the manually coded titles should be added to the coding tool. This approach would ensure that across countries the same job titles are coded similarly at the four-digit level, as described in the second approach in the previous section. However, the coding process would still be time-consuming and costly.

Alternatively, EU-SILC could collect and pool the coded verbatim responses from previous national surveys, preferably in a joint effort with organisations responsible for other European-wide surveys, such as Eurofound, SHARE, ESS and others ([175]). In a next step, this pool could be used in a dictionary approach, possibly supplemented with a machine learning approach, for batch coding of verbatim responses in new surveys. In this case it is recommended that the central survey agency manages the coding software and asks national survey agencies to code the remaining hard-to-code job titles. This approach, however, does not ensure that across countries the same job titles are coded similarly, as described in the first approach in the previous section. In the end, the coding process will be less time-consuming and less costly. Once this approach has matured, the software could also be used for coding during interviews, which is advantageous in terms of time and costs.

29.6.2. Closed question: using look-up tables with translations

Instead of further improving the verbatim coding of the open format questions, EU-SILC could explore using a closed format question with look-up tables. For web surveys, the WISCO database can be used for free ([176]). For CAPI surveys, SHARE wave 6 has shown that its Job Coder is a feasible tool to be used with Blaise. As described earlier, the Job Coder is derived from the WISCO database of occupations and includes 4 233 titles coded to ISCO-08, to be used for 99 locales including 47 languages in total ([177]). For other CAPI software, SHARE partner Centerdata should be contacted, as implementation of the Job Coder requires adaptations to the particular CAPI software in question. This approach ensures that across countries the same job titles are coded similarly, as described in the second approach in the previous section. Some surveys may prefer to use their own coding index in a closed format question instead of one multi-

([175]) It is, however, the author's experience that in multicountry surveys the verbatim responses are not always collected, or that survey organisations are not eager to share the verbatim responses with the international survey manager. This option assumes agreement among the national survey organisations that references to personal identification are to be removed, for example 'I work for McDonald's' should be replaced by 'I work for //' or 'I work for a fast food chain'.

([176]) https://www.surveycodings.org/articles/codings/occupation/

([177]) Note that some locales do not include all 4 233 titles because two English titles may be translated as one title and for a few locales the translations include a reduced number of job titles.

country database, for example because they aim for comparability over time. By doing so the validity of cross-country codes may decrease.

The current WISCO database for web surveys and the Job Coder for CAPI surveys could be improved. A simple improvement would be to include the option 'Occupation cannot be identified' or 'My job title is not in your list', followed by a text box and office coding. A second improvement could be to implement rules for the look-up table. A respondent entering the word 'clerk' will have to select from a long list of clerk occupations. Based on five-digit frequencies in survey data, for example from SHARE or WageIndicator, these look-up tables could be presented not in alphabetical order but ordered according to the highest frequencies. Here, the survey holder has to make a decision regarding the trade-off between user-friendliness and capturing rare occupational titles. A third improvement could be to implement extra questions in case respondents report highly aggregated job titles, such as clerk, operator, manager or teacher, so that the accurate ISCO-08 code can be identified. Specifically for the growing group of managers, this would be useful for accurate coding. A fourth improvement could be to explore approaches for selecting a second occupation for respondents in composite occupations. A final improvement could be to implement error messages in cases of unlikely combinations of occupation and education level or type of industry. This will be discussed in the next section.

29.6.3. One internet-based multicountry survey

The third, and most far-reaching, recommendation for EU-SILC is to change its CAPI, computer-assisted telephone interviewing and web modes into an integrated internet-based approach that can be used for web, face-to-face or telephone surveys equally ([178]). As a survey manager for the multicountry WageIndicator internet-based surveys in 140 countries and more than 40 languages for almost two decades, as well as for several other multicountry surveys, the author can strongly recommend the internet-based survey mode (see also Tijdens, 2020). Internet-based surveys can be used for both random samples and volunteer samples. This survey mode has one piece of XML or similar software to operate the survey in all countries jointly, and no country-specific software. Hence, routing and web tools are similar across countries. Such surveys also have one piece of software to allow communication between the user and the server. The surveys can be duplicated in applications for downloading onto a tablet or smartphone, allowing for face-to-face interviews. Applications do not require an internet connection during interviews because the completed interviews are stored and are uploaded at a WiFi point later on. Such surveys use one database or spreadsheet for all survey questions and answers, including their translations. The database or spreadsheet includes columns to identify routing, variables, values and labels. For any long-list question, a web survey can use an application programming interface tool to generate the national look-up table needed to measure a variable, such as occupation, education, industry, region, religion or country of birth ([179]). A survey2csv script can convert the data directly into a CSV file. Data cleaning can be partly design based ([180]) and partly rules based in the survey2csv script. The QuestAnalyser tool shows how many respondents have answered the survey at any point in time, allowing survey managers to track the response. The WageIndicator survey uses all these features.

