Report on the MTMM experiments in the pilot studies and Proposals for Round 1 of the ESS

Willem E.Saris and Irmtraud Gallhofer

University of Amsterdam

on behalf of

The Central Co-ordinating Team of the ESS

Introduction

After some deliberations in the CCT a design for the MTMM studies in the context of the ESS pilots has been agreed upon. The chosen design of the experiments in the Netherlands (NL) and Great Britain (GB) is presented in table 1.

Table 1 The design for the meta-analysis of MTMM designs

TOPIC	MAIN QUESTIONNAIRE		SUPPLEMENTARY QUESTIONNAIRE		
	Beginning	**End**	**Beginning**	**Middle**	**End**
Media use	Open frequency		7 cat. numeric 7 cat. verbal		
Social contact	7 point high-low	7 point low-high 7 verbal cat.			
Political action	dichotomous	k/n items		22 items dichotomous	
Political efficacy	5 point agree/ disagree		5 point force choice		5 point disagree / agree
Social trust	11 point scale		5 point scale		2 points
Satisfaction with economy/demo-cracy/government	4 point bi-polar		11 point bi-polar		4 point unipolar
Schwarz values scale			importance/ feelings	Complete Schwarz scale	Importance feeling complete items

As the table shows the first three topics concerned different behaviors. The reason for this choice was that so far few MTMM experiments have been done for behavioral measures so that it was not known how good these measures were and what the best measurement procedure was. In this respect 'political action' has a special position because in that case the original set of questions was drastically reduced so that we did not know how the new version related with the old one. For the experiment an intermediate form specifying a choice of k out of n actions was formulated which is formally equivalent with the original scale.

The first three subjective variables were chosen to study the effect of different formulations on reliability and validity. The political efficacy items are used to study the quality of the forced choice format compared with the agree/disagree format. The Social Trust items are used to test the effect of the number of categories. Especially the 11 point (0 -–10 points) scale is tested. In the British pilot this is done with and without a show card. The satisfaction item have the

same purpose but now the 11 point scale was placed in the self completion part of the questionnaire.

Finally, three different versions of the Schwartz value scale have been formulated : one using only the importance part, one using only the feeling part and one using the combination. In this case the purpose is to see if these forms measured the same or not.

In the first section we will discuss the results of the experiments with respect to the different variables and make suggestions for the formulation of the questions in the first round of the ESS. In the second section we will discuss some general findings and warn against too quick generalizations. In the third section we will suggest some changes in the questionnaire for the first round given the results of the analysis of the pilot studies. In the final section a proposal for Round 1 of the ESS is provided.

1. The results of the MTMM studies

1.1. Measurement of the use of the media

The purpose of the media questions is to register the use of the media and to determine how much time people spend looking, watching or reading about news or politics and current affairs relative to the total amount of media use. In order to determine the relative amount of time calculations are necessary that require numeric measures should be obtained. Therefore in the main questionnaire numeric measures for media use (b1-b8) were specified. In the drop off form a category scale was used with verbal categories (n7,n8,n9) and one with clearly specified numeric categories (n10,n11,n12). For the exact formulation we refer to the questionnaires.

The results of this study are rather disappointing with respect to the numeric measures. First of all many responses had to be dropped for each question because people specified amounts like 10 till 24 hours per day watching TV. Secondly, looking at the responses one can see that only very few people specified the media use in more detail than in half hours. This means that the precision was not better than for the category scale with numeric labels. Thirdly, while the validities were approximately the same and high (.95 and higher) for all three forms there are considerable differences in the reliabilities between the different measures as the table below shows.

Reliabilities	TV			Radio			Newspaper		
Method	NL	GB		NL	GB		NL	GB	
Numeric	.73	.85		.82	.87		.43	.87	
Verbal 7 cat	.84	.72		.93	.89		.76	.78	
Numeric 7 cat		.94	.96		.96	.99		.83	.82

This table shows that for all three topics the reliability of the category scales with numeric labels was the best in both countries while also the reliability is quite similar in that case in both countries. The numeric measure in hours and minutes is clearly the worst in both countries for all three topics.

This result suggest the use of the numeric 7 points category scale for the measure of media use. Before we draw this conclusion let us look more precisely at these measures. In the next table we show the link between the mean proportion of time spend on different media to get political information for different subgroups with respect to political interest.

Political interest	mean proportion of time spend on politics and current affairs					
	TV		Radio		Newspaper	
Political interest	NL	GB	NL	GB	NL	GB
Very interested	.54	.41	.36	.36	.73	.46
Quite interested	.42	.27	.31	.21	.70	.18
Hardly interested	.28	.16	.12	.06	.50	.03
Not at all	.09	.08	.11	.05	.25	.08

The table shows clearly the relationship between media use for political news and political interest. Similar relationships can be found between media use for political news and interest in other organizations than the government. These results show that these questions measure indeed what they are supposed to measure.

Conclusion: We suggest to use in the main questionnaire the question with the 7 categories specifying the categories in numeric ways using half hour categories (n11-n13). In this way we get the information people can provide about media use with not too much error and the use of the media for information about politics and current affairs can also be estimated.

1.2 The measurement of frequencies of contacts

The second set of measures for behavior concern the frequency of the use of internet, the frequency of contacts with family and friends and the frequency of participating in activities of organizations or clubs. To measure the frequency of the different activities three methods have been used. The first method specifies 7 categories with numeric specification of the activity in the categories (B7, F1 and F3). These frequencies are ordered from every day (1) to never (7). The second measure (L1 to L3) uses also 7 categories but with verbal category labels and ordered from never (1) to very often (7). Finally, the third method (L4-L6) uses

again a 7 points category scale but now the category labels are again numeric and ordered from never (1) to every day (7).
All three measures have been presented to the respondents in the main questionnaire by the interviewer.

The quality of these different methods is presented below. Since the method effects are very small and nearly the same, only the reliabilities are presented.

Reliabilities	Internet		family/friends			Organizations/clubs		
Method	NL	GB	NL	GB		NL	GB	
Numeric 7 cat		.98	.95	.79	.66		.85	.80
Verbal 7 cat	.94	.97	.76	.85		.88	.93	
Numeric 7 cat		.99	.98	.80	.87		.94	.97

In both countries for all three topics the last method is the best one which is a 7 point numeric scale going from a low frequency to a high frequency. Although this is true the differences with the verbal scale going from low to high are not very large. Partially this may be due to the fact that these questions were asked very quickly after each other. However we should mention also that this result is in agreement with the result of the first set of behavioral variables where also the method with numeric labels turned out to be the best.

Before to make a recommendation we checked if the preferred measures also had the expected relationships with other variables. We expected the frequency of internet use to go together with education; the frequency of contact with family and friends should vary with the importance of family and friends and the frequency of participation in organizations should go together with membership of such organizations (we chose a sports club). The data showed that the expected relationships are indeed present in the British as well as the Dutch data as shown in the next table.

Relationship between		frequency of contact with Internet	family/ friends	organizations/clubs
and				
education	GB	.325		
	NL	.271		
Importance	GB		.119	
Friends	NL		.159	
Doing sport	GB			.322
	NL			.368

All these expected relationships are significantly different from zero even though we did not correct for measurement error. These results support the validity of these measures.

