UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 22

2011

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici ESCI-Born - Despatxos 19.517-19.527
Passeig Pujades, 1 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

Is there anything wrong with the MTMM approach to question evaluation?

Willem Saris

RECSM/UPF

Barcelona

In 1959 Campbell and Fiske introduced the Multitrait Multimethod (MTMM) approach to evaluate the convergence and discriminant validity of measurement instruments. 1984 Frank Andrews adjusted the procedure to estimate the quality of questions for survey research. He also did the first meta-analysis across many experiments to evaluate the effect of the different question characteristics on the quality of the questions. After his death my research group continued his valuable work by doing more experiments, doing a new meta analysis and developing a program for prediction of the quality of questions (Oberski et al 2005) on the bases of the findings of the meta analysis (Saris and Gallhofer 2007).

In 2001 the European Social Survey (ESS) decided to include in its data collection MTMM experiments to evaluate the quality of questions, to compare the quality across countries and to correct for possible differences in quality across these countries. In order to apply this approach in the ESS the design of the MTMM experiment was adjusted to reduce memory effects by repetition of the same questions. Saris, Satorra and Coenders (2004) developed for this purpose the Split Ballot MTMM design. This design has been used in all rounds of the ESS in more than 20 countries.

In a recent book, edited by Madans, Miller, Maitland and Willis (2011) published by Wiley in the series on survey research three authors criticize the MTMM approach. Because these criticisms come from three very well known and respected scholars, Duane Alwin, Jon Krosnick and Peter Mohler and they have been published in the very prestigious survey research series of Wiley I feel obliged to indicate that there is nothing wrong with the MTMM approach but that the criticisms are unjustified.

Peter Mohler emphasizes in his chapter in the reader of Madans et al (2011) that survey research is a complex process consisting of many different steps. He suggests that one should evaluate for each stage whether the result passes predefined quality benchmarks. If that is not the case one should first correct this stage. This approach is also put forward by Biemer and Lyberg 2003, Mohler et al 2010 and Pannell et al 2010. On page 307/8 Peter Mohler writes:

Guidance is given by current standards and best practices that indicate what should be avoided (negation, double barreled, context effects) in designing questions. Thus they are no prescriptions or cookbooks for "good" questions but there are well established procedures governing the questionnaire design process. This process must be quality controlled using well-

defined assessments and benchmarks (Presser et al 2004). Whether a question is "good or bad" is, in the end, not the question. In quality terms , questions/ items must optimally meet predefined measurement properties.

If such benchmarks have been established that would certainly be a useful approach. However it would also not harm to give survey researchers suggestions how they can avoid serious problems or at least make them aware of the choices they make. Given our experience with questionnaire design in general and in the ESS, we tried to do so in the chapter 1-9 of our book (Saris and Gallhofer 2007). Others have also contributed a lot in this context by their publications (Dillman et al. 2009 , Couper 2008).

However, no matter how much we try to prevent errors in each stage of questionnaire design, they will always be there and so the variable we measure will not be identical to the variable we want to measure. Therefore it would also make sense to know how strongly these two variables are related. This gives an indication of the quality of the measure used for the concept of interest. The MTMM approach tries to estimate this quality indicator denoted by q^2 .

We think that both approached make sense: we should try to maximize the quality in each step of the question design but we also should know what the final quality is. The latter is important because errors have strong effects on the means of variables and the relationships between variables. Let me illustrate this point. In table 1 three questions from the main questionnaire of round 3 the ESS have been presented.

Table 1 the three questions with respect to the consequences of immigration

B38 CARD 15 Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries? Please use this card.

Bad for the economy									**Good for the economy**		**(Don't know)**
00	01	02	03	04	05	06	07	08	09	10	88

B39 CARD 16 And, using this card, would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?

Cultural life undermined									**Cultural life enriched**		**(Don't know)**
00	01	02	03	04	05	06	07	08	09	10	88

B40 CARD 17 Is [country] made a worse or a better place to live by people coming to live here from other countries? Please use this card.

Worse place to live									**Better place to live**		**(Don't know)**
00	01	02	03	04	05	06	07	08	09	10	88

These questions are designed to measure the evaluations of the people with respect to consequences of the immigration for the economy, the cultural life and the live in general.

Imagine that we are interested in the correlations between these three evaluations, then the problem is that each of these questions will contain measurement errors i.e. if we would repeat these questions after some time, people will not give the same answers even if their opinion has not changed. So we can say that there is a difference between the opinion of the person in his mind and his response. For instance if the strength of the relationships between the opinion in the mind and the observed variable is .7 and the true correlation between the two opinions is .6 then the observed variables will have a correlation around .3 ($=.7^2 x.6$) and not .6. On the other hand it is well known that, if one knows the quality of the measures, one can recalculate the true correlation corrected for measurement error which is .6 ($=.3/.7^2$). In psychology the effect of measurement error and correction for measurement error has been discussed at length (see for example Lord and Novick 1968).

However there is another problem that was discussed by Campbell and Fiske (1959), one of the most cited paper in psychology. They would argue in this case that in all questions of table 1 the same method is used and people may react to this method in a different way but systematically over the different questions. For example some people use extreme values while others do not do so but both groups do this systematically. In that case the used method will cause a correlation between the observed variables. This correlation was called the Common Method Variance (CMV). In some fields reviewers suggests rejecting papers which try to make causal statements between variables which are measured with the same procedure (Campbell (1982). See the discussion by Lance et. al. (2011). We have presented these arguments in the model in Figure 1.

