UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 2

2009

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici de França - Despatxs 70.374-70.382
Passeig de Circumval·lació, 8 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

The MTMM approach to coping with measurement errors in survey research

Willem E. Saris

Universitat Pompeu Fabra

Barcelona, Spain

Abstract

Designing a survey involves many more decisions than most researchers realize. Some survey specialists therefore talk about the art of designing survey questions (Payne 1951). Designing a survey in a scientific way requires knowledge of how the many decisions that researchers take in survey design affect the quality of questions. Many studies have been done in this area. Inspired by the work of Frank Andrews (1984), we elaborated his Multitrait Multimethod (MTMM) approach to evaluating the quality of questions. On the basis of a meta-analysis of a large number of MTMM experiments, we developed a program (SQP) that can predict the quality of questions before they are used in the field (Saris and Gallhofer, 2007). In this paper we will briefly explain the approach we have chosen, illustrate the method and discuss its advantages and disadvantages.

Introduction

Over the last 40 years, many studies have been performed to evaluate the quality of survey questions. Most studies use random assignment of respondents to different question forms to see whether the form of the question makes a difference. These so called "split ballot experiments" have been used by Schuman and Presser (1981) and many others in the social sciences. Molenaar (1986) studied the quality of questions using nonexperimental research. In official statistics, test-retest models have been popular in evaluating questions (Forsman 1989). Heise (1969), Wiley and Wiley (1970), Alwin and Krosnick (1991) and Alwin (2007) used the quasi-simplex model based on panel data to evaluate the quality of questions. The testing of questions in cognitive laboratories has recently received a great deal of attention. As well as all these approaches, an alternative was applied by Frank Andrews (1984) which is called the Multitrait Multimethod or MTMM approach. After the death of Frank Andrews, his work was continued by European researchers (Scherpenzeel 1996, Scherpenzeel and Saris 1997, Coenders and Saris 2002, Corten and Saris, Aalberts and Saris 2002, Saris, Satorra and Coenders (2007), and finally led to a summary of this research in a book by Saris and Gallhofer (2007) which also introduces a computer program (SQP) that can predict the quality of questions before data are collected in the field (Oberski, Kuipers and Saris 2005). In this paper, we concentrate on the MTMM approach. We will first explain what we mean by quality of a question, and then we will introduce the MTMM design and model. We will illustrate the approach and discuss its advantages and disadvantages.

Quality criteria for survey measures

The first quality criterion for survey items is *item non-response*. This is an obvious criterion, because missing values have a disrupting effect on the analysis, which can lead to results that are not representative of the population of interest.

A second criterion is *bias*, which is defined as a systematic difference between the real values of the variable of interest and the observed scores corrected for random measurement errors[1]. Real values can be obtained for objective variables and therefore the most preferable method is the one that provides responses corrected for random errors which are closest to the real values. A typical example comes from voting research. Participation in the elections is known after the elections. This result can be compared with the results obtained from survey research performed using various methods. It is a well-known fact that participation is overestimated when standard survey methods are used. A new method that does not overestimate the participation or produces a smaller bias is therefore preferable to the standard procedures.

In the case of subjective variables, in which the real values are not available, it is only possible to study the various distributions of responses for different methods. If differences between two methods are observed, at least one method is biased; however, it is also possible that both are biased.

These two criteria have received a lot of attention in split-ballot experiments. See Schuman and Presser (1981) for a summary. Molenaar (1986) studied the same criteria while focusing

[1] This simple definition serves the purpose of this text. However, a precise definition can be found in Groves (1989).

on non-experimental research (1986). In short, these criteria describe the observed differences of nonresponse and differences of response distributions.

Other quality criteria which have also been discussed at length are *reliability*, *validity,* and the *method effect*. Reliability is the complement of random errors and validity is the complement of systematic errors. Both criteria have been discussed extensively in psychology and other social sciences as criteria for the quality of measures. There are many different definitions of these criteria. Below e give the definitions which have been used in the MTMM literature for some considerable time, starting with a paper by Saris and Andrews (1991)

In order to do so we present a measurement model for two variables of interest, such as "satisfaction with the government" and "satisfaction with the economy." The measurement model for the two variables is presented in Figure 1.

