The results of the MTMM experiments[1] in round 4

Melanie Revilla
Willem Saris
Irmtraud Gallhofer

RECSM
Universitat Pompeu Fabra
Barcelona

In the fourth round again 6 MTMM experiments have been done to evaluate the quality of the questions. In this report we discuss the differences in quality of the responses in the different countries. But before we are going to discuss these results we will first indicate the quality criteria that we use.

The quality criteria

In Figure 1 we show the basic response model we are using as our starting point. For details of this approach we refer to our earlier papers and to Saris and Gallhofer (2007).

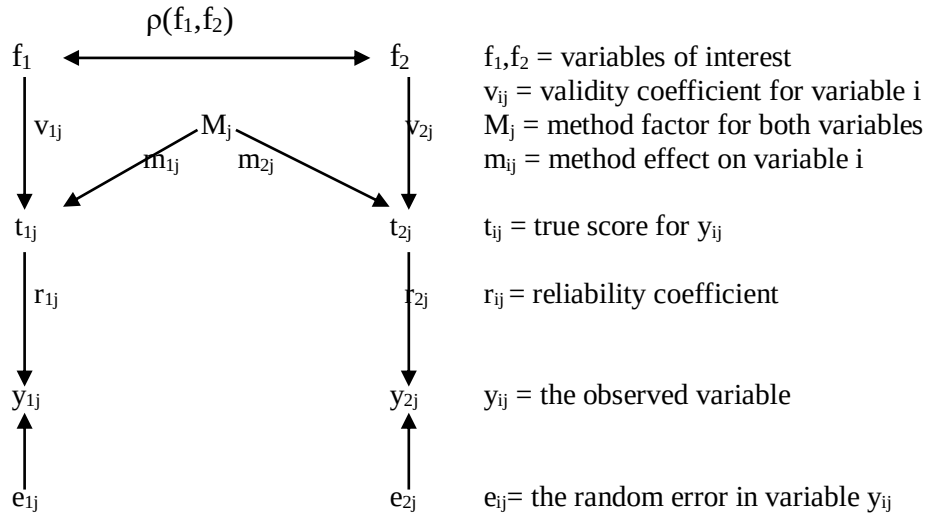

Figure 1: The response model used in the MTMM experiments

The difference between the observed response (*y*) and the so called "true score" (*t*) is random measurement error (*e*). So the coefficient *r* represents the reliability coefficient and r^2 is the reliability.

The difference between the true score and the concept by intuition (f_1) is systematic effects due to the method (*m*). So the coefficient *v* represents the true score validity coefficient and v^2 is the true score validity.

The quality of a measure (q^2) is defines as $q^2 = r^2.v^2$ and *q* is the quality coefficient.
The correlation between the variables of interest is denoted by $\rho(f_1,f_2)$

[1] We are very grateful for the work that has been done by Daniel Oberski to develop a program to collect the estimates of the parameters from a SEM output and bring them in a database.

Several remarks should be made. The first is that the correlation (r_{ij}) between two observed variables is:

$$r_{ij} = q_i \rho(f_1, f_2) q_j \quad (1)$$

A second point is that this means that this correlation between the observed variables can never be larger that the product of the quality coefficients.

A third point is that one cannot compare correlations across countries without correction for measurement errors if the measurement quality coefficients are very different across countries. This follows directly from the above equation (1).

In this paper we concentrate on the variation in measurement quality across different types of questions and across countries as far as this can be studied on the basis of the MTMM experiments of Round 4 of the ESS.

The experiments

In the fourth round the following experiments have been done:

- the time spent on different media on an average weekday ("media")
- satisfaction ("satisf")
- political orientation ("polor")
- social trust ("soctrust")
- political trust ("trustin")
- the left-right orientation ("leftright")

Each experiment contains three items measured with several methods (usually three, but sometimes, one or two items were measured with less than three methods). Table 1 gives more information about the different items and methods.