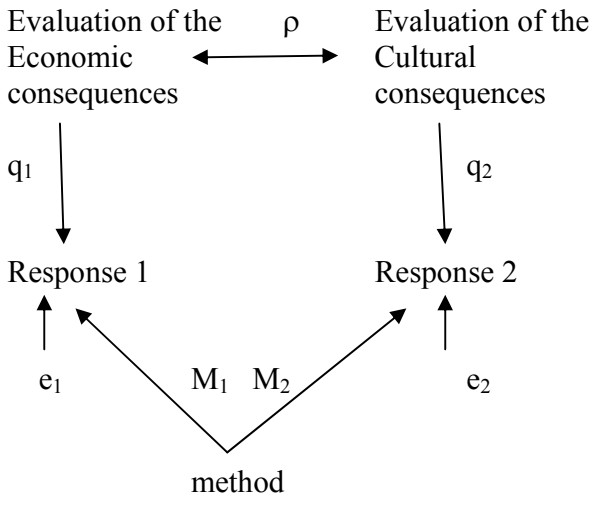

Figure 1 The model indicating the distinction between observed and latent variables and method effects

If the variables are standardized the path analysis suggest[1] for the correlation between Response1 and 2 (ρ_{21}) is

$$\rho_{21} = q_1 \rho q_2 + m_1 m_2 \qquad (1)$$

This model illustrates the well know fact that random errors will reduce the observed correlations and systematic errors or method effects can increase the observed correlations. So the correlation between the observed variables and the latent variables will not be the same. And as a consequence it is difficult to say something about correlations and regression coefficients between variables without knowing the size of these errors and correcting for them.

We would like to add to this that differences in error structure across countries can cause differences in correlations between variables across countries which have no substantive basis and make cross national comparison impossible without correction for these errors.

Given this situation, we think it makes sense to say a bit more about these measurement errors or their complement the quality of questions. The quality coefficient q is supposed to be an estimate of the strength of the relationship between the variable we want to measure and the observed variable. The square of this coefficient (q^2) is then an estimate of the variance of the observed variable explained by the variable we want to measure. So the value varies between 0 and 1 where 0 means no correlation at all and 1 means a perfect relationship. The higher the q^2 the better the measure is for the concept of interest. The q^2 gives a summary of the problems in designing the question and could be used as a benchmark, for example, if a question does not pass the level of say .7 one could decide that the quality of the question is not good enough and try to improve the measure.

I hope that by now the importance of this quality coefficient is clear. So the next topic should be the estimation of this quality coefficient.

The estimation of q

A crucial point in the use of the coefficient q is of course the quality of its estimation procedure. In this case three important requirements have to be discussed

The first requirement is that one needs repeated observations. Without repeated observations the quality cannot be evaluated. We know from factor analysis that at least three observed variables are needed in order to be able to estimate the loadings and the disturbance terms of the questions in a one factor model.

A second important requirement is that the three questions should measure the same variable and should vary only by the method used across the questions. If that is so then the loadings present

[1] This model is a bit too simple but we will come back on this issue later.

estimates of the quality coefficients q and the disturbance terms represents the errors due to the method used for the question. A typical example is that the question remains the same but that the response scale is varied, for example a 5 , 7 and an 11 point scale are used for the different questions. Table 2 gives 4 different versions of the question B38 of the ESS (see Table 1) about the consequences of emigration for the economy.

Table 2 Compilation of the alternative items used in the 2006 ESS MTMM experiment with reference to item B38

Now some questions about people from other countries coming to live in (country). Please read each question and tick the box on each line that shows how much you agree or disagree with each statement.

HS4 It is generally bad for (Country´s) economy that people come to live here from other Countries

Agree Strongly	agree	neither agree nor Disagree	disagree	disagree strongly
1	2	3	4	5

HS16 Now some questions about people from other countries coming to live in (country) How much do you agree or disagree that is generally bad for (country´s) economy that people come to live here from other countries?

disagree strongly										agree strongly
0	1	2	3	4	5	6	7	8	9	10

Now some questions about people from other countries coming to live in (country). Please read each question and tick the box on each line that shows how much you agree or disagree with each statement.

HS 25 It is generally bad for (Country´s) economy that people come to live here from other Countries. Please tick one box.

Disagree Strongly						agree strongly
1	2	3	4	5	6	7

All three questions measure without doubt the evaluation of the consequences for the economy of immigration. The question B38 is an item specific question i.e. the scale is specific for this question. The other three questions are items using the agree /disagree format. These items vary with respect to the number of categories but also with respect to the direction of the scale. Question B38 was presented to all respondents together with the questions B39 and B40 (see Table 1) in the main questionnaire of the ESS while the other three questions were presented to the respondents after all the questions of the main questionnaire were asked. These questions were part of an MTMM experiment to compare the quality of item specific questions with agree/disagree questions. Besides that we wanted to see the effect of the number of categories on the quality of agree/disagree questions.

If all three questions are asked to all respondents a correlation matrix of 3 variables is obtained which are measuring the same variable (evaluation of the consequences of immigration for the economy). So in this case in principle the two first requirements would be fulfilled and one could estimate a one factor model where the loadings would be the estimates of the quality coefficient q and the disturbance terms are estimates of the error variances. However, besides other problems, to be discussed later, we would not be able to separate the random errors and the systematic errors. Therefore also a third requirement has to be fulfilled.

The third important requirement is that three different concepts are included in the experiment, each with three questions which are the same but vary only in the method used and these methods are the same across the different concepts. This requirement is fulfilled if we use, for example for each question of table 1 the different methods presented in table 2. In that case for each method (5,7 and 11 point scale) the systematic effect on the three items of the different concepts can be estimated and the random error as well. So this is exactly what was done in the ESS for the questions of Table 1. In the main questionnaire the question presented in Table 1 were asked. The three other forms to measure the same were asked in the supplementary questionnaire after the main questionnaire was finished. These different forms have been presented in Table 3.

In the table we see that the three questions have been repeated in different forms. This means that the questions asked remained the same and so the three questions for each concept measure definitely the same variable (factor). The form of the questions varies for each concept but these forms are the same across topics. So each form appears three times. This is the minimum[2] requirement to be able to estimate the effect of the reaction to the form of the questions as well. This last characteristic has as consequence that the variance of the disturbance term can be split up in random errors and systematic errors and that one can estimate the Common Method Variance (CMV) which is caused by the method used between the observed variables (Campbell and Fiske 1959, Alwin 1974, Andrews 1984, Lance et al. 2011).