The business model of an internet-based multicountry survey can be different from the business models of CAPI surveys carried out in multiple countries. Web surveys require initial investments in software development and the drafting of the databases needed for the surveys. A web mode generates savings because national survey agencies do not need to develop their own software, because data-cleaning costs can be reduced due to in-built dependent routing and survey2csv scripts, and because coding costs for long-list questions are no longer relevant once applica-

([178]) The phrase 'internet-based approach', rather than 'web survey', is used on purpose because 'web survey' is often associated with non-randomly sampled, volunteer surveys with non-generalisable and sometimes even poor data.

([179]) These tools are freely downloadable from surveycodings.org.

([180]) In dependent routing, a dropdown menu to identify age at first job is dependent on the respondent's age, so no respondent can start working before being born.

tion programming interfaces are used. However, the most important feature of a web survey is that it can be managed centrally. During fieldwork QuestAnalyser can monitor the daily data intake. Countries that want to ask additional questions can easily be provided with these by adding the questions for these countries only. Finally, survey quality can be improved as fine-tuned routing commands can easily be implemented in the XML script; for example, a self-employed respondent without staff will not have to answer questions about supervisory tasks in their job or about managerial occupations. This may challenge the business model of survey agencies, as this is mostly based on the number of questions in the questionnaire and not on the number of questions asked per respondent. A web survey can easily identify how many questions are asked per respondent.

A final benefit of an internet-based survey, but to the author's knowledge not yet implemented in the WageIndicator survey or other surveys, relates to design improvements for the measurement of occupations. For example, Belloni and Tijdens (2017) have modelled occupation-to-industry predictions derived from a pooled multicountry data set of 1.2 million observations with four-digit ISCO-08 occupation and two-digit NACE Rev. 2 industry variables. Based on these predictions, the most likely industries can be shown to respondents, as well as an option 'other industry' followed by a long list of industries. The need for such a feature is based on the author's observation that respondents find it more difficult to self-identify their industry than their occupation from a look-up list; they tend to respond with the name of the company or institute that they work for instead of the name of the industry. The same occupation-to-industry predictions can be used to generate error messages during survey completion in case of unlikely combinations of responses or inconsistent reporting. For example, respondents who have selected 'primary school teacher' as an occupation and then 'agriculture' as an industry will be shown the message 'Are you sure?', allowing them to correct their answer if they so wish. In a similar way, error messages could be developed for very unlikely combinations of education and occupation. A respondent with primary education only who indicates that they are a 'medical doctor' can be shown the message 'Are you sure?'. Similarly, error message scripts for inconsistent reporting can be developed for self-employed respondents without personnel selecting 'department manager' and so on. Implementing these scripts will largely contribute to the quality of the survey data and will be less costly in an internet-based survey than in a CAPI survey.

References

Belloni, M. and Tijdens, K. G. (2017), 'Occupation > industry predictions for measuring industry in surveys', deliverable 8.11 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221, doi:10.13140/RG.2.2.31328.02566.

Belloni, M., Brugiavini, A., Meschi, E. and Tijdens, K. G. (2016), 'Measurement error in occupational coding: an analysis on SHARE data', *Journal of Official Statistics*, Vol. 32, No 4, pp. 917–945, doi:10.1515/JOS-2016-0049.

Bethmann, A., Schierholz, M., Wenzig, K. and Zielonka, M. (2014), 'Automatic coding of occupations using machine learning algorithms for occupation coding in several German panel surveys', in Statistics Canada (ed.), *Beyond Traditional Survey Taking: Adapting to a changing world*, Proceedings of Statistics Canada Symposium, 2014, Quebec (http://fdz.iab.de/342/section.aspx/Publikation/k151124301; accessed 22 June 2020).