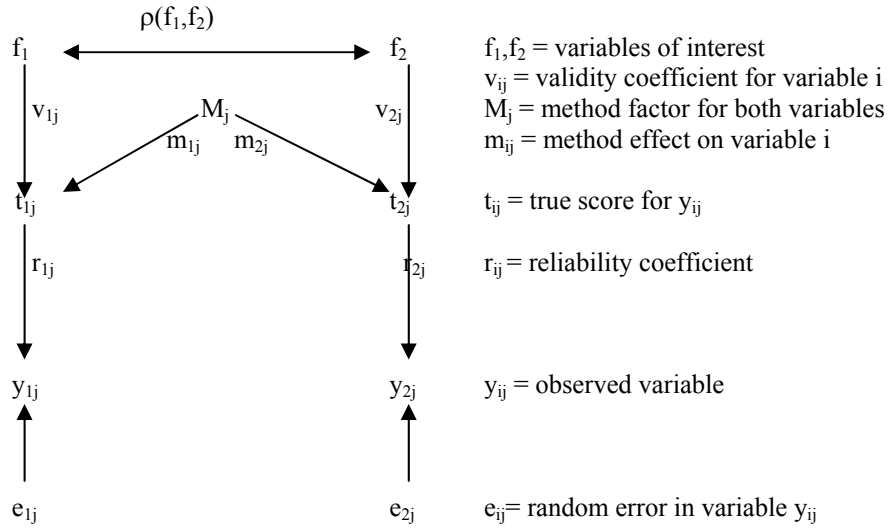


Figure 1: **The measurement model for two traits measured using the same method.**

In this model it is assumed that

- f_i is the trait factor i of interest measured by a direct question.
- y_{ij} is the observed variable (variable or trait i measured by method j).
- t_{ij} is the "true score" of the response variable y_{ij}.
- M_j is the method factor that represents a specific reaction of respondents to a method and therefore generates a systematic error.
- e_{ij} is the random measurement error term for y_{ij}.

The r_{ij} coefficients represent the standardized effects of the true scores on the observed scores. This effect is smaller if the random errors are larger. This coefficient is called the *reliability coefficient*. *Reliability* is defined as the strength of the relationship between the observed response (y_{ij}) and the true score (t_{ij}), that is r_{ij}^2 .
The v_{ij} coefficients represent the standardized effects of the variables of interest on the true scores for the variables that are in fact measured. This coefficient is therefore called the *validity coefficient.* V*alidity* is defined as the strength of the relationship between the variable of interest (f_i) and the true score (t_{ij}), that is v_{ij}^2.

The m_{ij} coefficients represent the standardized effects of the method factor on the true scores, called the *method effect*. An increase in the method effect results in a decrease in validity and vice versa. It can be shown that for this model $m_{ij}^2 = 1 - v_{ij}^2$, and therefore the method effect is equal to the invalidity due to the method used. The *systematic method effect* is the strength of the relationship between the method factor (M_j) and the true score (t_{ij}) denoted by m_{ij}^2.

The *total quality of a measure* is defined as the strength of the relationship between the observed variable and the variable of interest, that is $(r_{ij}v_{ij})^2$.
The *effect of the method on the correlations* is equal to $r_{1j}m_{1j}m_{2j}r_{2j}$.

The reason for using these definitions as quality criteria becomes evident after examining the effect of the characteristics of the measurement model on the correlations between the observed variables.

It can be shown that the correlation between the observed variables $\rho(y_{1j},y_{2j})$ is equal to the combined effect of the variables that we want to measure (f_1 and f_2) plus the spurious correlation due to the method factor as demonstrated in formula (1):

$$\rho(y_{1j},y_{2j}) = r_{1j}v_{1j}\,\rho(f_1,f_2)v_{2j}r_{2j} + r_{1j}m_{1j}m_{2j}r_{2j} \qquad (1)$$

Note that r_{ij} and v_{ij} , which are always less than 1, will decrease the correlation (see first term) while the effects of the method, if they are not zero, can generate an increase in the correlation (see second term).

If there are only two observed variables, the quality criteria and the correlation between the variables of interest cannot be estimated. A design for data collection is therefore needed that provides more information so that the parameters of the model can be identified.

The classical MTMM design and model

Campbell and Fiske (1959) suggested using multiple traits and multiple methods (MTMM). The classic MTMM approach recommends using at least three traits that are measured with three different methods, leading to nine different observed variables. An example of such a design is presented in Table 1.

Table 1. The classic MTMM design used in the ESS pilot study

The three traits were presented by the following three questions:
- *On the whole, how satisfied are you with the present state of the economy in Britain?*
- *Now think about the national government. How satisfied are you with the way it is doing its job ?*
- *And on the whole, how satisfied are you with the way democracy works in Britain?*

The three methods are specified by the following response scales:

(1) Very satisfied; (2) Fairly satisfied; (3) Fairly dissatisfied; (4) Very dissatisfied

Very dissatisfied *Very satisfied*
0 1 2 3 4 5 6 7 8 9 10

(1) Not at all satisfied; (2) Satisfied; (3) Rather satisfied; (4) Very satisfied

Using this MTMM design, data for nine variables are obtained and a correlation matrix of 9×9 is obtained from those data. The model formulated to estimate the reliability, validity, and method effects is an extension of the model presented in Figure 1. Figure 2

illustrates the relationships between the true scores and the general factors of interest. Figure 2 shows that each trait (f_i) is measured in three ways. It is assumed that the traits are correlated but that the method factors (M_1, M_2, M_3) are not correlated because the reactions will be different for different methods. To reduce the complexity of the figure, no indication is given that for each true score there is an observed response variable that is affected by the true score and a random error, as was previously introduced in the model in Figure 1. However, these relationships, although not made explicit, are implied.