In the classical MTMM approach the estimation of the quality coefficients, the random measurement error variance and the method effects is done on the basis of the correlation matrix for the 9 variables of an experiment like the one in Table 3. In order to get these correlations one has to ask all 9 questions to each respondent. We will come back to this issue below.

[2] At least if we are not allowing for correlations between the method factors or introduce other restrictions on the effects of the methods.

Table 3 The questions used in the three supplementary questionnaires of the ESS

Supplementary questionnaire 1:

HS4[23]	It is generally bad for [country's] economy that people come to live here from other countries	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
HS5[24]	[Country´s] cultural life is generally undermined by people coming to live here from other countries	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
HS6[25]	[Country] is made a worse place to live by people coming to live here from other countries	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Supplementary quetionnaire 2

HS16[54] How much do you agree or disagree that it is generally bad for [Country] ´s economy that people come to live here from other countries?
Please tick one box.

Disagree strongly										**Agree strongly**
0	1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

HS17[55] And how much do you agree or disagree that [Country] ´s cultural life is generally undermined by people coming to live here from other countries?
Please tick one box.

Disagree strongly										**Agree strongly**
0	1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

HS18[56]How much do you agree or disagree that [Country] is made a worse place to live by people coming here from other countries?
Please tick one box.

Disagree strongly										**Agree strongly**
0	1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Supplementary questionnaire 3

HS28[85]	It is generally bad for [country's] economy that people come to live here from other countries	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07
HS29[86]	[Country´s] cultural life is generally undermined by people coming to live here from other countries	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07
HS30[87]	[Country] is made a worse place to live by people coming to live here from other countries	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07

The estimation is done with a factor model with three substantive factors and three method factors. The model is presented in Figure 1.

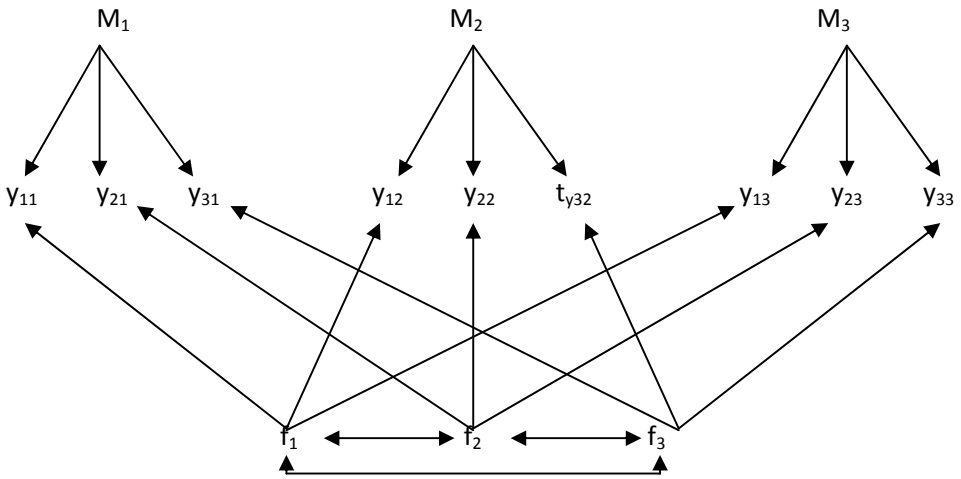

Figure 1 the classical MTMM model with three traits (f1-f3) , three methods (M1-M3) and 9 observed variables (y11-y33)

I hope that I have shown above that the estimation of the quality coefficient q is based on standard statistical methods which are generally recognized as acceptable procedures for many decades (Jöreskog 1969, Lawley and Maxwell (1971), Alwin 1974, Andrews 1984) and that q^2 indicates a very useful predetermined measurement property of a question i.e. the strength of the relationship between the observed variables and the variables that one wants to measure.

Peter Mohler clearly does not understand the fundamental basis of the MTMM approach presented above because on page 302 he gives his adaptation of an MTMM model. However, he says that he adapted a model in my book on page 311 but that model presents relationships between substantive variables and is not at all an MTMM model. On page 303 he suggests that he gives a calculation of the quality coefficient. However he presents a calculation of the quality of a composite score presented in my book on pages 296-297. That is a very different issue than the estimation of the quality of a single question. So I think that his criticism with respect to the estimation cannot be taken seriously.

This does not mean that there are no problems connected with this approach. Alwin, Krosnick and Mohler in the same book mention the problem of the memory effect due to the fact that one has to ask all respondents three times approximately the same questions. It will be clear that we have also been aware of this problem. We have tried to cope with this problem in two ways.

How to cope with memory effects

First of all we have suggested to reduce the memory problem by the development of the Split Ballot MTMM design (Saris, Satorra and Coenders 2004). Splitting up the sample randomly in subgroups, one can reduce the number of repetitions of the same question to 2 for each respondent. This can be done by giving group 1 method 1 and 2, providing group 2 with method 1 and method 3 and eventually group 3 with method 1 and method 4. This is the design used in the European Social Survey (ESS) but one can also use other designs, see Saris et al 2004). In that paper we have shown that the quality coefficients can still be estimated even though we have no complete correlations matrix for all variables in all groups[3]

It will be clear that repetition of the same questions is still present. So the question remains how much time should there be between the two measures of the same variables in order to avoid memory effects. Because this is indeed a fundamental issue in the MTMM design let me say more about the experiment that Van Meurs en Saris (1990) did to determine the minimal time gap between the repetitions of approximately the same questions.