	Var.	Wording of the questions	M1	M2	M3
media	tvtot rdtot nwsptot	On an average weekday, how much time, in total: - do you spend watching television? - do you spend listening to the radio? - do you spend reading the newspapers?	8 categ. hours	Hours and min	7 categ grl
satisf	stfeco stfgov stfdem	How satisfied are you with: - the present state of the economy in NL? - the way the government is doing its job? - the way democracy works?	11 pts (extr)	11 pts (very)	5 AD
polor	gincdif freehms ?	- The government should take measures to reduce differences in income level - Gay men and lesbians should be free to live their own life as they wish - The government should ensure that all groups in society are treated equally	5 AD	5 pts	5 AD
soctrust	ppltrst pplfair	- Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? - Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?	11 pts	6 pts	2 pts

	?	- Would you say that most people deserve your trust or that only very few deserve your trust?			
trustin	trstprl trstlgl trstplc	How much do you personally trust each of the institutions: - Dutch parliament - The legal system - The police	11 pts Batt	6 pts	11 pts score
leftright	lrscale ? ?	In politics people sometimes talk of "left" and "right". - Where would you place yourself on this scale? - Where would you place the party you most like? - Where would you place the party which you most dislike?	11 pts	11 pts (fix)	11 pts= m1

Table 1: The six SB-MTMM experiments

The questions without a name but with a question mark (column "var") have only been asked in the supplementary questionnaire. It concerns in this case new questions which could not be added to the main questionnaire in order to avoid changes in the main questionnaire. The purpose of introducing these questions was in the case of political orientations and social trust to see whether the quality of the scale for this concept could be improved. The extra questions for left right were introduced to try to get an identified model so that the standard left right scale could be evaluated on quality. Unfortunately the latter effort was not successful so that we will only provide the results for five experiments.

For each experiment, the estimates are obtained from Lisrel by Maximum Likelihood estimation for multi-group analysis. In order to test if there are misspecifications, we use the JRule software (Van der Veld, Saris, Satorra, 2009) based on the procedure developed by Saris, Satorra and Van der Veld (2009). JRule has the advantage of taking into account both type I and type II errors (analysis of the power), but also to test the misspecifications at the parameter level (i.e. test if each specific parameter is misspecified, and not test the model as a whole). This leads in many cases to the introduction of corrections with respect to the general model presented earlier. Principally, the changes consist in adding a correlation between two methods when they are really similar, or fixing one of the method effects (respectively error variance) to zero if method variance (respectively error variance) is not significantly different from zero, or allowing unequal effects of one method on the different traits. In order to be able to compare results across countries, we try to make the same corrections in all countries for one specific experiment. However, this is not always possible and we have sometimes to do different corrections in the different countries.

Results

The results concentrate on the quality of the first method, i.e. the one that is included in the main questionnaire, since this is the main concern for the ESS.

Figure 2 presents first the mean quality in each of the five experiments (mean of all the traits in all the countries). The different experiments have quite similar qualities in average, around 0.7, with the satisfaction experiment being the less good.