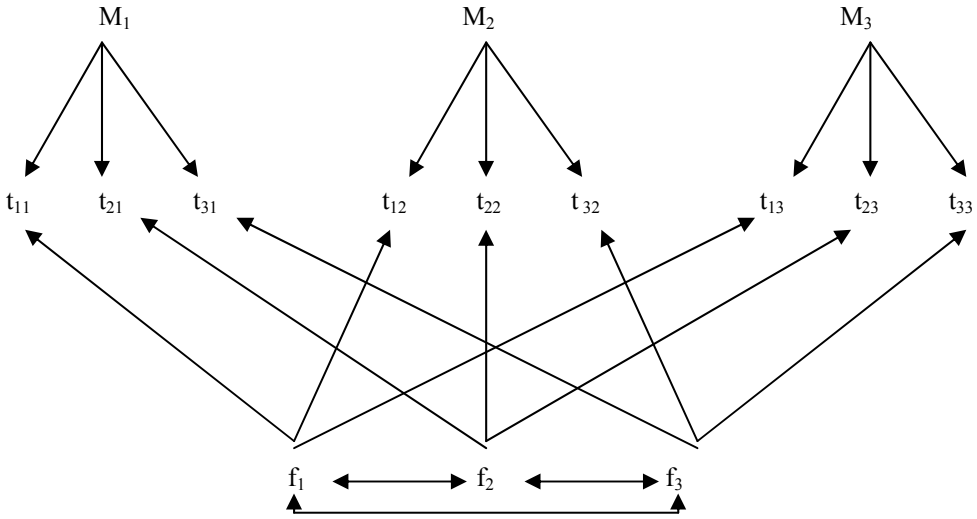


Figure 2: *MTMM model illustrating the true scores and their factors of interest.*

It is normally assumed that the correlations between the factors and the error terms are zero, but there is some debate regarding the actual specification of the correlations between the different factors. Some researchers allow for all possible correlations between the factors, while mentioning estimation problems[2] (Kenny and Kashy 1992; Marsh and Bailey 1991; Eid 2000). Andrews (1984), Saris (1990) and Saris and Andrews (1991) suggest that the trait factors can be allowed to correlate, but should be uncorrelated with the method factors, while the method factors themselves are uncorrelated. When this latter specification is used, combined with the assumption of equal method effects for each method, almost no estimation problems occur in the analysis. This was demonstrated by Corten et al. (2002) in a study in which 79 MTMM experiments were reanalyzed.

The MTMM design of 3 traits and 3 methods generates 45 correlations and variances. In turn, these 45 pieces of information provide sufficient information to estimate 9 reliability and 9 validity coefficients, 3 method effect coefficients and 3 correlations between the traits. There are a total of 24 parameters to be estimated. This leaves 45 ─ 24 = 21 degrees of freedom, meaning that the necessary condition for identification is fulfilled. It also can be shown that the sufficient condition for identification is satisfied, and given that df=21, a test of the model is possible.

Many alternative models have been suggested for MTMM data. A review of some of the older models can be found in Wothke (1996). Among these is the *confirmatory factor analysis* model for MTMM data (Althauser et al. 1971; Alwin

[2] This approach lends itself to non-convergence in the iterative estimation procedure or improper solutions such as negative variances.

1974; Werts and Linn 1970). An alternative parameterization of this model was proposed as the *true score* (TS) model by Saris and Andrews (1991), while the *correlated uniqueness* model has been suggested by Kenny (1976), Marsh (1989), and Marsh and Bailey (1991). Saris and Aalberts (2003) compared models presenting different explanations for the correlated uniqueness. Models with *multiplicative method eff*ects have been suggested by Campbell and O'Connell (1967), Browne (1984), and Cudeck (1988). Coenders and Saris (1998, 2000) showed that the multiplicative model can be formulated as a special case of the correlated uniqueness model of Marsh (1989). We suggest the use of the *true score (TS) MTMM model* specified by Saris and Andrews (1991) because Corten et al. (2002) and Saris and Aalberts (2003) have shown that this model has the best fit for large series of data sets for MTMM experiments. The classic MTMM model is locally equivalent with the TS model, meaning that the difference is only in its parameterization. See Appendix 1 for more details on why we prefer this model.

The Classical MTMM approach has its disadvantages. If each researcher performed MTMM experiments for all the variables of his/her model, it would be very inefficient and expensive, because he/she would have to ask six more questions to evaluate three original measures. In other words, the respondents would have to answer the questions about the same topic on three different occasions and in three different ways. This raises the questions of whether this type of research can be avoided; if this research is really necessary, and whether or not the work of the respondents can be reduced.