For several different topics we repeated exactly the same questions within the same interview and two weeks later. In both cases before the question was repeated the respondents were asked whether they could remember the previous answer. Some people said "yes" and others "no". Both groups were asked to make a guess what the previous answer was It turns out that for the group which said that they could not remember their previous answer 36% nevertheless gave the right answer. This was true for the group who answered this question within the same interview as well as for the group who got this question after 2 weeks. This is interesting because it seems that people who say they do not remember their previous answer do not guess their previous answer better after 9 minutes in average within the same interview as after two week or other activities. So this seems to be a percentage of correct answers if people don´t remember their previous answer. The authors comment on this result: "This phenomenon is not surprising because most respondents are not likely to have changed their opinion in such a short time period." I would like to add that at least some people will have guessed the answer correctly.

In the same study one more thing was studied namely: How much time should pass before people are in the situation of having forgotten the previous answer so that the next answer is independent of the previous answer? This is an important question because it determines how MTMM experiments have to be organized especially with respect to the time between the two repetitions of the approximately the same question. In this study we took into account

- the time interval

[3] For an elaborate discussion of this issue we refer to the paper of Saris et al 2004 and a recent working paper of Revilla et al 2011.

- the extremity of the first response
- the topic of the questions between the repetitions : is the topic of these questions similar or not.

In the computer assisted data collection with different skip patterns the repetitions occurred in 0-3, 3-6,6-9,9-12, 12-15,15-20 and more than 20 minutes. Regressing the the percentage correct answers on the time interval in different conditions we got the following results.

1. If the people had an extreme opinion expressed in the first interview they always gave the same answer no matter the time interval between the repeated questions. So it would not help to make the time interval larger.

This is not surprising if they have an extreme opinion. In that case it is not so that they give the same answer because they remember their previous answer. It is more likely that they do so because they have an extreme opinion.

2. If the people had no extreme opinion and the kind of questions between the repeated question was similar to the kind of question repeated then the relationship was as follows:

$$C = .59 - .94T$$

Where C is the percentage correct answers and T is the time between the two questions. In this case every extra minute in the time interval will reduce the percentage correct answers with approximately 1%. This means that after approximately 15 minutes that percentage correct answers will be around 36% which is the percentage you get if people do not remember their previous answer as we have seen above.

3. If the people had no extreme opinion and the kind of questions between the repeated questions was different from the kind of repeated question then the relationship was as follows:

$$C – 75 - .50T$$

So in this case the extra minute of delay of the repeated question reduced the memory with only a half percent. Therefore the level of 36% of correct answers would be reached only after 80 minutes.

As we have shown above we did a careful analysis of the memory effect problem: we have first determined how much agreement one can expect if there is no memory effect anymore. If one has determined that and one knows the decay of memory than one can determine the necessary time interval between the repetitions. That is what has been done in our experiment and we have specified for different groups of respondents and topics what the time interval should be.

Peter Mohler just says that there will be memory effect with a reference to Duncan and Stenbeck (1988,p 523) who said: " In the course of a single interview … there is evidence that responses to one question may contaminate responses to another or , alternatively, that several questions may be vulnerable to common but evidently transitory sources of contamination." They have studied another issue in their paper. So their remark is not based on their research but it is just a general remark without any empirical evidence. So this remark can not be used against our detailed study.

Duane Alwin discusses a word memory experiment that he has done where people had the task to remember 10 words and after that had an immediate memory task and one more after 10 minutes. On the basis of this experiment he derives the conclusion.." there is some memory decay across the 10-15 minutes elapsed between the two tasks. However, if one looks at the delayed task and focuses solely on those words produced in response to the immediate recall task, the impression one gets is that within the context of the survey, people remember what they said earlier. This goes against the claims of Saris and Van de Putte (1988) who argued that repeated measurement at the beginning and end of a 20 minute-long questionnaire can be considered free of memory effects and therefore a basis for estimating reliability (see also Van Meurs and Saris 1990). Those claims were, however based on conjecture rather than empirical evidence …"

I have shown above that our conclusion was a more detailed than the statement of Duane Alwin and was not "based on conjecture". We took this issue very seriously and have done some careful empirical study while the study of Alwin is not comparable with the normal survey situation.

Finally Krosnick writes: " Unfortunately, it is difficult to know on theoretical grounds or based on practical evidence how long the necessary minimum time interval must be. Van Meurs and Saris (1995) reported evidence suggesting that 20 minutes is sufficiently long. … However , it seems unlikely that the needed time interval to assure forgetting of a prior question is uniformly 20 minutes. Most likely, this time interval varies as a function of the particular topic and forms of the questions and intervening events." After that he illustrates his point of view with typical examples which we have also included in our study.

It seems that all three authors have remembered that we said that 20 minutes were enough for decay of memory but not under which conditions this was the case. We have shown above that our study was much more precise and would have deserved more attention. Our conclusions were not as simple as they remember. Below we will discuss the design of the ESS MTMM experiments and we will show that the time interval between the repeated measures in those studies is at least three time longer than the 20 minutes that has been criticized.

Some other problems

Besides the problem of memory effects one can also raise some other issues. One of them is the possibility of effects of item specific factors next to the trait and the method, the possibility of order effects and correlations between the method factors.

Item specific factors

Duane Alwin suggests that item specific factors next to the trait and the method can affect the responses. This is in general true in factor analysis models with different items. However in our MTMM experiments we keep the trait we measure always the same. So the questions differ only in the formulation characteristics like the scale or other form characteristics. This was illustrated in table 2 and table 3. Therefore we argue that the response is indeed only affected by the trait measured and the reaction of the respondents to the form of the question which is the method used.

Order effects

A second issue mentioned by Alwin is order effects. The idea is that people answer questions not independently of each other but they are taking the response to the previous question into account while formulating the answer to the next question. There are indeed experiments done which showed that this is the case. But these questions were especially chosen for this purpose because they were related to each other. For example in Schuman and Presser (1981) one can read about the famous experiments of allowing journalists in Russia and in the US. The order of these questions has an effect on the responses. In such cases with clearly related questions this can be expected but in most surveys such connections between the answers to the sequentially asked questions does not exist. In one of our studies to determine the best model for MTMM data we tested the hypothesis of Alwin in 7 different studies and in all cases it turned out that the MTMM model presented in Figure 1 fitted the data much better than the model with memory effects between responses (Aalberts et al, 2003).