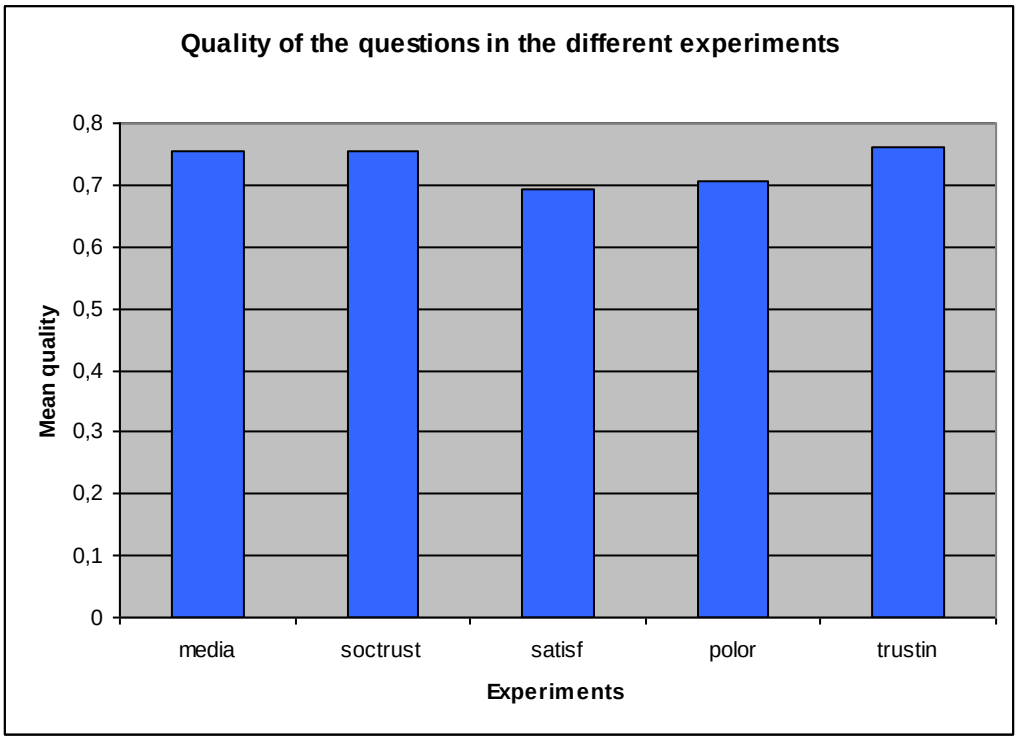

Figure 2: Quality of the questions in the different experiments

Figure 2 however is not very informative: it gives very aggregated results that hidde a lot of differences. Therefore, figure 3 separates the quality of the different traits in each experiment. Figure 3 still gives results overall countries.

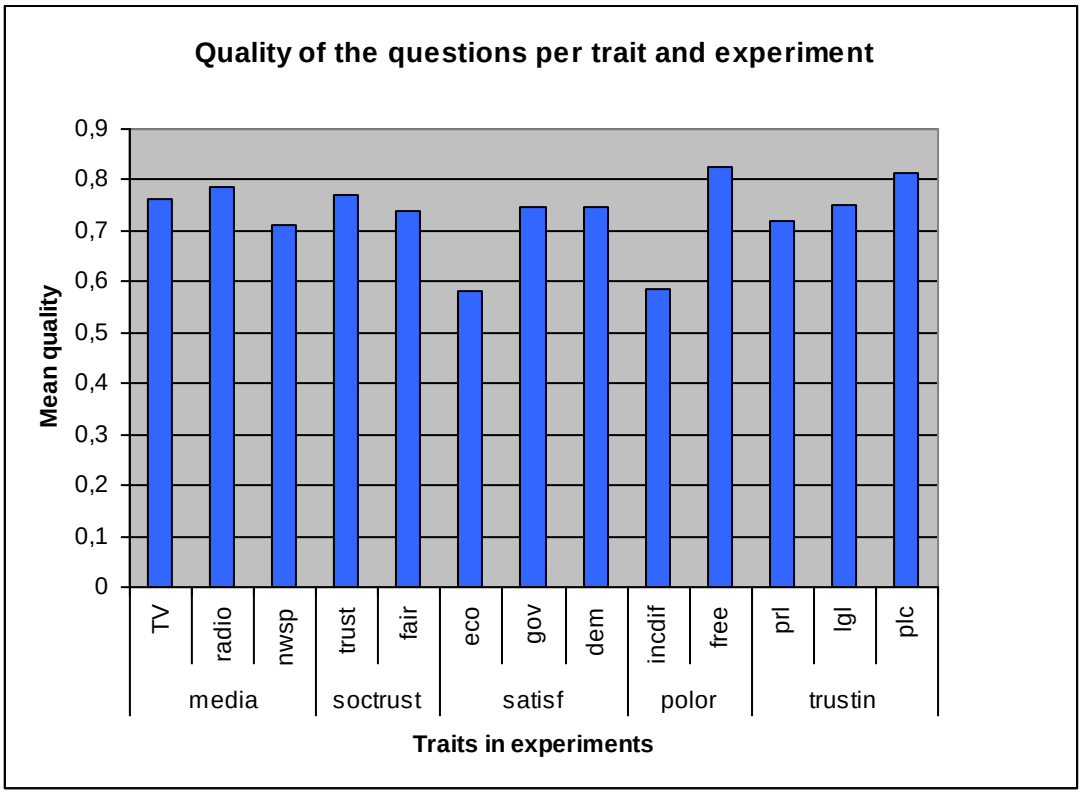

Figure 3: The mean quality of questions across all countries

The question about reading newspapers is a bit less good than the other two but not dramatic. The two social trust questions have rather high quality across all countries. The question on the economy in the set of satisfaction questions has much lower quality than the other two satisfaction questions. This may be due to the uncertainty of the economy at the time of the data collection. Also the first question in the set on political orientations concerning reduction of income inequality has a much lower quality than the question on freedom of lifestyle for homosexuals. For the items on trust in institutions the differences are not so big. It seems that especially the questions on the economic situation and policies about it have a much lower quality than the other questions.