Most MTMM experiments to date have used the classic MTMM design or a panel design with two waves, in which each wave had only two observations for the same trait, while at the same time the order of the questions was random for the different respondents (Scherpenzeel and Saris 1997). The advantage of the latter method is that the response burden of each wave is reduced and the strength of opinion can be estimated (Scherpenzeel and Saris 2006). The disadvantages are that the total response burden is increased by one extra measure and that a frequently observed panel is needed to apply this design. Although this MTMM design has been used in many studies because of the presence of a frequently observed panel (Scherpenzeel 1995), we feel that this is not a solution that can generally be recommended. Given the limited possibilities of this particular design, other types of designs have therefore been produced, such as the split-ballot MTMM design (Saris, Satorra and Coeders 2004), which will be discussed in the next section.

The split-ballot MTMM design

In the commonly used split-ballot experiments, random samples from the same population receive different versions of the same questions. In other words, each respondent group gets one method. The split-ballot design makes it possible to compare the response distributions of the various questions and to assess their possible relative biases (Schuman and Presser 1981; Billiet et al. 1986).

In the split-ballot MTMM design, random samples of the same population are also used but with the difference that these groups receive two different forms of the same question. In total there is one less repetition than in the classical MTMM design and one more than in the commonly used split-ballot designs. We will show that this design combines the benefits of the split-ballot approach and the MTMM approach in that it enables researchers to evaluate measurement bias, reliability, and validity simultaneously, and that it does so while reducing the response burden. The suggestion to use split-ballot designs for structural equation models can be traced back to Arminger and Sobel (1991).

The two-group split-ballot MTMM design is structured as follows. The sample is split randomly into two groups. One group has to answer three survey items formulated using method 1, while the other group is given the same survey items presented in a second form, called "method 2." in the MTMM literature In the last part of the questionnaire all respondents are presented with the three items, which are now formulated in method 3 format. The design can be summarized as shown in Figure 3.

	Time 1	**Time 2**
Sample 1	Form 1	Form 3
Sample 2	Form 2	Form 3

Figure 3 *The two-group split-ballot MTMM design.*

In short, in the two-group design the researcher draws two comparable random samples from the same population and asks three questions about at least three traits in each sample: once with the same method and once with another form (method) of the same questions (traits) after sufficient time has elapsed. Van Meurs and Saris (1990) have demonstrated that the effects of memory are negligible after 20 minutes. This time gap is enough to obtain independent measures in most circumstances.

The design in Figure 3 matches the standard split-ballot design at time 1 and thus provides information on the differences in response distributions between the methods. Combined with the information obtained at time 2, this design provides extra information. The question of whether the reliability, validity and method effects can be estimated from this data still remains, since each respondent answers only two questions about the same trait and not three, as required for the classical MTMM design. The answer is not immediately evident, since the information necessary for the 9×9 correlation matrix comes from different groups and is by design incomplete (see Table 2). Table 2 shows the groups that provide data for estimating variances and correlations between questions using either the same or different forms (methods).

Table 2: Samples providing data for correlation estimation

	Method 1	**Method 2**	**Method 3**
Method 1	Sample 1		
Method 2	none	Sample 2	
Method 3	Sample 1	Sample 2	Sample 1+2

In contrast to the classical design, no correlations are obtained for form 1 and form 2 questions, as they are missing by design. Otherwise, all correlations in the 9×9 matrix can be obtained on the basis of two samples, but the data come from different samples.

Each respondent is given the same questions only twice, reducing the response burden considerably. However, the correlations between forms 1 and 2 cannot be estimated, leading to a loss of degrees of freedom when estimating the model on the now incomplete correlation matrix. This might make the estimation less efficient than

the standard design in which all correlations are available, as in the three-group design. In large surveys the sample can be split into more subsamples and more than one set of questions hence evaluated. For more details of this approach, see Saris et al. (2004)

Estimating and testing models for split-ballot MTMM experiments

The split-ballot MTMM experiment differs from the standard approach in that different equivalent samples of the same population are studied instead of just one. Given that the random samples are drawn from the same population, it is natural to assume that the model is exactly the same for all respondents and the same as the model specified in Figure 2, which includes the restrictions on the parameters suggested by Saris and Andrews (1991). The only difference is that not all questions were asked in every group.

Since individuals were assigned to groups at random, and there is a large sample in each group, the most natural approach for estimation is the multiple -group SEM method (Jöreskog 1971). This approach is available in most SEM software packages. We refer to this approach as a multiple-groups structural equation model or MGSEM[3]. As stated above, a common model is fitted across the samples, with equality constraints for all the parameters across groups. With the current software, and applying the theory for multiple-group analysis, estimation can be made by using the maximum likelihood (ML) method or any other standard estimation procedure in SEM. In the case of non-normal data, robust standard errors and test statistics are available in the standard software packages. For a review of multiple-group analysis in SEM models as applied to all the designs, see Satorra (1992, 2000).

The incomplete data set-up we are dealing with could also be considered as a missing data problem (Muthen et al. 1987). However, the approach for missing data assumes normality, while this design does not provide the theoretical basis for robust standard errors and corrected test statistics that are currently available in MGSEM software. Since the multiple-group option therefore offers the possibility of standard errors and test statistics which are protected from non-normality, we suggest that the multiple-group approach is preferable.