Conclusion: On the basis of these results we recommend for the first round of the ESS to use the numeric 7 point scale with the scores going from low to high in the main questionnaire.

1.3.Measurement of political action

The last set of behavioral variables concern political actions. Normally a battery with 21 items is used where people have to indicate whether they did an activity or did not. Because of the fact that it was expected that this set of questions would take too long (expected duration 5 minutes) it was decided that an alternative measure with only 8 items should be tried in the main questionnaire. Because of the fact that we were wondering what the relationship would be between the new measure and the old one we created an experiment including these two forms plus one where people had to specify how many action out of three sets they had done. In the last case the sets were created on the basis of a classification of actions as conventional, unconventional and new social movement actions. The original list of 21 items provides a possibility to check such groupings and changes in the groupings. That was the reason why Thomassen preferred the original 21 item scale. In the version with the different sets such a test is not possible any more.

First of all, we will present the results obtained from the MTMM analysis. This is done in the table below. The measures are compared for three possible subsets of political actions. These sets have been measured with three different methods. In the first method a single item (c20,c22,c24) was used for each kind of actions. In the second measure three questions asked how many of n action the respondent has done in the last 12 month (L7, L8, L9). In the third method a sum score is computed over the items which also have been specified in the sets of method 2. The results with respect to reliability and validity were as follows.

Reliability and validity of measures of political action measured in three different ways

Reliability	Conventional		Unconventional		New social movement	
Method	NL	GB	NL	GB	NL	GB
Single action	.79	.53	.17	.26	.84	.79
K/n actions	.77	.96	.99	.99	.90	.89
Sum over n	.89	.84	.73	.81	.90	.82

Validity

Single action	.99	.97	.95	.88	.99	.99
K/n actions	.83	.99	.89	.99	.86	.99
Sum over n	.86	.98	.74	.98	.85	.98

It will be clear that the single item method now used in the main questionnaire is the worst method to collect information about the three sets of political actions. This is also obvious because the other two methods use more than one item to measure different kinds of political actions. The choice between the other two measures is not so simple. In the British survey the k/n measure is clearly better than the aggregation in three sets of the original question battery. In the Netherlands there is no clear winner with respect to reliability and both have considerable method effects indicating that people react quite differently to the different methods. This is also what can be seen if the correlations between the different scales without correction for measurement error are compared with the correlations after correction for measurement error.

Correlations between the K/n measures Correlations between the sum scores per set

	Conv	unconv	new		conv	unconv	new
Conv	1.00			1.00			
Unconv	.23	1.00		.31	1.00		
New	.17	.52	1.00	.19	.43		1.00

Correlations after correction for measurement errors

	Conv	unconv	new
Con	1.00		
Unconv	,10	1.00	
New	.05	.38	1.00

These results show that the correlations are reduced by correction for measurement error which is due to the fact that the systematic effect of the method on the correlations is larger than the effect of the random errors. The matrices also show that the conventional actions are hardly correlated with the other kinds of actions while unconventional activities and new social movement activities are somewhat related.

Conclusion: Given that the single item method is not good enough the choice is between the other two. In the Netherlands the two are equally good . In Britain the k/n method is clearly better. Nevertheless, we think that a choice should be made for the original procedure. The reasons are the following:

1. The original battery takes in average only one and a half minute and not 5 minutes as expected. The k/n procedure contains only 3 questions but the questions are complex and take more than 3 minutes

2. The original battery provides the possibility in the long run to detect changes in acceptance of different political actions which would lead to different ordering of the action. This can not be detected if the subsets are formed a priori.
3. The use of the original battery gives the possibility to make comparisons through time which will not be possible if one of the new methods would be chosen.
4. The correlations with other variables measuring political interest are approximately the same and all significant.

1.4. Measurement of political efficacy

Political efficacy is measured in most of the political science studies. Nevertheless, the measurement of the concept is not at all clear. The correlations between the different variables is normally very low and the structure within the 5 questions with respect to internal and external or individual and system efficacy is not at all clear. Therefore in the pilot study a suggestion of Vetter has been followed to use two items as substitutes for older items. Furthermore, we have decided to study the quality of different formats of the questions. The first format used in the main questionnaire is the commonly used agree/disagree format with 5 categories (C3-C7). The second format presented in the drop off was a forced choice format where no statement was used (N13,N14,N15) still with 5 response categories. The last format was expected to be the same as the first but in Britain the scale was reduced to a 4 point scale and the ordering of the categories of the first set was reversed from low to high to from high to low. In the Netherlands the last set was the same as the first set (N44.N45,N46).

Let us start with the comparison of the reliabilities because the validities are equally high for all three formats. The results are presented in the table below.

Reliabilities	Complexity		Active role		Understand	
Method	NL	GB	NL	GB	NL	GB
A/D 5 cat	.65	.83	.66	.71	.69	.78
FC 5 cat	.88	.70	.94	.86	.86	.84
A/D 4/5 cat	.78	.73	.87	.82	.82	.80

This table shows that in the Netherlands the Forced choice format has a much higher reliability than the Agree/Disagree format. In Great Britain the size of this effect is much less clear but holds true for 2 out of three items. Given the low reliabilities in the first measure it is understandable why the correlations are normally so low between these items without correction for measurement error and why the structure is unclear. If the correlations are corrected for

measurement error, as we did for all 5 variables of this set the structure becomes much clearer. This can be seen in the table below for the Dutch data.

The correlations between the 5 political efficacy variables after correction for measurement error

	F1	F2	F3	F4	F5
F1	1.0				
F2	-.22	1.0			
F3	-.54	.38	1.0		
F4	-.04	.09	.10	1.0	
F5	-.11	.00	.10	.81	1.0

In this correlation matrix it is very clear that the first three items belong to one factor (individual efficacy) while the last two belong to a second factor (system efficacy) . This is in line with the theory about these measures.

In the previous report on MTMM experiments we also evaluated the relationship of the different measures with a variable one would expect to be correlated with these efficacy variables namely education. Below we have indicated the correlations with and without correction for measurement error . The results for the Dutch sample are as follows:

Variable		correlation with	
	quality	level of education	
A\D in main part		uncorrected	corrected
1	.58	-.223	-.38
2	.67	.291	.43
3	.66	,286	.43
A\D in drop off part			
1	.82	-.418	-.51
2	.91	.330	.36
3	.79	.341	.43
FC in drop off part			
1	.88	-.363	-.41
2	.92	.374	.41
3	*.87*	*.397*	*.46*

This table shows very clearly how important correction for measurement error is. Without correction for measurement error one assumes that all measures are without errors or equally good. As we have shown above both assumptions are very wrong. The differences in quality have a strong effect on the relationships between substantive variables. The weaker the measurement is the more biased the estimated correlations are. Therefore the first set of questions with the worst data quality has also the weakest relationship with the education variable and

the format using a forced choice format is the least biased because the quality of the measures is the best.

Conclusion: On the basis of these results that indicate that the forced choice has the best measurement quality and gives probably the best results within the set of variables and with other variables, we suggest to use in the main questionnaire the forced choice format of these questions.

