WORK PACKAGE 9 – ANALYSIS OF RELIABILITY AND VALIDITY AT THE MAIN STAGE

Responsibility: University of Amsterdam

The purpose of this work package is to design MTMM experiments for the main questionnaire and to analyze the results of these experiments in order to evaluate the data quality, especially, the reliability and validity of the questions in the different languages

So far all decisions with respect to the design of the questions were based on knowledge that was collected with questionnaires in English , German and Dutch. So it was not at all clear that the choices made with respect to the design of the questions for all countries would be the best. Therefore also in the final questionnaire used in all countries MTMM experiments were built in to evaluate the choices that were made with respect to the formulation of the questions.

It was clear that it was impossible to evaluate all questions of the core questionnaire at once. That would cost too much time and make the response burden for the respondents too high. The MTMM experiments can also not provide estimates of the effects of all variables on the reliability and validity of the questions. In our experiments we could only study a number of crucial variables. The estimates of data quality are affected by other factors such as the position in the questionnaire, the distance to another MTMM measure, the length of the text etc. To be able to take all these factors into account and to make predictions of the quality of questions which have not been studied explicitly one needs another approach which is called the meta analysis of MTMM experiments (Andrews,1984, Scherpenzeel and Saris 1997).

Therefore the MTMM experiments are chosen in such a way that they cover the most crucial and influential choices which are made in the design of the questions in the ESS. The experiments allow to see if our choices with respect to the forms of the questions for the main survey were a good choice leading to higher reliability and validity of the questions in the main questionnaire than for the alternative form in the supplementary questionnaire. Later a meta analysis of the results of these experiments can provide an estimate of all the effects of the the crucial choices on the reliability validity and method effects. If these results are obtained, we expect that predictions can be made of the quality of all instruments as we have done in the design phase of the ESS questionnaires. The predictions were based on more than 1000 survey questions.

In the proposed MTMM experimental design the following crucial factors have been suggested for evaluation and have been evaluated in the pilot:

a. open questions asking frequencies or amounts versus 7 point category scales
b. dichotomous versus 5 points and 11 point scales
c. 5 point agree/disagree items with statements versus direct questions with construct specific responses
d. 11 point bipolar scales with show cards or without them
e. 4 point bipolar scales versus 4 point unipolar scales and 11 point bipolar scales
f. use of agree/disagree batteries compared with direct questions with construct specific responses

In this approach the choice of the topic is not so important but in the ESS we have to select for the experiment those topics which are in the questionnaire already. The following choice was made for the different experiments:

a. media use
b. political efficacy
c. social trust
d. satisfaction with the economy, democracy and government
e. trust in political institutions
f. socio-political orientations

Other topics could have been chosen, but, we have chosen for sets of questions from the core questionnaire because they should get priority in the evaluation of their quality.

Due to the fact that we have to use questions present in the questionnaire not just the above mentioned 6 characteristics will vary across the experiments but also other characteristics for example:

- the position in the questionnaire,

- the distance to the next MTMM question
- the mode of data collection etc, (see restrictions) but also
- the length of the question text,

the number of sentences, the number of labels etc.

In the data collection the split ballot MTMM design has been used in order to reduce the efforts for the respondents and to reduce memory effects. Recent research of Saris, Satorra and Coenders (forthcoming) has shown that problems occur if the correlations between the traits are too close to zero. In that case non-convergence or improper solutions are obtained. Only for the topics media use and socio-political orientations the correlations between the traits was so low that problems could have been expected.
The results described below are based on the data of 14 countries[1] while in each country 6 experiments are done. In principle that would give 84 MTMM matrices but in two countries the analysis could not be done for the media data and in three other countries no analysis could be done for socio-political orientations. This means that in total 79 data sets have been analyzed. In total each experiment was based on 9 questions , so the analysis is based on the data of 711 questions. The analysis is done using the model developed by Saris, Satorra and Coenders (forthcoming). The program Lisrel is used for the estimation of the reliability and validity from the 79 correlation matrices.

The quality of the data
In table 1 the mean reliability and validity coefficients are given for the forms used in the main and supplementary questionnaires in the different countries. The best result in each country has been underlined.

	Reliability coefficient			Validity coefficient		
Country	main	sup 1	sup2	main	sup 1	sup 2
UK	.818	.855	.852	.973	.929	.937
Ireland	.868	.816	.799	.955	.932	.912
Netherlands	.871	.808	.816	.988	.919	.934
Sweden	.814	.843	.856	.962	.875	.931
Norway	.836	.823	.779	.987	.887	.884
Finland	.846	.817	.831	.957	.922	.941
Spain	.870	.843	.819	.976	.925	.913
Portugal	.911	.875	.874	.957	.929	.899
Greece	.908	.904	.878	.982	.958	.936
Czech	.897	.845	.819	.932	.929	.937
Poland	.841	.888	.860	.948	.975	.963
Slovenia	.853	.864	.868	.930	.958	.986
Switzerland	.822	.855	.843	.982	.923	.941
Israel	.863	.856	.842	.985	.926	.931
Means	.858	.849	.837	.966	.927	.934

Table 1 The mean reliability and validity coefficients over six experiments in 14 countries

This table shows very clearly that in most cases the right choices have been made. With respect to reliability in 9 out of 14 countries the reliability coefficient of the question in the main questionnaire was higher than the reliability coefficient of the questionnaires in the supplementary questionnaire. With respect to validity this was also the case : 11 out of 14 countries had a higher validity coefficient in the main questionnaire than in the supplementary questionnaire.