Given this situation, we suggest the MGSEM approach for estimating and testing the model using SB-MTMM data. In doing so, the covariance matrices are analyzed while the data quality criteria (reliability, validity coefficients and method effects) are obtained by standardizing the solution.

Although the statistical literature suggests that data quality indicators can be estimated using the SB-MTMM designs, we need to be careful when using the two group designs with incomplete data, because they may lead to empirical underidentification problems (Saris et al 2004). However under normal circumstances the model is identified and all parameters can be estimated. We will illustrate this approach below.

Application: Comparison of Agree / disagree scales with item specific scales

We will illustrate the MTMM approach using a recent example in the European Social Survey (ESS). This survey is nowadays performed in most European countries. Methodological evaluation of the questions has been built into the data collection in this

[3] Because each group will be confronted with partially different measures of the same traits, some software packages for multiple-group analysis will require some tricks to be applied. This is the case for LISREL, where the standard approach expects the same set of observable variables in each group. A simple trick to handle such a situation was described in the early work of Jöreskog (1971) and in the manual of the early versions of the LISREL program; such tricks are also described in Allison (1987). Multiple-group analysis with the software EQS, for example, does not require the same number of variables in the different groups.

large scale project, in a supplementary questionnaire which follows the main questionnaire. In this supplementary questionnaire, 6 MTMM experiments are normally specified for 6 sets of 3 questions in the main questionnaire. 54 questions are therefore in each round of the ESS evaluated using a Split Ballot MTMM design, in order to reduce the response burden for the respondents and also to reduce the memory effects.

The example given here is linked to the comparison of Item-specific scales with agree/disagree scales. Respondents are often asked, using a battery format, how much they agree or disagree with different statements. It is also possible to ask such questions directly. For example, in the third round an 11-point item-specific-scale was used for three items in the main questionnaire of the ESS. The questions in the main questionnaire were formulated as follows:

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

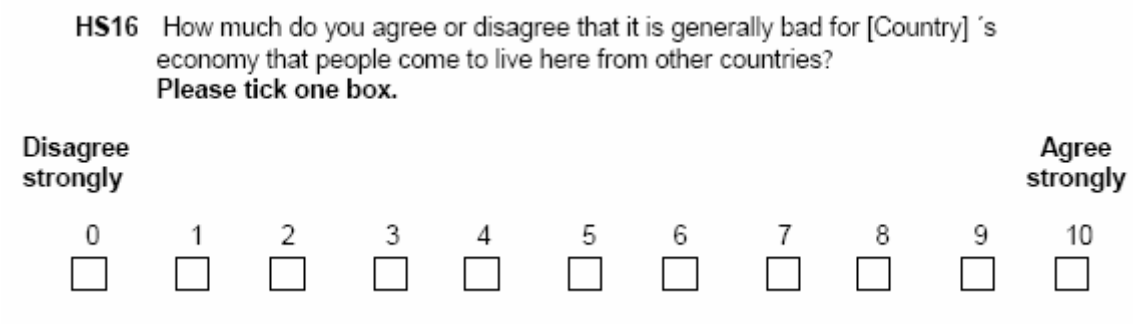
This scale was compared with a 5 point, 7 point and an 11 point A/D scale. These three scales were presented to three random subgroups in the sample in the supplementary questionnaire.

The following statements were presented to the first subgroup in combination with a standard 5 point agree/disagree scale:

> 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
> HS5 [Country´s] cultural life is generally undermined by people coming to live here from other countries

HS6 [Country] is made a worse place to live by people coming to live here from other countries

The second subgroup was also confronted with a battery of agree/disagree statements but now the scale was an 11-point scale. The formulation was as follows:



The same scale is used then for two other items:

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?
How much do you agree or disagree that [Country] is made a worse place to live by people coming here from other countries?

An agree/disagree scale was also used in the third group, but this time it was the same 7-point scale as in the previous experiment, while the items were formulated in the same way as in the first group of this experiment.

In order to give an impression of the differences in quality of the different scales in the different countries, we present the quality for the three questions for the four types of scales in each country in table 3.

Table 3 about here

This table shows that the IS scale is much better than any of the other measures for all questions in all countries studied.

In order to give an impression of the differences in quality across the different countries, the average quality across all countries is presented in table 4

Table 4 about here

This table shows that the difference in quality between the item specific scales and the agree/disagree scales is very big.

In a recent paper, we summarized the results of several such experiments (Saris, Revilla, Krosnick and Schaefer 2009). They all give the same picture across all European countries. We believe that this is a very strong indication that the social sciences have relied too much on agree/disagree scales, because their quality is much worse than item-specific scales. This is also a nice example of the type of results that can be obtained by MTMM experiments.

Many MTMM experiments have been carried out in recent decades (Scherpenzeel 1995). These experiments have provided information about the reliability and validity of 1087 questions. These questions were coded with respect to their characteristics and a meta-analysis was subsequently performed to determine the effect

of the question characteristics on the quality criteria. The results of the meta-analysis have been reported in the book by Saris and Gallhofer (2007) which also introduces a program (SQP) for the prediction of the quality of questions based on this meta-analysis (Oberski et al 2005).