1.5 The measurement of social trust

For the measurement of social trust three questions are used. The first concerns whether one can trust people or has to be careful. The second questions asks whether people are fair or will try to take advantage. The third concerns the question whether people try to help or only look for themselves. These judgments are asked using three different methods. The first method requires judgements on a bipolar 11 point scale (C15,C16,C17). The second requests in the drop off form judgements on a bipolar 5 point scale (N16.N17,N18). The third method asks judgements in a forced choice format with only two categories (N47,N48,N49). In the British pilot two versions of the questionnaire were used. The first version of the main questionnaire provided show cards for the 11 point scales while the second version did not. In that case the full instruction was given by the interviewer and no show card was used at all. Given this situation in the table evaluating the quality of the different measures two British questionnaires are presented and one Dutch questionnaire.

Reliability and validity of the social trust measures

	Economy			Government			Democracy		
Reliability	NL	GB1	GB2	NL	GB1	GB2	NL	GB1	GB2
11 pts cat	.89	.80	.81	.84	.77	.86	.73	.85	.79
5 pts cat	.91	.96	.90	.93	.85	.80	.90	.85	.88
2 pts cat	.88	.75	.83	.84	.79	.79	.81	.82	.83
Validity									
11pts cat	.92	.94	.93	.91	.93	.93	.88	.94	.92
5 pts cat	.96	.93	.87	.96	.91	.83	.96	.91	.86
2 pts cat	.90	.89	.92	.89	.90	.91	.89	.91	.92

This table shows that the forced choice two point scale is definitely worse in both countries than the 5 or 11 point scale with respect to validity while often the validity is the lowest i.e. the same as saying that the method effect is the largest. The evaluation of the 11 and 5 point scale is not so easy. One problem is that for an unclear reason the validity of the 5 point scale is lower in version 2 than in version 1 in Britain even though the questions were exactly the same and the

administration (self completion) is also the same. If we ignore this point the validity of the 5 point scale is systematically lower than the validity of the 11 point scale in Britain but the opposite is true for the Dutch sample. With respect to the reliability the 5 point scale is better in the Dutch sample and also most of the time in the British samples.

This result is a bit in contradiction to our expectations and previous results (Andrews 1984, Scherpenzeel and Saris 1997, Corten and Saris 2002) and the results for the measurement of satisfaction as we will see below. A possible explanation could be another factor on which the measures using the different methods vary. The most likely explanation is that the mode of administration has reduced the expected difference. In the next section we will show that self administration of the questions has a positive effect on the quality. This has also been reported in previous studies (Scherpenzeel and Saris 1997, Corten and Saris 2002). At this moment other explanations, like memory effects, can also not be excluded as we will discuss later. However, if it was only the memory effect one would expect even better results for the third method but this was not found. So we think that under normal circumstances the 11 point scale would be better and should be preferred.

Finally the table does not show a clear difference between the version 1 and 2 of the 11 point scale. The difference was that in version 1 show cards were used while this was not the case in version 2. The table shows that at least with respect to reliability and validity there are no significant differences between the two versions. So far we can not make a strong case pro or con the use of show cards. We will come to this question in the next section.

Conclusion: For the time being we would suggest to keep the measurement of social trust in the main questionnaire the same as it was in the pilot projects i.e. a 11 point scale. However, if one would like to get more assurance of the reasons of the less favorable quality of the 11 point scale the 5 point scale and the 11 point scale should switch position in the first round. That would provide the necessary evidence.

1.6. The measurement of satisfaction with the government

Three questions have been asked to evaluate the satisfaction with the results of the present government : satisfaction with the economy, satisfaction with the government it self and satisfaction with the way the economy functions in the country. It was planned that three methods would be evaluated in both countries for these measures: a 4 point scale (C33,C34,C35) going from very satisfied to very unsatisfied in the main questionnaire, a method asking an evaluation on a 11 point scale from very unsatisfied to very satisfied (N19,N20,N21) and a third

method asking an evaluation on a 4 point scale again but now going from not at all satisfied to very satisfied (N50,N51 and N52). By mistake the last method has been omitted in the Dutch pilot. Since in this case also a 11 point scale is involved we have used the same split of the British sample in two groups on the basis of the different versions of the questionnaire they got. However, in this case the 10 point scale was presented in the self administrated part so there was absolutely no difference in the way the two 10 point scales were presented to the two British samples.

The Reliability and validity of the satisfaction measures

	Economy			Government			Democracy		
Reliability	NL	GB1	GB2	NL	GB1	GB2	NL	GB1	GB2
4pts cat, h-l	.65	.72	.84	.81	.80	.87	.78	.77	.84
11 pts cat, l-h	.80	.94	.88	.91	.95	.94	.90	.93	.93
4pts cat , l-h	-	.79	.84	-	.84	.90	-	.84	.84
Validity									
4pts cat , h-l	.99	.94	.94	.99	.95	.95	.99	.94	.94
11 pts cat, l-h	.76	.90	.93	.81	.91	.94	.81	.90	.94
4pts cat l-h	-	.80	.90	-	.82	.91	-	.92	.90

As before the different samples don't show much differences with respect to the quality of the 11 point scale but they show differences for the other two methods. Why this happens requires further research.

Besides these less clear differences between the samples the table also shows very large differences in quality between the three methods. It turns out that the 11 point scale is now much better that the two other methods. This higher quality of the 11 point scale is found in the Dutch ample as well as the two British samples. The only less positive finding is that the validity of the 11 point scales for the Dutch sample is rather low compared with the British samples.

This is also an interesting case to show the difference it makes whether one is making corrections for measurement error or not. For this purpose we show the correlations between these three variable for the first (m1) and second method (m2) without correction for measurement error and after correction for measurement error for the first British and the Dutch sample.

Correlations between the satisfaction variables in the British and Dutch sample with and without correction for measurement error

Combination of	British		Dutch	
variables	no correction	corrected	no correction	corrected

	m1	m2	both	m1	m2	both
1 with 2	.43	.70	.73	.34	.54	.61
1 with 3	.28	.67	.71	.25	.51	.55
2 with 3	.48	.70	.76	.42	.64	.67

This table shows first of all the very large differences in correlations between the 4 point scale and the 11 point scale in both countries. The differences between the methods are much larger than the differences between the countries. Correction for measurement error gives for both methods the same corrected correlations which are very similar to the correlations for the 11 point scale. Using the 11 point scale or correction for measurement error shows that these variables are highly correlated. This was not at all clear if the 4 point scale would have been used.

Note that this result does not mean that there are no errors. In fact in this case the random errors and the systematic errors have nearly the same effect and therefore there is no large change in the correlations.

Conclusion: These analyses suggest very clearly that the 11 point scale should be used in the main questionnaire. In doing so even scholars who do not make corrections for measurement get results which are much closer to the proper values than with the 5 point scale.

4.7. Measurement of Schwartz human values

The items of the Schwartz value scale consist of two parts a judgment of the importance of a value and the feeling to do something in line with this value. Given that one could ask if these items are double barreled given that a reaction to two different things is asked we have suggested to test whether these assertions are indeed seen as different. In order to do so three Schwartz value statements have been decomposed in an importance assertion and a feeling assertion and these parts and the full items have been asked twice in the drop off form.