Correlation between the method factors

Another possibility for misspecifications in out model and so in the estimation of the quality coefficients is that we assume normally that the method effects are independent of each other. This seems a strong restriction if one uses for example in a experiment different forms of category scales. One can assume that the reaction of the people is similar independent of the type of the category scale. This would mean that the method factors would be correlated. This point is not only made by Mohler (2011) but has also been considered by Scherpenzeel in her thesis (1995). She looked at the consequences of such correlations. It turns out that the correlations have to be very high in order to be detectable in the data. Normally do these correlations lead to increases of not more than .02 or .03 in the correlation between the observed variables. So, one has to have very large samples to detect these errors. As a consequence our starting model is

always with uncorrelated method factors but we allow for such correlations if the analysis shows that it is necessary.

The critic of the questionnaire design of the ESS

The European Social Survey has decided from the very beginning that not only data will be collected but that also control on the quality should be applied. Besides the control on the sampling (Häder and Lynn 2007) , Nonresponse (Billiet et al. 2010) , Translation (Harkness et al. (2007) the comparability of the questions has also been a concern of the ESS (Saris and Gallhofer 2007b). In the section about the quality coefficient q we have argued that the q^2 in a study has a big effect on the observed correlations. So if the q^2 is different across countries then the correlations obtained for the observed variables will be different as well even if the correlation between the variables of interest (without measurement error) are the same. So in order to be able to compare relationships between variables across countries the quality of the questions should be the same or one should correct for the differences in quality as was indicated above.

Given this argument the ESS decided to introduce in the supplementary questionnaire alternative forms of some questions present in the main questionnaire following the Split ballot MTMM design mentioned above using different randomized subgroups in order to estimate the quality (q^2). In round 3 even four different formulations of the same questions are tested: one in the main questionnaire and three alternatives in three different subgroups of the sample in each country. Peter Mohler discusses the experiment for one trait (concept) in detail. In table 3 the different forms of these questions were presented. Question B38 was presented in the main questionnaire. The other three questions were placed in the supplementary questionnaire, one for each subgroup. Peter Mohler complains about possible memory effects and that a time interval of 20 minutes is probably too short. However in this specific case there were 241 questions between the first measure and any repetition. Assuming that people answer 3- 4 question per minute the time interval between the repeated questions is between 80 and 60 minutes[i]. This is at least 3 times larger than we suggested that is needed for the decay of the memory of the previous answer if there are similar questions in between. In round 3 there were many similar questions with respect to topic and form. So we think that most people have no idea any more what they have said before after more than 60 minutes. This is even more so because the questions in the supplementary questionnaire are formulated differently and so the memory does not help much. That does not mean that we do not expect a correlation between the responses but mainly because of the stability of the opinions of the respondents and not because of memory effects.

Besides this point Peter Mohler mentions that in the ESS not sufficient attention is paid to the similarity of the instruments across countries. He indicates that across the countries the following differences can be found:

- the supplementary questionnaire is self or interviewer administered

- the response scales on the show cards for B38 have very different forms in different countries
- in the self completion version “don´t know” is in some countries mentioned in others not

I agree with Peter Mohler that it would have been much better if the form of the questions in all countries would have been the same. However, unfortunately that is not the case in several papers my research group has indicated that there are differences in the questionnaires in the different countries and as a consequence also in the quality of the questions (Saris and Gallhofer 2007, Oberski et all 2010, Zavala 2011). Consequently it is important to estimate these differences in quality in order to able to correct for these differences. It seems that Mohler thinks that these differences in formulation will devaluate the experiments, however, this is not at all the case because we estimate the quality in each country. So as a matter fact we estimate the consequences of these differences and by that make the correction possible. I would like to add to this that even if the formulation would have been completely comparable I think the experiments would still make sense because we do not know if the reaction of the people on these comparable questions would be the same. So I think that his criticism on these experiments is completely unjustified. He should be in favor of these experiments given the differences that exist.

Criticism of the results of an SRM paper

Peter Mohler presents in his chapter a table with quality estimates of question B38 and questin HS4 metioned in table 2. This table comes from a paper of Saris, Revilla, Krosnick and Schaefer (2010). That paper makes an argument that “ item specific scales” like B38 have much better quality than the very common Agree/disagree items like item HS4. That paper presents first theoretical arguments for this point of view and shows afterwards with several examples that the empirical results are in agreement with that. The paper presents 4 studies of which 3 come from the ESS and one of them is comparing several questions like B38 with questions like HS4,HS5 and HS6. A copy of the table presenting the resulting quality estimates (q2) is presented in table 4. For the full discussion of this topic we refer to the original paper.

Table 4. The quality of the different scales for three different questions in each country