Alternative approaches to estimating quality or questions

While the MTMM approach is an attractive approach for subjective variables, it is less attractive for objective variables such as the background variables age, education, income, occupation etc. The problem is that it is difficult to repeat these questions in a different form and without a memory effect. It is for this reason that other procedures have been developed for these questions. We will discuss these below and comment on their advantages and disadvantages.

Test-retest reliability

A very popular idea is that the reliability of a question can be determined by repeating the same observation twice using the model shown in Figure 4 for the analysis.

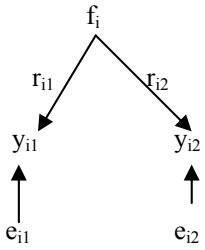


Figure 4: *The standard test-retest model.*

Here, f_i is the variable to be measured and y_{i1} and y_{i2} are the responses to the question used to measure this variable. This approach requires that the same method be used on two occasions. If the model holds true, then the correlation between the two variables can only be due to the product of the two reliability coefficients of the two measures:

$$\rho_{yi1,yi2} = r_{i1} \cdot r_{i2}$$

However, since the same measure is used twice, we can assume that $r_{i1} = r_{i2}$ and then it follows that the reliability = $r_{i1}^2 = r_{i2}^2 = \rho_{yi1,yi2}$. In this case, the reliability of the measure is equal to the test-retest correlation.

However, the model above is too simple when discussing subjective variables. In this case, it is better to start with the model shown in Figure 5.

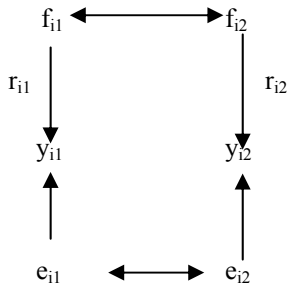


Figure 5: *A more realistic test-retest model.*

The difference with the previous model is that a distinction is made between the latent variable for the first and second measures, accounting for a change that might have occurred while conducting the two observations. In addition, the possibility that respondents might remember their first answers is left open, and indicated by a correlation between the error terms. In order to move from this model to the earlier model, the following assumptions have to be made:

1. No change in the variable of interest between the first and the second measurements
2. No memory effects
3. No method effects
4. Equal reliability for the different measures of the same trait

The standard test-retest approach is unrealistic for subjective variables. If the time between the repetitions is too short, we can expect a memory effect (assumption 2) and if the time is too long, the opinion may be changed (assumption 1). Finally, any possible method effects cannot be detected (assumption 3). This approach is therefore not an accurate representation of reality for subjective variables. Although many people think that it is a robust procedure, it is based on a number of unattainable assumptions, and a less restricted approach is needed.

However if one can be sure that the situation is not changing rapidly, as is the case with background variables, the test-retest approach is not impossible because the variable of interest remains the same (assumption 1) and the repetition of the question can be delayed for long enough so that the second answer does not depend on the previous answer (assumption 2). The reliability remaining the same also seems plausible (assumption 4). The only question that therefore remains is whether there is a method effect. If there is a method effect, the reliability of the questions will be overestimated.

The quasi-simplex approach

In 1969, Heise suggested that the approach mentioned above can be made more manageable for subjective variables by using three observations instead of two. His approach has been improved upon by Wiley and Wiley (1970) and used by Alwin and Krosnick (1991) to evaluate measurement instruments for survey research. Its advantage is that it is no longer necessary to assume that no change has occurred, and it is suggested that the memory effect can be avoided by making the time gap between the observations so long that a memory effect can no longer be expected. Figure 6 shows the suggested model. In Figure 6 "s" is the stability coefficient and "r" is the reliability coefficient.

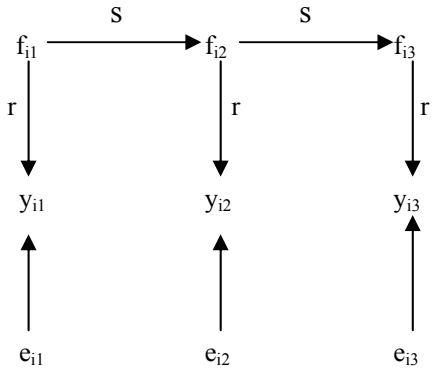


Figure 6: *The quasi-simplex model for three repeated observations.*

However his approach has two major problems with regard to subjective variables. First, it assumes that it is not possible that considerations associated with the variable of interest are forgotten at time 2, but return at time 3. This would suppose that there is an effect of f_{i1} on f_{i3}.

However, this effect is not possible for technical reasons in his model. However, if these effects exist, incorrect estimates of the quality of the measures will be obtained, as discussed by Coenders et al. (1999).