Here we present only the analysis of the Dutch data because we think that these are convincing enough. First we present the results of a standard MTMM analysis.

Reliability and validity of the Schwartz values and its components in the Dutch sample

Reliability of	item 1			item2			item3		
	1	2		1	2		1	2	
Importance	.87	.84		.94	.87		.96	.86	
Feeling		.89	.93		.89	.91		.97	.88

Complete item		.82	.91		.76	.93		.81	.97

Validity of	item 1			item2			item3		
	1	2		1	2		1	2	
Importance	.99	.92		.99	.93		.99	.93	
Feeling		.99	.97		.99	.97		.99	.97
Complete item		.99	.98		.99	.98		.99	.99

This table does not give an indication that one of the measures is better than the other.

The next test which can be done is to check if the correlations between the three items would be the same after correction for measurement error. This test is presented in the next table.

Correlation of complete item	importance	feeling	
1 with 2	.74	.70	.71
1 with 3	.55	.52	.49
2 with 3	.50	.50	.49

Without a formal test one can say that there is no substantive relevant difference between these correlations. Although this is very strong evidence of the equality of these measures. One can still argue that the variables are different but possibly the intercorrelations between these items are the same.

Given that there is hardly any difference in validity , i.e, there is hardly any method effect, one can directly test the equality of the measures by testing if the correlations between the importance and the feeling assertion for each separate item are equal to 1 using the congeneric test model of Joereskog (1971). The results of this test are presented in the table below.

Number of	assumption corr=1				corr= free			
The value item		chi2	df	n		chi2	df	corr
1	26.0	2	200		4.0	1	.85	
2	16.6	2	200		0.1	1	.91	
3	15.6	2	200		9.4	1	.95	

Formally the assumption that the correlation=1 between the importance judgment and the feeling after correction for measurement error has been rejected. However due to the high loading of these items the power of the test is very high. If this correlation is estimated it varies between .85 and .95 which most people would see as not substantially different from 1.

Conclusion: For this topic we draw two conclusions. The first is that we are inclined to accept that the importance judgement and the feeling assertion measure the same for all practical purposes. This means that it also does not matter if these two items are combined in one statement. The second is that we have no indication that the combination of the two assertions in one combined statement improves the quality of the statement with respect to reliability and validity. We have also seen that the underlying structure can be discovered very exactly by all three types of statements.

Combining the two different assertions is one statement should therefore be based on an other argument than data quality.

2. Useful results for the first round of the ESS

In the previous section we have explicitly discussed the results for specific variables. For these variables we could show that one measure was better than another. In this section we would like to make suggestions for the full questionnaire not only for the questions discussed above. In doing so we try to generalize the results from these and previous studies. However, the results can not be generalized in a simple way. This was for example clear in the case of the 11 point scale. For the variable satisfaction with the government the 11 point scale was much better than the 4 point scales it was compared with. However, for social trust the 5 point scale was slightly more reliable than the 11 point scale. There we made the argument that the last result was in contradiction with many results seen before because normally the scales with more categories are at least a bit more reliable. So other factors may have suppressed this effect. We have suggested that one possibility is the effect of the administration mode because we have seen in the past and in this study that self administration has a positive effect on the reliability of measures all other factors remaining equal. This effect has also been found in this study.
Taking into account the limited possibilities of generalization in case of a small number of cases we will try only to say something about the following choices in the ESS questionnaire:

1. mode of data collection: self administered (visual) versus interviewer administered with and without show cards
2. the number of categories used
3. Direct questions or statements in a battery
4. Ordering of the categories: low to high or high to low

We will discuss the different topics in sequence starting with the mode of data collection. In discussing these results we are not making a distinction between the two countries because the results were in general the same. Even in the final multivariate analysis using regression the language had no significant effect on the reliability and the validity.

2.1 Mode of data collection
In this study mainly three modes of data collection have been used:

1. self administered interviews (visual)
2. interviewer administered interviews with show cards and
3. interviewer administered interviews without show cards

Let us first look if these data collection modes in the pilot studies led to differences in reliability and validity.

Mean reliability and validity across modes of data collection

	Mean reliability	mean validity	N
Administered by			
Self	.86	.93	87
Interviewer with show cards	.84	.97	32
Interviewer without show cards	.79	.97	40

The most remarkable result is certainly the low reliability of the interviewer administered questionnaire without show cards compared with the other two forms. In this specific study one problem is that the self administered form is always the second or third method but this does not hold true for the interviewer administered form using show cards. An other problem is that other factors can be confounded with the mode. One of them is the use of open questions for behavior. This is done in two forms. First of all, numeric values have been asked in the main questionnaire and not in the drop off form. Secondly, yes no questions with respect to political action have been asked in the main questionnaire and not in the drop off form. Both had rather low reliability and certainly contributed to the low score. However , we have also seen that 4 point scales for satisfaction had also very low reliability when the answers were only presented orally. Finally we have also seen an exception where the use of show cards had hardly any effect, namely in case of the experiment in Great Britain with the social trust items.

This overview of the present information shows that the conclusions are not so clear yet. In a multivariate analysis presented below we will show that the administration mode causes a lot of difference in the effects of the other variables on the quality indicators while we have already seen that the data collection mode can have a considerables effect on the quality. This is also in agreement with previous results of MTMM experiments. Our most recent meta analysis shows that visual presentation has a positive effect on the quality of the measures (Corten and Saris 2002).

Conclusion: On the basis of the results obtained in this study and previous meta - analyses we would suggest that the use of show cards or self administration can improve the quality of the measures certainly with respect to reliability. In case of completely oral presentations the results suggest that one can get less good quality but that is not always the case.

2.2 The number of categories used

The second choice concerns the number of categories to be used. In this respect the results are rather clear as can be seen in the table below:

The mean reliability and validity related with the number of categories in the scale

Administered by	an Interviewer			respondent self			total		
Number of categories		rel	val	n	rel	val	n	rel	val
n									
2		.56	.96	6	.82	.90	9	.72	.93	15
4		.79	.96	9	.84	.86	6	.81	.92	15
5		.72	.97	6	.85	.94	20	.82	.95	26
6		.89	.99	12	.88	.96	25	.89	.97	37
7		.87	.99	18	.88	.95	11	.87	.98	30
11		.82	.92	7	.91	.88	9	.87	.90	16

In the table a distinction is made between the interviewer administered part and the self administered part. Between the two parts there is a clear distinction in absolute level of quality but in both parts we see a clear increase in quality if the number of categories increases. The only exception is possibly the 11 point scale with respect to reliability but this might be partially due to the limited number of measures of this type included in the experiment. The drop in validity for this scale is probably more realistic because one can expect more method effect and consequently less validity for scales with a larger number of response categories due to the different ways people interpret the scale. This effect is called variation in response functions (Saris 1986). This problem can partially be reduced by the use of fixed reference points.

The results found with respect to the effect of the number of categories are in complete agreement with the results obtained in previous results reported by Andrews (1984, Költringer 1995, Alwin 1997, Scherpenzeel and Saris 1997, Corten and Saris 2002).