Brugiavini, A., Belloni, M., Buia, R. E. and Martens, M. (2017), 'The "job coder"', in Malter, F. and Börsch-Supan, A. (eds), *SHARE Wave 6: Panel innovations and collecting dried blood spots*, Munich Center for the Economics of Ageing, Munich, pp. 51–70 (http://www.share-project.org/uploads/tx_sharepublications/201804_SHARE-WAVE-6_MFRB.pdf; accessed 23 June 2020).

Cheeseman Day, J. (2014), 'Using an autocoder to code industry and occupation in the American Community Survey', presentation, Federal Economic Statistics Advisory Committee Meeting, 13 June 2014, Suitland, MD (https://apps.bea.gov/fesac/meetings/2014-06-13_day.pdf; accessed 23 June 2020).

Conrad, F. G., Couper, M. P. and Sakshaug, J. W. (2016), 'Classifying open-ended reports: factors affecting the reliability of occupation codes', *Journal of Official Statistics*, Vol. 32, No 1, pp. 75–92, doi:10.1515/JOS-2016-0003.

de Vries, J. and Ganzeboom, H. B. G. (2008), 'Hoe meet ik beroep? Open en gesloten vragen naar beroep toegepast in een statusverwervingsmodel (How to measure occupation? Open and closed questions about occupation applied in a status attainment model)', *Mens & Maatschappij*, Vol. 83, No 1, pp. 71–96, erratum *Mens & Maatschappij*, Vol. 83, No 2, pp. 190–191.

Elias, P. (1997), 'Occupational classification (ISCO-88) concepts, methods, reliability, validity and cross-national comparability', *Labour Market and Social Policy Occasional Papers*, No 20, Organisation for Economic Co-operation and Development, Paris, doi:10.1787/304441717388.

European Commission (2009), Commission Regulation (EC) No 1022/2009 of 29 October 2009 amending Regulations (EC) No 1738/2005, No 698/2006 and No 377/2008 as regards the International Standard Classification of Occupations (ISCO) (https://circabc.europa.eu/sd/a/9c0cea63-ac58-40f7-b2db-74e4fa8b13a4/ISCO-08%20Commission%20Regulation.pdf).

Gallup Europe (2010), *5th European Working Conditions Survey, 2010 – Technical report*, working document, European Foundation for the Improvement of Living and Working Conditions (https://www.eurofound.europa.eu/sites/default/files/ef_files/surveys/ewcs/2010/documents/technical.pdf; accessed 23 June 2020).

Ganzeboom, H. B. G. (2014), 'Coding and scaling of parental occupations in the European Social Survey', presentation, InGRID Workshop, Free University Amsterdam, 10 February 2014.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2017), 'Three methods for occupation coding based on statistical learning', *Journal of Official Statistics*, Vol. 33, No 1, pp. 101–122, doi:10.1515/JOS-2017-0006.

Hacking, W. and Willenborg, L. (2012), *Method Series Theme: Coding; interpreting short descriptions using a classification*, Statistics Netherlands, The Hague (https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/throughput/throughput/-/media/183ba8c5bd30473f8492b8adb656bfd6.ashx; accessed 23 June 2020).

Hoffmann, E., Elias, P., Embury, B. and Thomas, R. (1995), *What Kind of Work Do You Do? Data collection and processing strategies when measuring 'occupation' for statistical surveys and administrative records*, International Labour Organization Bureau of Statistics, Geneva (https://www.ilo.org/wcmsp5/groups/public/–dgreports/–stat/documents/publication/wcms_087880.pdf; accessed 23 June 2020).

ILO (2012), *International Standard Classification of Occupations 2008, Volume 1 – Structure, group definitions and correspondence tables*, International Labour Office, Geneva (https://www.ilo.org/public/english/bureau/stat/isco/isco08/; accessed 23 June 2020).

Kalton, G. and Stowell, R. (1979), 'A study of coder reliability', *Journal of the Royal Statistical Society, Series C*, Vol. 28, No 3, pp. 276–289.

Lyberg, L. (1982), *Coding of Occupation and Industry: Some experiences from Statistics Sweden*, International Labour Office, Geneva.

Mannetje, A. T. and Kromhout, H. (2003), 'The use of occupation and industry classifications in general population studies', *International Journal of Epidemiology*, Vol. 32, pp. 419–428.