The second problem is that any temporary component in the variables that is not present on the next occasion will be treated as an error, even if it is a substantive part of the latent variable at a given point in time. For example, if we ask about "life satisfaction" and the respondent is in a bad mood, that person's score will be lower than if the same respondent is in a good mood on a different occasion. The mood component is a real part of the satisfaction variable, but because mood changes rapidly, this component will end up in the error term. The error term therefore increases and reliability decreases. This is not because of a lack of data quality, but because of the instability of a component within the variable of interest. For further discussion of this point, see Van der Veld (2006). However, this would not occur if the measures were conducted quickly in the same survey, but then memory effect might emerge again. For these reasons this approach is not preferable for defining the reliability coefficient for subjective variables.

As regards objective variables which only change very slowly and are not affected by incidental changes, this approach could be a good procedure because the change can be estimated and there is no problem of incidental fluctuations. A panel studied with a low frequency of measurement such as observation on a yearly basis therefore works very well for the evaluation of the reliability of the question for objective variables. The only problem left with this approach is that it is impossible to detect a method effect using this approach and the quality of the questions may therefore be overestimated.

Testing external validity

In order to evaluate the validity of different measures for the same variable, suggestions have included using the correlation with other variables that are known to correlate with the variable of interest. The measure with the highest correlation is then the best estimate. According to this line of reasoning, this approach is modelled in Figure 7. In this Figure, ρ is the correlation between the variable of interest and the external criterion variable (x). The other coefficients retain the meanings discussed above.
From this model it follows that:

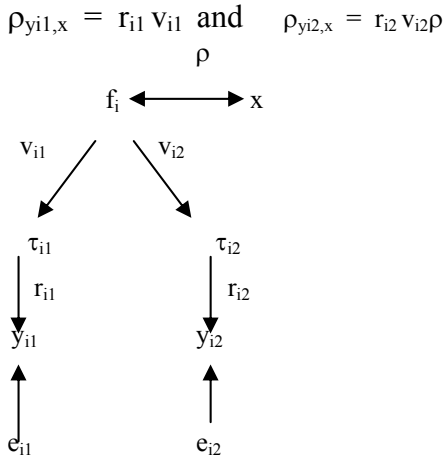


Figure 7: *A standard model to evaluate validity.*

This demonstrates that correlations can be different because of differences in validity, differences in reliability, or both. It also suggests that these correlations are not the proper criteria for evaluating the validity of measures. The validity of a measure should be evaluated by comparing the validity coefficients presented in the previous sections, in order to avoid confusion between reliability and validity, as occurs when the correlation is used with a criterion variable.

Conclusion and discussion

In this paper we hope we have shown that the Multitrait Multimethod approach to measurement problems in the social sciences can provide relevant information in terms of the reliability and validity of survey questions. In case of the use of the split ballot MTMM design, the approach can also provide information about the items missing values and bias, as well as other split ballot studies.

We argue that this approach is especially useful for subjective variables. It is often difficult to formulate alternative questions for objective variables, and to know whether memory effects can be excluded. The test-retest approach or the panel approach using the quasi simplex model is probably better for these variables.

We have also illustrated that relevant results can be obtained with the MTMM approach, suggesting that it is better to made use of item-specific scales than batteries of agree / disagree scales.

In Saris and Gallhofer (2007), we also presented the results of a meta-analysis of 87 MTMM experiments and a program (SQP) to predict the quality of survey questions. So far, this program can only predict the quality of questions in English, German and Dutch. Thanks to the experiments included in the ESS, it may be possible in the future to develop a new version of the SQP program that can predict the quality of questions in many other European languages.

The results of the MTMM experiments and the predictions of the program can be used to improve questions before the data are collected or for correction for measurement error after the data have been collected.

REFERENCES

Allison P. D. 1987. Estimation of linear models with incomplete data. In C. C. Clogg (ed.), *Sociological Methodology,* Washington DC: American Sociological Association, 71─103.

Althauser R. P., T. A. Heberlein, and R. A. Scott 1971. A causal assessment of validity: The augmented mulitrait-multimethod matrix. In H. M. Blalock Ir. (ed.), *Causal Models in the Social Sciences*. Chicago: Aldine, 151─169.

Alwin D. F. 1974. An analytic comparison of four approaches to the interpretation of relationships in the multitrait-multimehod matrix. In H. L. Costner (ed.), *Sociological Methodology*, San Francisco: Iossey Bass, 79─105.

Alwin D. F., and I. A. Krosnick 1991. The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research*, 20, 139─181.

Alwin D.F. 2007. Margins of error: A study of reliability in survey measurement. Hoboken, Wiley.

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

Arminger G., and M. E. Sobel 1991. Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association***,** 85, 195─203.

Campbell, D. T., and D. W. Fiske 1959. Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychological Bulletin*, 56, 81─105.

Campbell D. T., and E. I. O'Connell 1967. Method factors in multitrait-multimethod matrices: multiplicative rather than additive? *Multivariate Behavioral Research*, 2, 409─426.

Coenders G., and W. E. Saris 1998. Relationship between a restricted correlated uniqueness model and a direct product model for multitrait-multimethod data. In A. Ferligoi (ed.), *Advances in Methodology, Data Analysis and Statistic., Metodološki Zvezki* 14. Liubliana: FDV, 151─172.