Conclusion: These results suggest that one should use as many categories as possible but when the number of categories becomes larger it is more important to introduce fixed reference points.

2.3 Direct questions or statements in a battery

One has a choice between asking direct questions or agree - disagree questions with respect to statements. Mostly the last kind is questions are presented in batteries which have the advantage that the questions and the response categories have to be presented only once. Also writing such survey items is easier than writing direct questions. The question is , however, if this format also leads to better quality of the responses. In the table below we have summarized the results for this study.

Mean reliability and validity related to the use of batteries in different modes of data collection

	Administered by Interviewer			respondent self		
	Rel	val	n	rel	val	n
Direct question	.83	.96	48	.86	.92	53
Use of statements	.77	.98	24	.87	.95	34

The most remarkable result is obtained when statements are used in interviewer administered interviews. In that case it seems that the reliability is much lower than for direct questions and in self administered questionnaires. If one takes into account that the questions with respect to behavior are also included in this category of questions, this result is even more remarkable.
In previous research we have not found such a difference for direct questions and questions using statements. Therefore one should consider the possibility of an effect which is specific for the questions in the ESS

Conclusions: In order to avoid reduction of quality one should try to avoid as much as possible the use of statements. We would like to add that this rule should not be applied so strictly that only direct questions are used. Using different formats in the questionnaire keeps the respondents awake and avoids serious method effects.

2.4 Ordering of the categories: low to high or high to low

The last choice we would like to discuss concerns the choice to present the categories from low to high or from high to low. In our previous studies a positive effect on quality was found when the response categories were presented from low to high. In this study we found the opposite as can be seen in the table below.

This table shows that in this study the scales with the ordering of the categories from high to low received a better score on reliability and validity than the scales

with an ordering from low to high. The difference is not large but looks systematic.

Mean reliability and validity for different orderings of the categories and the mode of data collection

	Adminstered by					
	The interviewer			respondent self		
Categories	rel	val	n	rel	val	n
Ordered from						
Low to high	.80	.96	42	.85	.92	63
High to low	.84	.98	30	.88	.96	24

However, it is not impossible that this variation goes together with other characteric which causes this unexpected result. In this respect I would like to mention that in the previous section we discussed two examples where explicitly for the same topic the ordering of the response categories were varied to see what the effect was. In both cases (frequency of contacts and satisfaction with the government) the items with the ordering from low to high had better quality. So it seems that this result should come from other characteristics of the items than the ordering of the categories. In the following multivariate analysis we will check his possibility.

Conclusion: On the basis of the data of this experiment mentioned above no clear suggestion can be done. So far the results look contradictory.

2.5 A multivariate analysis of the data from the pilot study

So far we have analyzed the effects of the different variables on the reliability and validity, mainly using one or two variables. At several occasions we have mentioned that the relationships are more complex due to relationships with other variables.

Therefore we are now presenting a multivariate analysis using regression analysis to determine whether the different variables have also an effect on the reliability controlling for more variables. We concentrate in this analysis on the reliability because the above given tables show that the chosen variables mainly influence the reliability. After the analysis of the reliability we will try to find some other variables which have more influence on the validity.

In the analysis for reliability we did not only include the question characteristics discussed before but also some characteristics which have to do with the design of the study. One of these characteristics is the language used in the questionnaire (English =0 or Dutch =1). Another is the distance between the questions for the same trait. The last variable is included because the smaller the distance the larger the correlations will be and consequently the reliability.

Another variable which has been introduced is the variable specifying if an absolute or comparative judgment was asked. This variable has more effect than a variable like 'objective versus subjective' or 'frequency versus category scale' because the 'absolute/comparative' variable covers both where questions asking objective information and/or frequencies are coded as absolute while the subjective variables are split in absolute and comparative judgments. If this variable is introduced with the variable 'behavior or not' the latter has no significant effect and is therefore omitted.

In this regression analysis we have done a separate analysis for questions in the main questionnaire using an oral presentation by the interviewer and questions in the drop off form using visual presentation by the respondent self. The results of the regression analyses are presented below.

Linear regression where the dependent variable is the reliability of the different questions

	Administration by Interviewer (71 questions)	Respondent (86 questions)
Intercept	.768*	.807*
Language	-.055	.022
Distance	.000	-.000
In a battery	-.095*	-.049*
Gradation	.170*	.051*
Ordering cat.	.007	-.016
Comparative	.143	.084*
No show cards	-.039	-
Adjusted R^2	.39	.09

* means significant on the 5% level

This table shows clearly (R^2) that we have not enough data to make very strong predictions with respect to the reliability of the questions[1] . We also see that the results in the main questionnaire, using mainly oral presentation, are very different from the results in the drop off form using completely visual presentation. Looking at the specific results we see that the language has a non-significant effect on the reliability and the same is true for the effect of distance between the questions in the two forms. The last phenomenon can be explained by the fact that the largest difference exist between the items of the different forms and not within the different forms.

[1] . This is only possible if these results are combined with our previous results but that can not be done at this moment.

Two effects discussed above as rather strong clearly remain significant on 5% level in this analysis. One has to do with the choice to place an item in a battery or not. This characteristic is strongly related with the approach using agree/disagree questions and statements. The alternative is the use of direct questions where the basic assertion is incorporated in the question. It is clear from these results that in the ESS the use of statements has a strong negative effect on the reliability in both modes of data collection.

The second variable which has a significant effect in both modes of data collection is the use of gradation or not. For category scales this means the use of more categories , for frequency questions it refers to the number of values that can be given. These analyses suggest that more categories always lead to better results. The analysis is, however , not precise enough to show that in case of 11 categories or frequencies above 100 the reliability did decrease again. We have mentioned that this is a phenomenon that can be explained by variation in response functions (Saris 1986).

The ordering of the categories in the scales does not have a significant effect , even the sign in reversed for the different modes of data collection. This result indicates that this variable is too confounded in this data set to lead to a clear result.

Questions asking a comparative judgement have in general a higher reliability than questions asking an absolute judgement or a frequency. However the standard errors are very large due to correlations with other characteristics and therefore the effect is only significant in case of self administered questionnaires. Finally not using show cards has a rather large negative effect on the reliability of the questions. This variable only plays a role if the interviewer administers the questions. It is in the expected direction but this effect turned out not to be significant on the 5% level. Again the large standard error is due to relationships with other variables and the small sample size.

Conclusion: On the basis of these results and previous meta analyses it seems that one should be careful with the use of too few categories in scales and with agree/disagree question formats with statements in a battery. It is less certain if one can use questions without show cards. To be sure we would recommend to use show cards when ever possible. There is no indication that the show cards can have a negative effect while there are several indications that not using them can have a negative effect at least at some questions.

Special attention is required for the switch of mode of data collection. The previous results have clearly indicated that self administration will have a positive effect on the reliability of the question. This was also found in previous meta-analyses. This point requires further attention because a switch of mode of

data collection in the direction of self administration (mail or Web-surveys) is the most likely possibility for the future. The negative effect of a completely oral mode of data collection would suggest not to go in the direction of telephone interviewing. This conclusion is also in agreement with previous meta analyses where telephone interviewing led to the lowest data quality.