Country	Q1	Q2	Q3
Austria			
IS(11)	,81	,83	,79
A/D (5)	,46	,51	,56
A/D (11)	,32	,37	,46
A/D (7)	,32	,33	,32
Belgium			
IS(11)	,72	,79	,64
A/D (5)	,51	,48	,63
A/D (11)	,24	,35	,41
A/D (7)	,29	,38	,47
Bulgaria			
IS(11)	,71	,81	,85
A/D (5)	,30	,31	,33
A/D (11)	,13	,18	,22
A/D (7)	,22	,29	,32
Switzerland			
IS(11)	,71	,85	,67
A/D (5)	,50	,60	,60
A/D (11)	,20	,46	,36
A/D (7)	,49	,57	,57
Cyprus			
IS(11)	,81	,86	,83
A/D (5)	,47	,55	,47
A/D (11)	,53	,55	,41
A/D (7)	,36	,43	,42
Germany			
IS(11)	,77	,79	,79
A/D (5)	,43	,49	,56
A/D (11)	,32	,41	,51
A/D (7)	,38	,48	,59
Denmark			
IS(11)	,74	,83	,79
A/D (5)	,61	,59	,60
A/D (11)	,40	,53	,55
A/D (7)	,41	,44	,50
Estonia			
IS(11)	,55	,77	,81
A/D (5)	,41	,37	,35
A/D (11)	,17	,22	,25
A/D (7)	,22	,24	,31
Spain			
IS(11)	,83	,77	,69
A/D (5)	,46	,56	,51
A/D (11)	,24	,17	,27
A/D (7)	,21	,28	,43
Finland			
IS(11)	,71	,76	,74
A/D (5)	,60	,52	,63
A/D (11)	,38	,36	,51
A/D (7)	,37	,14	,36
France			
IS(11)	,79	,85	,77
A/D (5)	,55	,64	,61
A/D (11)	,31	,52	,48
A/D (7)	,25	,44	,43
A/D (7)	,31	,36	,42
Ireland			
IS(11)	,77	,77	,81
A/D (5)	,37	,33	,3
A/D (11)	,02	,09	,14
A/D (7)	,16	,12	,27
Latvia			
IS(11)	,81	,90	,86
A/D (5)	,24	,28	,24
A/D (11)	,05	,07	,08
9			
United Kingdom			
IS(11)	,81	,83	,83
A/D (5)	,41	,49	,59
A/D (11)	,28	,38	,44
A/D (7)	,10	,11	,13
Netherlands			
IS(11)	,72	,69	,62
A/D (5)	,38	,35	,47
A/D (11)	,23	,24	,30
A/D (7)	,29	,23	,32
Norway			
IS(11)	,72	,79	,77
A/D (5)	,67	,57	,58
A/D (11)	,09	,32	,43
A/D (7)	,36	,42	,38
Poland			
IS(11)	,69	,81	,67
A/D (5)	,33	,31	,39
A/D (11)	,10	,13	,18
A/D (7)	,19	,20	,18
Portugal			
IS(11)	,83	,81	,86
A/D (5)	,47	,39	,43
A/D (11)	,18	,22	,27
A/D (7)	,40	,35	,45
Romania			
IS(11)	,88	,85	,79
A/D (5)	,29	,39	,44
A/D (11)	,08	,14	,22
A/D (7)	,17	,19	,20
Russia			
IS(11)	,77	,83	,83
A/D (5)	,42	,46	,44
A/D (11)	,36	,33	,34
A/D (7)	,27	,33	,29
Slovenia			
IS(11)	,81	,79	,74
A/D (5)	,37	,36	,38
A/D (11)	,01	,10	,22
A/D (7)	,13	,20	,22
Slovakia			
IS(11)	,67	,69	,56
A/D (5)	,32	,31	,26
A/D (11)	,12	,14	,15
A/D (7)	,14	,22	,16
Ukraine			
IS(11)	,81	,88	,83
A/D (5)	,44	,49	,46
A/D (11)	,17	,20	,25
A/D (7)	,12	,26	,27

Peter Mohler was willing to allow question B38 to pass his benchmarks, without saying what his benchmarks were but he thought that HS4 was inacceptable. The reason was the following (pp 309).
"The context of presentation for HS4 is quite different from item B38 which was part of a series, not a matrix like item battery. That is the first change of "method" in terms of the MTMM design. The next is a dramatically different wording of the question. While B38 asks whether it is "good or bad" for the country, HS4 item asks for "bad " only. The attached agree-disagree answer scale camouflages the uni-dimensionality of item HS4. In doing so, the designers assume that the underlying construct "effect of immigration" appears to have only one – negative dimensionality- or only the negative dimensionality is attempted to be measured by item HS4."

Let me first say that the designers did not think at all that the effects of immigration are only negative. If that was so we had not pressed to use B38 in the main questionnaire. It is also not necessarily true that such an agree- disagree statement can only be seen as measuring the negative effects of immigration. People can still say completely disagree which suggests that they are probably completely at the other end of the scale. What his comment, however, shows is that he does not realize what kind of experiment is done. We wanted to compare the quality of "item specific scales" and "agree/disagree scales using batteries of statements". The latter approach is much more common in the social sciences than the former. A problem of the use of the agree/disagree format with statements is that one has to specify a position of the scale. One can´t use statements like:

How much do you agree or disagree with the statement:

It is generally good or bad for (Country´s) economy that people come to live here from other countries

In an agree disagree statement one has to chose for a position on the relevant scale. This is true for all statements. In this case the scale goes from bad to good. So the designers could have chosen "good" as well instead of "bad". In the paper where the table comes from such an experiment varying the position of the statement with respect to the scale of interest has also been done. This experiment confirmed the idea of Peter Mohler that the quality of a negative formulated items is indeed worse than of a positively formulated item but even in the case of a positive statement the quality is much lower than the quality of the same question formulated with an item specific scale . However I am not aware that this kind of experiment using the q^2 as criterion has been done before.

In the experiment that Peter Mohler discussed and the other 3 experiments discussed we wanted to show that questions using a statement and an agree disagree scale have a rather bad quality and are much worse than items which do not make a choice but ask the respondents to make this choice on an "item specific scale. The

experiment discussed above showed this very clearly because in all countries the difference in quality (q^2) is very large.

Peter Mohler gave three reasons to raise questions about the quality of this experiment: memory effects, order effect and the form of the questions. In the paper it is clear that there were indeed many differences planned and unplanned across the experiments in the 4 experiments and across the countries. Nevertheless in all experiments, in all countries the item specific scales had a much better quality than the agree-disagree forms of the same questions. That brought us to the following conclusion in the paper:

"Many aspects of the design have been manipulated, but still the same conclusion is drawn. It did not matter whether the IS scale (Item Specific scale) was asked before or after the A/D scale (Agree/disagree scale) or at the same time. Even if the A/D scale had more answer categories the IS scale with fewer categories was still of higher quality. The mode in which the questions were asked (face to face or self completion) also did not change this general tendency. So the better quality of IS scales is a quite general and robust result, which holds across different topics, countries, modes and ordering of the questions in the experiments."