Another way of looking at our results is to compare countries rather than experiments. Comparing the mean quality of the questions across topics for the different countries we got the results presented in figure 4. When different languages are used for the interviews in one country, we differentiated the respondents answering in the different languages. In some countries however, the samples in one language were too small to analyse them. For instance, in Finland, some respondents completed interviews in Swedish, but not enough to apply a MTMM approach on that group of respondents. In that case, we excluded them, and focused only on the Finnish questionnaires in Finland.

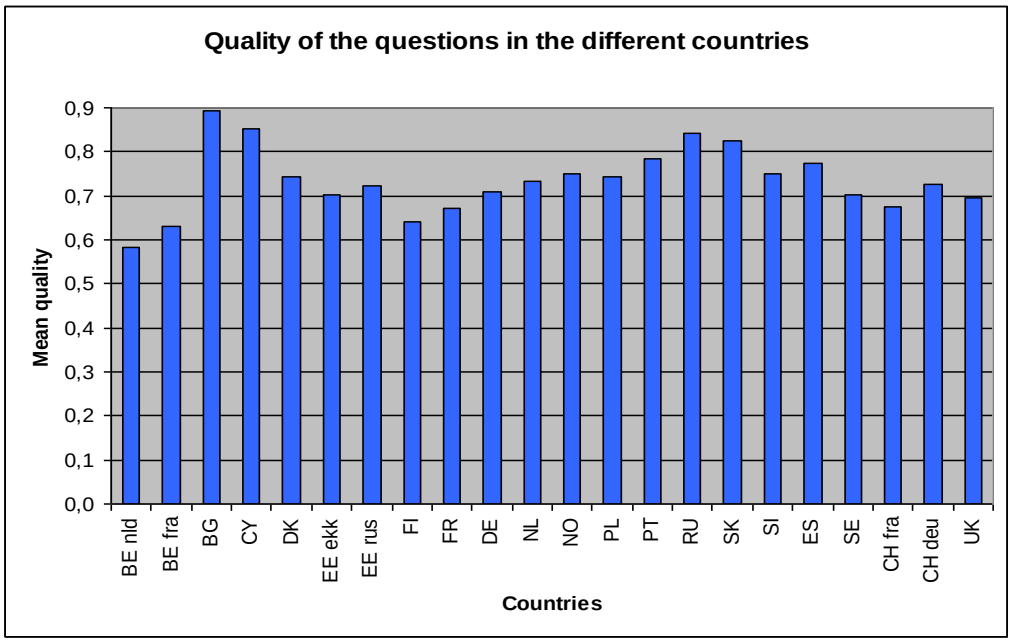

Figure 4: The mean quality across questions for the different countries and languages.

This figure shows that there are dramatic differences in quality across the countries. Starting with the best countries, we have to mention Bulgaria, Cyprus, Russia and Slovakia which all have a mean quality between .8 and .9. On the other hand, we see that the questions in Belgian Dutch as well as in Belgian French have the lowest quality and not just a bit lower but around .6. One of the reasons is that in both Belgian subsamples the media questions had extremely low quality (.5 and lower). However this is not the only topic where the quality was low in Belgium. Also for other topics the quality was below the mean over all countries.

Other countries where the mean quality score is below .7 are France and the Swiss French. This may be a consequence of deviant formulations of the questions as we have detected by coding the questions using SQP. In a later report we will come back to this issue.