In order to explain differences in validity the variables mentioned above have only a minor effect and their effects are often confounded with the effect of other variables. Therefore we also looked for other variables. One we have mentioned above is the number of fixed reference points in scales. Further we expect a possible effect of the way the don't know option is provided. Furthermore we expected a effect of the variables called 'range' which indicate whether the range of the theoretical variable is the same as the range of the observed variable, especially one can have that the variable theoretically is bipolar while the response scale used is uni-polar. With the variable characterizing the data collection mode these variables have been used as explanatory variables in the analysis. In the analysis the variables which characterize the design of the experiment namely the language used and the distance between the questions have also been included in the analyses and also kept in the equation even though they had no significant effect. We also expected effects of the complexity of the question for example the number of words in the question or introduction or the number of abstract nouns on the total number of nouns. However these variables all had only small non-significant effect and are therefore omitted in the analysis presented below.

Linear regression where the dependent variable is the reliability of the different questions

	Administration by Respondent (71 questions)	Interviewer (86 questions)
Intercept	.785*	1.063*
Language	.017	.002
Distance	-.000	-.000
# of fixed ref. points	.004	.005*
don't know	.067*	-.035*
range	-.026*	-.024*
No show cards	-	.037*
Adjusted R^2	.28	.38

* means significant on the 5% level

Also in this case we see that the language has no significant effect. The same is true for the variable distance between the measures of the same traits.

The number of fixed reference points has a positive effect which is approximately equally large for the two modes of data collection but in the main questionnaire the effect is significant. This means that the validity is higher the more fixed reference points there are. Although we have specified an additive model we expect this variable to play a role mainly if the number of categories is larger that 7 which means that we should in fact specify an interaction effect but the sample size is too small to do so. It is therefore even more remarkable that this effect has been found. The small size of the effect is due to the miss-specification of the model.

Another surprising effect has been found for the 'don't know option'. The coding here is 1= dkn option present 2 =dkn registered not mentioned 3= no dkn option. One should realize that a positive effect on validity means smaller method effects or less variation in the way the question is interpreted and respondent. So the significant positive effect of the don't know variable in case the respondent is administering the questions to him self suggests that this lead to significant less method effect if the don't know option is absent while in case the interviewer presents the questions the method effect is increased if this option is not provided. In case of no don't know option the score of the variable is 3 and therefore the validity will be reduces with .1 in the interviewer administered questionnaire which is a very large reduction while in a self administered interview it would lead to an increase in validity of .2.

The effect of the variable 'range' is significant and approximately the same in both data collection modes. This result suggests that we should avoid questions where the variable which we want to measure is bipolar while the scale used is unipolar. An example is the concept satisfaction which goes from 'very satisfied' to 'very dissatisfied' while sometimes a measurement scale is used going form 'not at all satisfied' to 'very satisfied'. The latter is an unipolar scale while the former is a bipolar scale.

The last variable to discuss has again to do with the data collection mode. The table shows that the validity is higher if only an oral presentation is used compared with the alternative of using show cards. Here we see an opposite effect of the same variable on reliability and validity. In practice it can make quite a difference if the validity is increased by not using show cards or the reliability is increased by using them. It can easily be proven that one should use show cards if the validity is larger that the reliability if one would like to improve the total quality of the instrument[2]. Since normally the validity is higher than the reliability one better can use show cards to get optimal quality.

Conclusions: On the basis of these analyses we would suggest that

[2] Total quality= $(rxv)^2$ so the questiion is when $[(r+.04)(v-.04)]^2 > [(r-.04)(v+.04)]^2$ working out this inequality will lead to the result that this is the case if $v > r$..

- in case of scales with more than 7 categories fixed reference points should be use such as extremely good and bad in stead of good and bad
- in the main questionnaire in general a don't know option should be provided in order to avoid method effects due to the interviewer in case a don't know is needed not to use unipolar scales if bipolar scales are possible
- to use show cards if possible for all questions in the main questionnaire

3. Suggestions for the ESS questionnaire in the first round

In the first section suggestion for the questions involved in MTMM experiments have been specified. In the second section more general remarks have been made. On the basis of these two sections I will now make suggestions for changes in the ESS questionnaire for the first round. In doing so the following general rules have been applied which follow from the analysis:

1. use of scales with as many categories as reasonable but above 7 categories fixed reference points should be provided.
2. Avoid the agree/disagree format with statements in batteries as much as possible
3. Use of show cards under all circumstances
4. Provide a don't know option in the main questionnaire
5. Avoid to use unipolar scales if bipolar scales are also possible

Taking these rules into account and the suggestions of the first section we come to the following suggestions for the main questionnaire of the ESS.

Media use: According to my analysis of the media questions the questions B1 to B6 should be substituted by the questions of the form of the questions N10 – N13.

Frequency of contacts: According to my analysis of the frequency of contact questions the questions B7, F1 and F3 should be substituted by the questions L4-L6.
Political interest: If more categories are possible it would be recommended. This holds also for C1 and C2.

Political efficacy: My analysis suggests that the questions C3- 7 should be rewritten in direct questions like questions N13, N14 and N15. For the last two questions a similar direct question form has to be developed.
Political trust: I would recommend to use for items C8- C14 in stead of 'very strong trust' the term 'complete trust' or a similar formulation of a fixed reference point.

Social trust: I would recommend to change only the label of C15 'one should always be careful' . The opposite pole would be "people can always be trusted ' but nobody would believe this. The same is true for the other items.

Participation in elections: no suggestions

Other forms of political participation: As I have mentioned in section 1 I would recommend to use the complete list N43A-R in stead of the short list C20-C27. It does not take much time and there are many more possibilities to compare with the past and to look for changes.

Party membership/ identification: No suggestions for changes in questions C28-C32.

Evaluation of the economy: It was recommended to substitute the questions C33-C35 by the questions N19-N21 using a 11 point scale. Substituting 'very' by 'most' would be recommended.

Multilevel government: no suggestions for items D1 – D10.

Socio-political orientations: With respect to items E1 – E16 the recommendation would be to change these agree/disagree battery into direct questions but that would give a lot of work and reduce the comparison with the past so I suggest to leave these questions as they are if they work well.
Social exclusion: Questions F1 and F3 are discussed before. I would recommend to keep the other questions in this block the same for comparison with previous measures.

Religion: I have no suggestions for changes in the G block, except that one could consider to change the ordering of the categories for questions G6 and G7.

Citizenship, ethnic identity..: No suggestions for questions H1 – H11. The questions H12 – H14 were supposed to measure prejudice. In order to check this I looked briefly to question H14. There is a big difference in responses between the British sample and the Dutch. The Dutch answer between 60% and 75% that they are a little bit prejudiced under nearly all circumstances checking with education (m5) and with questions about allowing people from outside Europe in the country (J3) and tolerance item J33. Given this situation the correlations between these variables is minimal in the Netherlands. This is in sharp contrast with the British sample where one can find much more variation in the responses between the groups on the other variables. Given this situation I fear that these questions will not work so well outside Britain. The question is if we need them.