Coenders, G., W. E. Saris, I. M. Batista-Foguet, and A. Andreenkova 1999. Stability of three-wave simplex estimates of reliability. *Structural Equation Modeling*, 6, 135─157.

Coenders, G., and W. E. Saris 2000. Testing nested additive, multiplicative and general multitrait-multimethod models. *Structural Equation Modeling* , 7, 219─250.

Corten I., W. E. Saris, G. Coenders, W. van der Veld, C. Albers, and C. Cornelis 2002. The fit of different models for multitrait-multimethod experiments. *Structural Equation Modeling*, 9, 213─232 .

Cudeck, R. 1988. Multiplicative models and MTMM matrices. *Iournal of Educational Statistics,* 13, 131─147.

Eid M. 2000. Multitrait-multimethod model with minimal assumptions. *Psychometrika***,** 65**,** 241─261.

European Social Survey 2002. *European Social Survey Round 1: Report of the First Round*

Forsman G. And I.Schreiner (1991) The design and analysis of Reinterview: An Overview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley,279-303

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

Heise D. R. 1969. Separating reliability and stability in test-retest-correlation. *American Sociological Review*, 34, 93─101.

Heise D. R., and G. W. Bohrnstedt 1970. Validity, invalidity and reliability. *Sociological Methodology*, 2, 104─129.

Iöreskog K. G. 1971. Simultaneous factor analysis in several populations, *Psychometrika* 34, 409─426.

Iöreskog K. G., and D. Sörbom 1989). *LISREL 7. A Guide to the Program and Applications*. Chicago: SPSS Inc.

Kenny D. A. 1976. An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology,* 12, 247─252.

Kenny D. A., and D. A. Kashy 1992. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112 , 165─172.

Marsh H. W. 1989. Confirmatory factor analysis of multitrait-multimethod data: many problems and few solutions. *Applied Psychological Measurement* , 13, 335─361.

Marsh, H. W., and L. Bailey 1991. Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement,* 15, 47─70.

Molenaar. N. I. 1986. *Formuleringseffecten in Survey-Interviews*. *PhD thesis*, Amsterdam: Free University.

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

Payne S. 1951. *The Art of Survey Questions*. Princeton: Princeton University Press.

Saris W. E. 1990. The choice of a model for evaluation of measurement instruments. In W. E. Saris, and A. van Meurs (eds.), *Evaluation of Measurement Instruments by Meta-analysis of Multitrait-Multimethod studies*, Amsterdam: North Holland, 118─133.

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 explanantions for correlated errors in MTMM studies. *Structural Equation Modeling* , 10, 193─214.

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

Saris W. E., and I. N. Gallhofer 2007. Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods,* 1, 31─46.

Saris W.E. and I.N. Gallhofer 2007 design, evaluation and analysis of questionnaires for Survey research, Hoboken, Wiley.

Saris W.E., M.Revilla, I. Krosnick and H.Schaefer (forthcoming) *Comparing questions with agree/disagree response options to questions with item-specific response options*

Satorra A. 1992. Asymptotic robust inferences in the analysis of mean and covariance structures. In P. V. Marsden (ed.), *Sociological Methodology 1992*. Oxford: Basil Blackwell, 249─278.

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

Scherpenzeel A. C., and W. E. Saris 1997. The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods and Research*. 25, 341─383.

Scherpenzeel A. C., and W. E. Saris 2006. Multitrait-Multimethod models for longitudinal research. In K.van Montford, H. Oud and A. Satorra (eds.), *Longitudinal Models in Behavioral and Related Sciences*, London: Lawrence Erlbaum, 381─403.

Schuman H., and S. Presser 1981. *Questions and Answers in Attitude Survey:*

Experiments on Question Form, Wording and Context. New York: Academic Press.

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.

Van der Veld W. 2006. *The Survey Response Dissected: A New Theory about the Survey Response Process*. *PhD thesis*, University of Amsterdam.

Wiley D. E., and I. A. Wiley 1970. The estimation of measurement error in panel data. *American Sociological Review*, 35, 112─117.

Wothke W. 1996. Models for multitrait-multimethod matrix analysis. In G. C. Marcoulides, and R. E. Schumacker (eds.), *Advanced Structural Equation Modeling: Issues and Techniques*, Mahwah NI: L. Erlbaum, 7─56.

Table 3: The quality of the different scales for three different questions in each country

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

Table 4 The mean quality of the three questions of experiment 2 in Round 3 of the ESS across 23 countries for the different methods (standard deviations in brackets)

Method	Q1	Q2	Q3
IS(11)	.76	.81	.76
	(.07)	(.05)	(.08)
A/D(5)	.44	.45	.47
	(.11)	(.11)	(.12)
A/D(11)	21	.28	.32
	(.13)	(.15)	(.13)
A/D(7)	.27	.31	.35
	(.11)	(.12)	(.13)