Finland has also a low mean quality score even though we now used only the data of people who filled in the supplementary questionnaire on the same day as the main questionnaire. The other Scandinavian countries, which had rather low quality in round 2, have a much higher quality now that we analyze only the people who answer the questions in the supplementary questions on the same day. So this cannot be the explanation for the low score in Finland. At least one reason for the low scores is again the very low quality of the media questions in Finland.

Given the large differences in quality for the media questions we have also made the same computations but now leaving out the quality scores for the topic media. Figure 5 shows the mean quality in the different countries when excluding the media experiment and focusing on the four other ones.

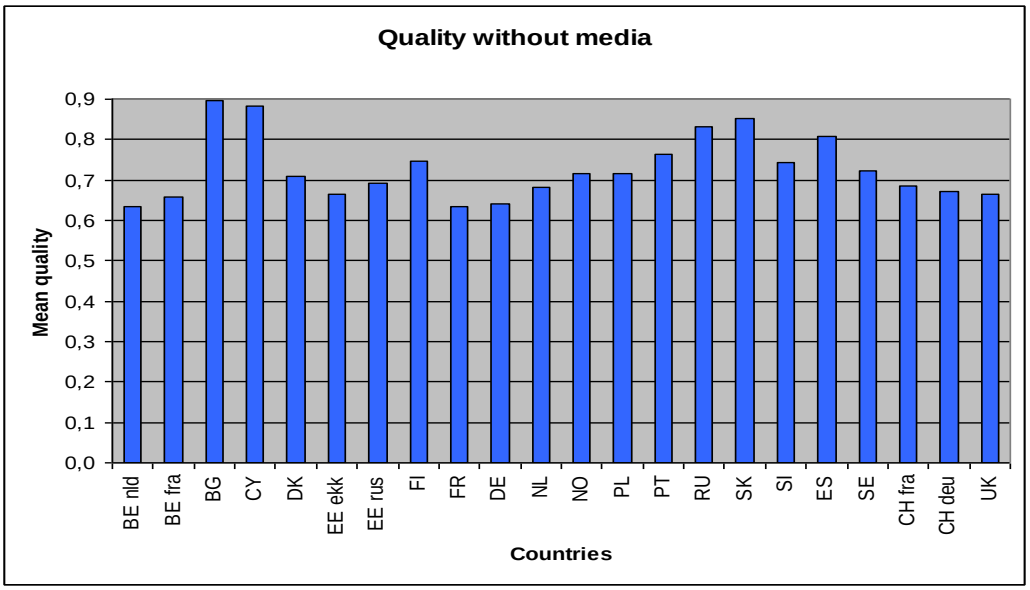

Figure 5: The mean quality across the topics, leaving out the media for all counties

Now we see that Finland has indeed a much higher score than before. So the problem was really the media questionnaire for this country. This was, however, not the case for Belgium because both Belgium samples still have the lowest quality scores of all countries/language areas.

Furthermore we see that France and Swiss French did not change much and are still relatively low with respect to quality but now they have received company of Germany and Swiss German, while also the UK and the Netherlands have quality score below .7. These last changes in the results are due to the fact that these countries had a very high quality for the media questions which compensated for lower scores elsewhere. If we take these high quality results away their mean quality clearly drops.

It is worthwhile to mention that this did not happen to the countries with the highest

quality (Bulgaria, Cyprus, Russia and Slovakia): they have also now a quality between .8 and .9. So their score is overall high. This effect can also be seen in figure 6 where we present the quality per country and experiment.