Immigration: I do not comment on the content of the questions. That will be done by Jaak and his assistant. I will only comment on the form of the questions.
With respect to questions J1 – J12 I would substitute 'very important' by 'extremely important'. I would also leave J13-J15 as they are for a change.
The questions J16 – J22 have a nice distribution and seem to work well but have a different format in the sense that the end points of the scales represent

consequences which do not exclude each other like important and unimportant / good and bad. Therefore it is not clear what people do if they think that both events will occur for example they will take away jobs and create new jobs. It would be interesting to know the relationship between the response to this item and two separate items about taking away jobs and one about creating new jobs. For the moment I do not see a reason for a change.
No suggestions for questions J23 and J24, this format is good for a change.
One could consider to change the labels 'mind a lot' in questions J25-J33 into 'would be against'. In doing so the contrast is made a bit stronger and possibly more differentiation is obtained between the types of immigrants.
No suggestions for questions J34-J43. Concerning questions J44 – J52 I would recommend to reverse the scale. No comments on questions J53- J56.

Citizenship, involvement and democracy: With respect to questions K2A-G the suggestion would be to use maximally important in stead of very important. The same suggestion could be made for the questions K7A-F.
The scale for items K9A – E is rather strange. One could better ask How much influence could you have on different aspects of your job:

No influence										complete control	DK/NA
00	01	02	03	04	05	06	07	08	09	10	88

working hours
organization of your work
etc

This applies for K9A-E and H and I while items E and F should be formulated on a scale of easiness as it is done.

For K10 and K13 one could think of fixed reference point: 'completely' in stead of 'very'

The L block will be omitted.
Demographics: With respect to the questions M1 to M25 I have no suggestions.
Questions 26a is the same as question K9E except for the three month period. Otherwise the questions M26a and b could also be formulated with a 11 point scale.
The same is true for question M7.
The questions M55 –M59 could be seen as an intrusion of the privacy. Besides, there are hardly people who answer these questions. If I remember correctly 33 people should answer question 57 and 38 have answered this question. Does it make sense to ask such questions ?

The consequences for the MTMM experiments

Given that several sets of questions in the main questionnaire have been changed also the drop off form with respect to the MTMM experiments has to be changed. I will provide the six sets of questions which could be provided to the different groups after the decision about the main questionnaire is made.

Here I would like to emphasize that the MTMM study is still needed for the following reasons:

1. We should get an estimate of the quality of the measures in other languages than in English and Dutch.
2. Without these estimates of data quality the results can not be compared across countries.
3. Also for English and Dutch it is relevant to see if the results are stable and whether the ordering of the questions plays an important role in the estimates.

With these arguments in mind and taking into account the changes in the questionnaire the drop off versions for the different groups will be reformulated

4. The MTMM questions for round 1 of the ESS

The MTMM questions can be presented to the respondents in three different ways:

1. At the end of the interview by the interviewer
2. At the end of the interview as a self administered questionnaire
3. In a drop off questionnaire that has to be send to the fieldwork agency

The questionnaire can contain 6 or 18 questions depending if a design is used with 6 groups (6 questions) or with 2 groups (18 questions). This leads to the following combinations:

Version	**Interviewer administered, face-to-face (with show cards)**	**Self completion**	
		Interviewer present	**Drop off form**
6 questions and 6 groups	F-F6, groups A to F	S-C6, groups A to F	
18 questions and 2 groups	F-F2, groups A & B	S-C2, groups A & B	S-C2, groups A & B

If the possibility of 6 questions is chosen the set of questions can be split in 6 parts as indicated below and each part can be presented to a different group. This means that each group gets only 6 extra questions besides the 21 value

questions. The two forms should be randomly assigned to the different respondents

The simplest way to realize randomization in this case is the make a booklet with 6 pages each containing a questionnaire for a group and print this book as many times as interviews will be done divided by 6. If the pages are not sorted and each interviewer is given as many pages as he/she will do interviews starting from the top of the pile of pages going down than the different forms are automatically randomly distributed over the interviewers and respondents.

Because of the fact that the use of a real drop off form may increase the non-response considerably, it is necessary that in case of a drop off form only two versions of the questionnaire are used. One made up of the questions specified below for groups 1,2,and 3 and another questionnaire consisting of the questions for the groups 4,5, and 6. This means that each group gets 18 questions besides the value questions. The two forms should be randomly assigned to the different respondents.

As the questions can be presented by the interviewer at the end of the interview or can be provided for self completion to the respondent two forms have been made; one with show cards and one without show cards.

Below is a mock-up of the supplementary questionnaire to be used in round 1 at the main stage. This shows the version where 6 groups get 6 questions each, to be completed by self-completion. The footnotes refer to the corresponding question in the main questionnaire, which will also form part of the experiment.

Questions for group 1 (self-completion)

HS1 On an average weekday, how much time, in total, do you spend watching television[3]?

WRITE IN HOURS: ☐☐ AND MINUTES: ☐☐

HS2 On an average weekday, how much time, in total, do you spend listening to the radio[4]?

WRITE IN HOURS: ☐☐ AND MINUTES: ☐☐

[3] See A1 in main questionnaire

[4] See A3

HS3 On an average weekday, how much time, in total, do you spend reading the newspapers[5]?

WRITE IN HOURS: ☐☐ AND MINUTES: ☐☐

Please indicate to what extent you agree or disagree with each of the following statements.

HS4 **"**Sometimes politics seems so complicated that I can't really understand what is going on[6]." **Please tick one box.**

Disagree strongly ☐ 1
Disagree ☐ 2
Neither disagree nor agree ☐ 3
Agree ☐ 4
Agree strongly ☐ 5

HS5 "I think I could take an active role in a group involved with political issues[7]" **Please tick one box.**

Disagree strongly ☐ 1
Disagree ☐ 2
Neither disagree nor agree ☐ 3
Agree ☐ 4
Agree strongly ☐ 5

HS6 "I find it easy to make my mind up about political issues[8]"

[5] See A5
[6] See B2
[7] See B3

Please tick one box.

- Disagree strongly ☐ 1
- Disagree ☐ 2
- Neither disagree nor agree ☐ 3
- Agree ☐ 4
- Agree strongly ☐ 5

Questions for group 2

HS7 On the whole how satisfied are you with the present state of the economy in [country][9]? **Please tick one box.**

- Very dissatisfied ☐ 1
- Fairly dissatisfied ☐ 2
- Fairly satisfied ☐ 3
- Very satisfied ☐ 4

HS8 Now thinking about the [country] government, how satisfied are you with the way it is doing its job[10]? **Please tick one box.**

- Very dissatisfied ☐ 1
- Fairly dissatisfied ☐ 2
- Fairly satisfied ☐ 3
- Very satisfied ☐ 4

[8] See B4
[9] See B30
[10] See B31

HS9 And on the whole, how satisfied are you with the way democracy works in [country][11]? **Please tick one box.**

- ☐ 1 Very dissatisfied
- ☐ 2 Fairly dissatisfied
- ☐ 3 Fairly satisfied
- ☐ 4 Very satisfied

HS10 Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people[12]? Please tick the box that is closest to your opinion, where 0 means you can't be too careful and 5 means that most people can be trusted.