Massing, N., Wasmer, M., Wolf, C. and Zuell, C. (2019), 'How standardized is occupational coding? A comparison of results from different coding agencies in Germany', *Journal of Official Statistics*, Vol. 35, No 1, pp. 167–187, doi:10.2478/JOS-2019-0008.

Meng, C., Wessling, K., Mühleck, K. and Unger, M. (2020), *Eurograduate Pilot Survey: Design and implementation of a pilot European graduate survey*, Publications Office of the European Union, Luxembourg, doi:10.2766/629271.

ONS (2016), *SOC2010 Volume 2: The structure and coding index*, last updated: 8 March 2016 (https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010/soc2010volume2thestructureandcodingindex; accessed 22 June 2020).

Paulus, W. and Matthes, B. (2013), *The German Classification of Occupations 2010 – Structure, coding and conversion table*, FDZ-Methodenreport 08/2013, Institute for Employment Research, Nurnberg (http://doku.iab.de/fdz/reporte/2013/MR_08-13_EN.pdf; accessed 23 June 2020).

Pignatti Morano, C. (2014), 'The advantages and shortcomings of different occupational categorizations', presentation, InGRID Workshop, Free University Amsterdam, 10 February 2014.

Schierholz, M., Gensicke, M., Tschersich, N. and Kreuter, F. (2017), 'Occupation coding during the interview', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 181, pp. 379–407, doi:10.1111/rssa.12297.

Tijdens, K. G. (2014), 'Reviewing the measurement and comparison of occupations across Europe', *AIAS Working Papers*, No 149, InGRID project, M21.2, Leuven, doi:10.13140/RG.2.2.10297.01123.

Tijdens, K. G. (2019a), 'Measuring job tasks by ISCO-08 occupational group', deliverable 8.6 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221, doi:10.13140/RG.2.2.17906.25286.

Tijdens, K. G. (2019b), 'Database of occupational titles, with explanatory note', deliverable 8.3 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221, doi:10.13140/RG.2.2.22100.55687.

Tijdens, K. G. (2020), *Managing Surveys: Ten lessons learned from web surveys*, WageIndicator Foundation, Amsterdam (https://wageindicator.org/documents/publicationslist/publications-2020/ten-lessons-learned-from-web surveys_20200411.pdf; accessed 23 June 2020).

Tijdens, K. G. and Kaandorp, C. S. (2018), 'Validating occupational coding indexes for use in multi-country surveys', *Survey Insights: Methods from the Field*, doi:10.13094/SMIF-2018-00007.

UN (2015), *Conference of European Statisticians recommendations for the 2020 censuses of population and housing*, United Nations Economic Commission for Europe, New York and Geneva (https://www.unece.org/fileadmin/DAM/stats/publications/2015/ECECES41_EN.pdf; accessed 22 June 2020).

UN and ILO (2014), 'National classifications', International Standard Classification of Occupations (http://www.ilo.org/public/english/bureau/stat/isco; accessed 22 June 2020).

US Census Bureau (2013), *Current Population Survey Interviewing Manual*, CPS-250, US Department of the Census, Suitland-Silver Hill, MD.

Westerman, S. and Offermans, M. (2014), 'Coding job titles from the Labour Force Survey in the Netherlands', presentation, InGRID Workshop, Free University Amsterdam, 10 February.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Improving the measurement of poverty and social exclusion in Europe: reducing non-sampling errors

EDITED BY PETER LYNN AND LARS LYBERG

Non-sampling error can seriously influence statistical estimates based on survey data. Almost any stage of the survey process can give rise to such statistical error, from initial decisions about the concepts to be measured by the survey through to the final stages of data editing. Two aspects of the implementation of data collection are particularly important: survey participation (or its counterpart, non-response) and survey measurement (the validity and accuracy of the answers provided by respondents). Data collection modes play an important role in determining the influence of these aspects. This book attempts to map out the influence of all possible types of non-sampling error on the European Union Statistics on Income and Living Conditions (EU-SILC) data, and to identify ways in which the error could be reduced. The majority of the chapters report research that formed part of the activities of the Third Network for the Analysis of EU-SILC (Net-SILC3), although there are also some guest chapters. The many practical conclusions include suggestions for improvements to documentation of procedures, improvements to guidance on survey procedures, capacity building in methods for dealing with error sources, and methodological studies, especially cross-national studies.

For more information
https://ec.europa.eu/eurostat/

Publications Office
of the European Union