[1] In this analysis the data of all countries have been analyzed that are released in September except Hungary. Hungary is omitted because for this country the data of the supplementary questionnaire was incomplete.

This is a very promising result because it suggests that information from studies in "Old European" countries can be applied in other countries as well. So it suggests some stability of these effects. However this should not close our eyes for the fact that quite large differences in results have been obtained for different topics. So far the results were aggregated across topics.

Let us now look at the differences between the different forms for each of the topics taking the means over the countries. In Table 2 these results are summarized.

	Reliability coefficient			Validity coefficient		
Topic	main	sup 1	sup 2	main	sup 1	SUP 2
Media (12) .	.938	.748	.826	.997	.878	.946
Political efficacy (14)	.797	.908	.833	1.00	.969	.966
Satisfaction (14)	.878	.816	.875	.907	.881	.939
Social trust (14)	.832	.880	.775	.966	.909	.857
Political trust (14)	.924	.944	.923	.957	.949	.940
Soio-pol orientations (11)	.771	.889	.793	.974	.980	.959

Table 2 The mean reliability and validity coefficients for different topics across countries. The number of studies the results are based on are mentioned within brackets.

Table 2 shows that the validity coefficients in the main questionnaire for all topics are rather close to 1 which means that there is hardly any method effect. That can not be said for all alternative measures. So this confirms again that the choice for the main questionnaire was rather good. With respect to the reliability coefficients table 2 shows much more variation also for the different questions in the main questionnaire. A reliability coefficient of .938 means that observed correlations are hardly effected by the random errors while a reliability coefficient of .748 would mean that observed correlations for the latter measure would be half the size of the former one.

To evaluate the effect of the different factor more studies are needed but some effects are easy to detect. For example , for media use a scale was chosen with 7 categories specified in minutes for the main questionnaire (.938). This question worked very well. In one group we used an open question where people had to indicate in hours and minutes how much time they spent using the media. It turns out that people specify minutes where hours are registered or only hours and no minutes. An interviewer can easily correct these errors or even prevent them but in a self completion form many of these errors are made. Without a lot of editing this open question creates a very low reliability (.748).

Another obvious case is the yes/no format which is often used for social trust questions. The second form of the supplementary questionnaire contains that question. We see that the reliability coefficient is very low .775. This is however not due to random errors but due to the dichotomous character of the scale which created lots of errors compared with the other scales with more data points. Such a scale can only be used if the analysis is adjusted to this level or the correlations are corrected for these effects for example by using tetrachoric correlations instead of Pearson correlations.

Remarkable is also the difference between the social and political trust items in the main questionnaire. Both sets of questions have been combined with an 11 point scale. The only major differences between the two approaches are the labels to define the end point of the scale. In the social trust items full sentences are used such as "You can't be too careful" and "Most people can be trusted" while for the political trust items short fixed reference points are used "No trust at all" and "Complete trust". One should be aware that only due to these differences in reliability the political trust variable can have a stonger relationship with other variables than social trust.

A last example that deserves some attention concerns the topic socio-political orientations. The questions in the main questionnaire and the first form of the supplementary questionnaire are by purpose formulated exactly in the same way. The difference is, however, that for many respondents the supplementary questionnaire was a self completion questionnaire. This is the only difference. It seems that this difference has created a difference in reliability of .1 between the two measures, which is a lot. That the second form of this set of questions has also a low reliability has a different reason. Here the point is that the question is formulated as a yes/no question while people have to indicate an extremity of their opinion. As Ongena(2003) has shown this can also lead to lower data quality.

All results mentioned here are in agreement with previous findings with respect to effects of these factors on reliability and validity of questions.

Conclusions

The above results show that the choices that have been made in order to improve the quality of the questions in the main questionnaire have in general had a positive effect on the quality of the data. In most countries the questions used in the main questionnaire are better than possible alternatives tested in the supplementary questionnaire. The average level of the quality is also quite good. Over all countries and topics the reliability coefficient is .858 and the validity coefficient is .966. This result means that of the total variance of an observed variable with this qualities 68,7% is due to the trait to be measured .4.9 % is due to the systematic error and 26.4 % is due to random measurement error. This is a good result. However it should be remarked that the quality of the questions varies with the topic studied as we have seen in table 2 and with the country studied as shown in table 1. This issue should be of concern for the cross cultural analysis even though the quality of the questions in general is quite good

17 july 2003

Willem E.Saris
Irmtraud Gallhofer