You can't be too careful					Most people can be trusted
0	1	2	3	4	5
☐	☐	☐	☐	☐	☐

HS11 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[13]? **Please tick one box.**

Most people would try to take advantage of me					Most people would try to be fair
0	1	2	3	4	5
☐	☐	☐	☐	☐	☐

HS12 Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves[14]? **Please tick one box.**

People mostly look out for themselves					People mostly try to be helpful
0	1	2	3	4	5
☐	☐	☐	☐	☐	☐

[11] See B32
[12] See A8
[13] See A9
[14] See A10

Questions for group 3

Please indicate on a score of 0 to 10 how much you <u>personally</u> trust each of the institutions below. 0 means you do not trust an institution at all, and 10 means you have complete trust[15].

Please tick the box that is closest to your opinion.

		No trust at all										Complete trust
HS13	[Country]'s parliament	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	10 ☐
HS14	The legal system	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	10 ☐
HS15	The police	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	10 ☐

Please indicate to what extent you agree or disagree with each of the following statements.

HS16 "The less that government intervenes in the economy, the better it is for [country][16]"
Please tick one box.

- Agree strongly ☐ 1
- Agree ☐ 2
- Neither agree nor disagree ☐ 3
- Disagree ☐ 4
- Disagree strongly ☐ 5

[15] See B7, 8 and 9

[16] See B43

HS17 "The government should take measures to reduce differences in income levels[17]".
Please tick one box.

- Agree strongly ☐ 1
- Agree ☐ 2
- Neither agree nor disagree ☐ 3
- Disagree ☐ 4
- Disagree strongly ☐ 5

HS18 "Employees need strong trade unions to protect their working conditions and wages[18]".
Please tick one box.

- Agree strongly ☐ 1
- Agree ☐ 2
- Neither agree nor disagree ☐ 3
- Disagree ☐ 4
- Disagree strongly ☐ 5

Questions for group 4

HS19 On an average weekday, how much time, in total, do you spend watching television[19]?
Please tick one box

- No time at all ☐ 01
- Very little time ☐ 02
- A little time ☐ 03
- Some time ☐ 04
- Quite a lot of time ☐ 05
- A lot of time ☐ 06
- A great deal of time ☐ 07

[17] See B44
[18] See B45
[19] See A1

HS20 On an average weekday, how much time, in total, do you spend listening to the radio[20]?
Please tick one box

No time at all	☐ 01
Very little time	☐ 02
A little time	☐ 03
Some time	☐ 04
Quite a lot of time	☐ 05
A lot of time	☐ 06
A great deal of time	☐ 07

HS21 On an average weekday, how much time, in total, do you spend reading the newspapers[21]?
Please tick one box.

No time at all	☐ 01
Very little time	☐ 02
A little time	☐ 03
Some time	☐ 04
Quite a lot of time	☐ 05
A lot of time	☐ 06
A great deal of time	☐ 07

Please indicate to what extent you agree or disagree with each of the following statements.

HS22 **"**Sometimes politics seems so complicated that I can't really understand what is going on.[22]" **Please tick one box.**

Agree strongly	☐ 1
Agree	☐ 2
Neither agree nor disagree	☐ 3
Disagree	☐ 4
Disagree strongly	☐ 5

[20] See A3
[21] See A5

HS23 "I think I could take an active role in a group involved with political issues.[23]"
Please tick one box.

- ☐ 1 Agree strongly
- ☐ 2 Agree
- ☐ 3 Neither agree nor disagree
- ☐ 4 Disagree
- ☐ 5 Disagree strongly

HS24 "I find it easy to make my mind up about political issues.[24]"
Please tick one box.

- ☐ 1 Agree strongly
- ☐ 2 Agree
- ☐ 3 Neither agree nor disagree
- ☐ 4 Disagree
- ☐ 5 Disagree strongly

Questions for group 5

HS25 On the whole, how satisfied are you with the present state of the economy in [country][25]? Please tick the box that is closest to your opinion, where 0 means extremely dissatisfied and 5 means extremely satisfied.

Extremely dissatisfied					Extremely satisfied
0	1	2	3	4	5
☐	☐	☐	☐	☐	☐

[22] See B2
[23] See B3
[24] See B4
[25] See B30

HS26 Now thinking about the [country] government, how satisfied are you with the way it is doing its job[26]? **Please tick one box.**

Extremely dissatisfied					Extremely satisfied
0	1	2	3	4	5
☐	☐	☐	☐	☐	☐

HS27 And on the whole, how satisfied are you with the way democracy works in [country][27]? **Please tick one box.**

Extremely dissatisfied					Extremely satisfied
0	1	2	3	4	5
☐	☐	☐	☐	☐	☐

HS28 Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people[28]?
Please tick one box.

You can't be too careful	☐ 1
Most people can be trusted	☐ 2

HS29 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[29]?
Please tick one box.

Most people would try to take advantage of me	☐ 1
Most people try to be fair	☐ 2

[26] See B31
[27] See B32
[28] See A8
[29] See A9

HS30 Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves[30]?
Please tick one box.

People mostly look out for themselves ☐ 1

People mostly try to be helpful ☐ 2

Questions for group 6

HS31 Please say on a scale of 0 to 10 how much you trust **[country]'s parliament**. If you have no trust at all give a score of 0. If you have complete trust, give a score of 10. The more you trust the parliament, the higher the score should be[31].

Your score: ☐☐

HS32 Please say on a scale of 0 to 10 how much you trust the **legal system**. If you have no trust at all give a score of 0. If you have complete trust, give a score of 10. The more you trust the legal system, the higher the score should be[32].

Your score: ☐☐

HS33 Please say on a scale of 0 to 10 how much you trust the **police**. If you have no trust at all give a score of 0. If you have complete trust, give a score of 10. The more you trust the police, the higher the score should be[33].

Your score: ☐☐

[30] See A10
[31] See B7
[32] See B8
[33] See B9

HS34 Is it generally good for [country] if government intervenes less in the economy[34]? **Please tick one box.**

- Definitely ☐ 1
- Probably ☐ 2
- Not sure either way ☐ 3
- Probably not ☐ 4
- Definitely not ☐ 5

HS35 Should the government take measures to reduce differences in income levels[35]? **Please tick one box.**

- Definitely ☐ 1
- Probably ☐ 2
- Not sure either way ☐ 3
- Probably not ☐ 4
- Definitely not ☐ 5

HS36 Do employees need strong trade unions to protect their working conditions and wages[36]? **Please tick one box.**

- Definitely ☐ 1
- Probably ☐ 2
- Not sure either way ☐ 3
- Probably not ☐ 4
- Definitely not ☐ 5

[34] See B43
[35] See B44
[36] See B45

Questions for all groups

At the end of the questionnaire for each of the 6 groups, where it is completed by self-completion, a number of identification questions will be asked, in order to check whether the designated respondent filled in the supplementary questionnaire, and when it was filled in. These are as follows:

Are you…

…male ☐ 1

or, female? ☐ 2

In which year were you born?

Write in year: **1 9** ☐☐

PLEASE ENTER TODAY'S DATE: DAY ☐☐ MONTH ☐☐ YEAR ☐☐