So far about this issue.

Peter Mohler also mentions: *"Finally, there is a real surprise in the data. Comparing the Q-values of the three MTMM items (HS4,HS16 and HS24), HS4 has the higher Q values in all countries (Table 2). Why is that surprising? Because HS4 uses a standard fully labeled 5-point likert response scale, while the other two use a 7-point or 11 point response scale with only the end points labeled (see table1) That a 5 point Likert response scale shows the best Q-values is counter to the common wisdom that 11 point scales are the best (see ESS documentation) Thus one might muse how much quality is either in Q or our common wisdom."*

I hope that I have shown above that the quality coefficient q2 (not Q as he suggests incorrectly) is not perfect because it remain an estimate, but it is a useful tool to evaluate the quality of questions. Sure one can have doubts about a single experiment but this result is repeated across many countries and data collection forms and also across different topics as has been shown in a paper which will appear in Sociological Methods and Research. He tries to discredit the MTMM approach with a vague reference to ESS documentation. We prefer to look at our empirical data and they lead us to the suggestion that the quality may increase with the number of categories in item specific scales but for Agree/disagree scales this seems not to be the case. If one analyzes the task the respondent has to do in the latter case that would not come as a surprise (Saris et al. 2009) but we have also empirical evidence that has to be taken seriously.

The quality of the design steps and question quality

I agreed that in the design of a questionnaire each step can go wrong and can lead to loss of quality of the questions. Therefore it is important to know what these decisions are and what their effect is on the quality of the final form of the question. This was the basic idea with which Frank Andrews started his MTMM experiments in 1984 and we continued his research after his death. The procedure is relatively simple but rather tedious. From the MTMM experiments one gets an estimate of the quality of the questions but these questions have many different characteristics due to the choices that have been made in the design process. So the obtained results are rather specific for each question. However by meta-analysis across all questions evaluated one can get an impression of the effects of the different choices on the quality of questions. This requires that all the choices made in the design of the questions are coded.

Such a coding system and meta analysis across MTMM experiments was first done by Andrews (1984) for experiments in the US . Next such a meta analysis was done by Költringer 1993 for Austrian data and by Scherpenzeel (1995) for Dutch data and by Saris and Gallhofer (2007) over all three data sets using the same coding schema for all three different studies.

The last meta analysis has also led to a program Survey Quality Prediction (SQP) that uses the results of the latter meta analysis to make predictions for the quality of new questions after they have been coded on the same characteristics (oberski etal. 2005). This program can be used to make predictions of the quality of questions before they are used in the field. If the quality is not good enough the program can give suggestions how the question can be improved. This program has been used in the ESS to test the quality of questions before they go into the field.

The nice thing of the ESS is also that in every round 6 MTMM experiments have been included in order to evaluate specific questions of the ESS. These experiments have been done in all participating countries. As a consequence we have now more than 4000 questions of the ESS for which the quality has been estimated while we have 1000 extra questions from the earlier studies. All the questions of the first third rounds which were involved in the MTMM experiments have been coded using the SQP program with respect to the questions characteristics and at the moment we are preparing a new version of the SQP program which will be based on the meta analysis of more than 3000 question from many different countries. This new version of SQP will be out before the end of the year and will make predictions of the quality of questions for more than 20 countries and languages.

In the coding of all the questions of the first four rounds of the ESS we have observed many differences in the implementation of the questions with respect to the

form of the scales, formulation of the labels, form of the show cards, use of batteries of not, specifying a non response option or not , the data collection made etc. In the meta analysis these differences have been taken into account. Besides that they have led to a discussion in the ESS about the comparability of the questions and the conclusion has been that now standard in each round a limited number of the translated questions as formulated in the different countries are coded with SQP in order to see, by comparing with the codes of the source questionnaire, whether the translated questions do not deviate in form from the formulation of the same questions in the source questionnaire. One has decided that this process is important because these form differences will lead to differences in quality of the questions across countries and that will make comparative research across countries impossible if no correction is made for these differences in quality.

This section indicates that the MTMM approach not only provides information about the quality of the question by the q^2 but that the MTMM research has lead to information about the quality of each step in the questionnaire design process. This information, provided in the program SQP, can be used to improve the decisions made in the different stages of the process.

Conclusions

I have felt obliged to comment on the criticism of the MTMM approach developed and used by my research group and used in the ESS. My opinion is that the criticism is partially incorrect by ignorance and partially by incomplete knowledge of the studies of difficult issues. All these studies have been published in reviewed journals. Because we think that these comments harm the MTMM approach, the valuable work done in the ESS and the results of a paper published in SRM, I have decided to give this reaction. I have used this opportunity to explain in a very simple way what the importance is of the q coefficient in general and especially in comparative survey research. I have also clarified what the basic idea behind the crucial study of the memory effect is. This is indeed an essential point in the evaluation of the MTMM approach. I hope that I have clarified these different issues. For more information I refer to book of Saris and Gallhofer (2007) where all these issues have been discussed in detail. I can also refer to a completely independent source Lance et. al. (2010).

References

Alwin D. (1974) Approaches to the interpretation of relationships in the multitrait.multimethod matrix. In : *Costner H.L. (Ed) Sociological Methodology 1973-74.* San Francisco, Jossey-Bass, pp 79-105

Alwin D. (2011) Evaluating the reliability and validity of Survey Interview Data Using the MTMM Approach.In Jennifer Madans, Kristen Miller, Aaron Maitland and Gordon Willis (Eds) *Question Evaluation Methods*, Wiley, 2011, 265-295

Andrews F. (1984) Construct validity and error components of survey measures: A structural equation approach. *Public Opinion Quarterly*, 48, 409─442.

Biemer PP and Lyberg L. (2003) *Introduction into Survey Quality*. Hoboken, NJ: Wiley

Campbell D.T. and Fiske D.W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychological bulletin,* 56 , pp 81-105.