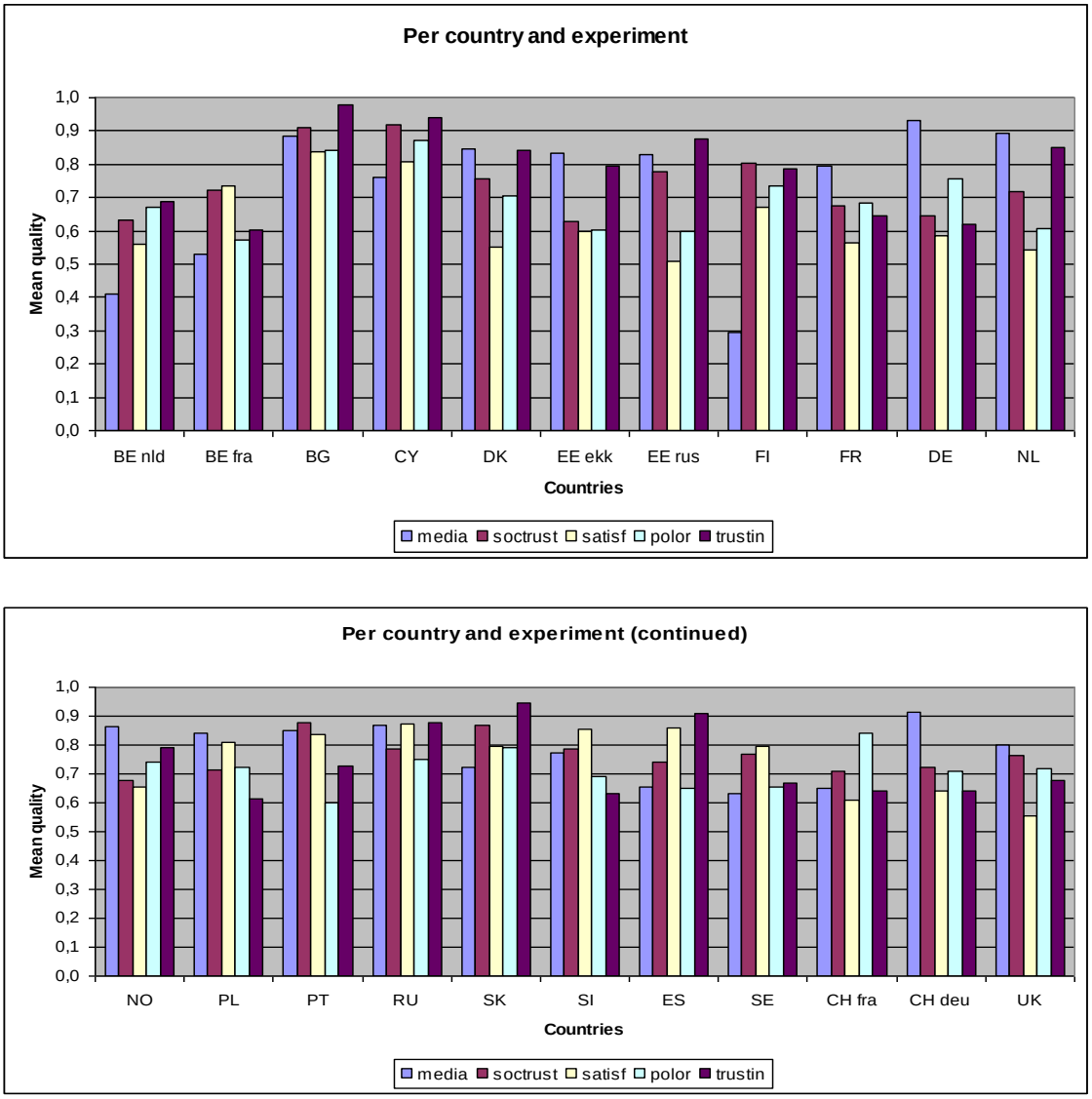

Figure 6: the mean quality of the questions per experiment and country/language area

It would lead to far to give a description of the whole figure but the trends we have mentioned above can clearly be observed.

Conclusions

These results show again how large the differences in quality of the questions are across countries. The consequences of these differences are important. This can easily be demonstrated using the results for this study. If two variables have an equal correlation of .6 in Bulgaria and in Belgium but the quality of the questions is as different as found here (.9 versus .6) then the observed correlation in Bulgaria will be .54 i.e. rather close to the real correlation but in Belgium the correlation will be .36. In general one would see these two correlations as very different and will try to explain this difference by a difference between the countries. However this makes no sense because the difference is due to difference in data quality.

On the other hand if the true correlation in Bulgaria is .4 and in Belgium .6 then the observed correlation will be .36 in Bulgaria and also in Belgium. Most people will conclude on the basis of such a result that there is no difference in correlation between these two countries but this may be true because of big differences in quality while the true correlations are very different.

This shows how important it is to have these quality estimates so that one can correct for the quality and estimate the true correlations i.e. the correlations corrected for measurement error.