Campbell J.P. (1982) Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology,* 67, 691-700.

Couper M. (2008) *Designing effective Web Surveys.* Cambridge, Cambrigde University Press

Dillman D. A. 2000. *Mail and Internet Surveys. The Tailored Design Method.* New York: Wiley.

Dillman D.A., Smyth J.D. Christian L.M. (2009) *Internet, Mail and Mixed Surveys. The Tailored design methods*, Hoboken , Wiley

Duncan O.D. and M.Stenbeck (1988) No opinion or not sure? *Public Opinion Quarterly*, 52: 513-525.

Häder S. and P.Lynn (2007) How representative can a multi-nation survey be? In Jowell R., C.Roberts, R.Fitzgerald and G.Eva (Eds) *Measuring Attitudes Cross-nationaly: Lessons from the European Social Survey*. London, Sage. 33-53.

Harkness J.A. (2007) Improving the comparability of translations. In Jowell R., C.Roberts, R.Fitzgerald and G.Eva (Eds) *Measuring Attitudes Cross-nationaly: Lessons from the European Social Survey*. London, Sage. 79-95.

Jöreskog K.(1969) A general approach to cinfirmatory maximum likelihood factor analysis. Psychometrika 34: pp 183-202

Költringer R. 1993. *Gültigkeit von Umfragedaten*. Wien: Bohlau.

Krosnick J. (2011) Experiments for evaluating questions. In Jennifer Madans, Kristen Miller, Aaron Maitland and Gordon Willis (Eds) *Question Evaluation Methods*, Wiley, 2011, 215- 239

Lawley D. N., and A. E. Maxwell 1971. *Factor Analysis as a Statistical Method*. London: Butterworth.

Lord F.M. and M.R.Novick (1968) *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lance C.E., Dawson B. Birkelbach D. and Hoffman B.J. (2010) Method effects, measurement error and substantive conclusions. *Organizational Research Methods* , 13, pp 435-455

Madans J. , Miller K, Maitland A and G.Willis (2011) (Eds) *Question Evaluation Methods*, Wiley, 2011

Mohler P. et al. (2010) A survey process quality perspective on documentation. In: Harkness J.A. Braun M., Edwards B. Johnson T.P. Lyberg L. Mohler P. Pennell B E. Smith T.W. (Eds) *Multinational Multicultural and Multiregia¡onal Survey Methods*. Hoboken, Wiley, pp 299-314

Mohler P. (2011) Response to Alwin´s Chapter: Evaluating the reliability and validity of Survey Interview Data Using the MTMM Approach. In Jennifer Madans, Kristen Miller, Aaron Maitland and Gordon Willis (Eds) *Question Evaluation Methods*, Wiley, 2011, 295-319

Oberski D., L. Kuipers, and W.E. Saris 2005. *SQP Survey Quality Predictor*. www.sqp.nl

Oberski D. Saris W.E. , Hagenaars J.A. (2010) Categorization errors and differences in the quality of questions in comparative surveys. In Harkness J.A. Braun M., Edwards B. Johnson T.P. Lyberg L. Mohler P. Pennell B E. Smith T.W. (Eds) *Multinational Multicultural and Multiregia¡onal Survey Methods.* Hoboken, Wiley, pp 435-453.

Pennell B.E. et al. (2010) *Cross Cultural Survey Guidelines* (CCSG) . Ann Arbor, MI, ISR.

Presser S. Rothgeb J.M. Couper M.P. Lessler J.T. Martin E. Martin J. Singer E. (Eds) (2004) Methods for testing and evaluating Survey Questionnaires. Hoboken, Wiley.

Saris W.E.. and Van den Putte B. (1988) True score of factor models: a secondary analysis of the ALLBUS test-retest data. *Sociological Methods and Research*, 17. pp 123-157

Saris W.E. and Van Meurs A. (1990) *Evaluation of measurement instruments by Meat-analysis of Multitrait Multimethod studies*. Amsterdam North-Holland

Saris W. E., and F. M. Andrews 1991. Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley, 575─599.

Saris W.E. and C.Aalberts (2003) Different explanations for correlated disturbance terms in MTMM studies. In *Structural Equation Modeling* , 10, 193- 214.

Saris W.E. and I.Gallhofer (2007) The results of the MTMM experiments in round 2 of the ESS. *RECSM working paper*.

Saris W. E., A. Satorra, and G. Coenders 2004. A new approach for evaluating quality of measurement instruments. *Sociological Methodology*, 3, 311─347.

Saris, W.E. and I. Gallhofer (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York, NY: John Wiley & Sons, Inc. 2007.

Saris W.E. and I.Gallhofer (2007b) Can questions travel successfully? In Jowell R., C.Roberts, R.Fitzgerald and G.Eva (Eds) *Measuring Attitudes Cross-nationaly: Lessons from the European Social Survey*. London, Sage. 53-79

Saris, Willem E., Revilla, Melanie, Krosnick, Jon A., Shaeffer, Eric M. (2010). "Comparing Questions with Agree/Disagree Response Options to Questions with Construct-Specific Response Options" *Survey Research Methods* 4(1):61-79

Scherpenzeel A. C. 1995. *A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies.* KPN Research: Leidschendam.

Schuman H and Presser S. (1981) *Questions and answers: Experiments in question wording, form and context.* NewYork: Academic Press.

Stoop I. Billiet J. Koch A. Fitzgerald R. (2010) *Improving nonresponse*, Hoboken, Wiley.

Van Meurs A., and W. E. Saris 1990. Memory effects in MTMM studies. In W. E. Saris and A. van Meurs (eds.), *Evaluations of Measurement Instruments by Metaanalysis of Multitrait-Multimethod Studies,* Amsterdam: North Holland, 134-146.

Zavala D. (2011) Avoiding deviations across countries due to questionnaire translations through SQP. *RECSM Working paper*

[i] An interval is given because in one part of the questionnaire skips were occurring