

Discussion

Evaluation Procedures for Survey Questions

*Willem E. Saris*¹

In this article, different criteria for the choice of an evaluation procedure for survey questions are discussed. Firstly, we mention a practical criterion: the amount of data collection the procedures require. Secondly, we suggest the distinction between personal judgments and model-based evaluations of questions. Thirdly, we suggest that it would be attractive if the procedure could evaluate the following aspects of the questions: 1. The relationship between the concept to be measured and the question specified; 2. The effects of the form of the question on the quality of the question with respect to: a. the complexity of the formulation, b. the precision, c. possible method effects, d. many other characteristics; 3. The social desirability of some of the response categories. Besides that, it would be desirable if the procedure could indicate the effect of respondents lack of the knowledge about the topic on their answers. We compare 13 procedures for the evaluation of questions with respect to these criteria and will derive some conclusions from this overview.

1. Introduction

In their article, Yan, Kreuter and Tourangeau mention a number of papers which compare the results of different evaluation procedures for survey questions: Fowler and Roman (1992), Presser and Blair (1994), Willis, Schechter and Whitaker (2000), Rothgeb, Willis and Forsyth (2001, 2004), DeMaio and Landreth (2004), and Jansen and Hak (2005). In these papers, the following evaluation procedures are mentioned: expert panels, focus groups, cognitive interviews, behavioral coding, three-step procedure of Jansen and Hak, standard pretests with debriefing, Quaid, SQP, latent variable models like test-retest, factor analysis and LCA, quasi-simplex design and model, MTMM design and model.

We would like to add to this list “the three step procedure” developed by Saris and Gallhofer (2007), “scaling procedures” developed by many people (see, for example Torgerson 1958), and item response theory (see, for example Hambleton et al. 1991).

We are not aware of papers discussing the criteria that could be used to select procedures for the evaluation of survey questions. Therefore, in the following pages we would like to suggest such criteria.

The first criterion we would like to suggest is a practical one: what one has to do to be able to use the different procedures. In this context, we distinguish between approaches that can be used without any data collection, procedures which require a small data set and

¹ Universitat Pompeu Fabra, Research and Expertise Centre for Survey Methodology, Passeig Pujades, 1, 08003, Barcelona, Spain. Barcelona. Email: w.saris@telefonica.net

Acknowledgment: I am very grateful for the useful comments of my colleagues of RECSM on an earlier version of this article.

those that require a more or less complete survey. It is clear that this criterion will play a role in the choice of an evaluation procedure.

As a second criterion to choose between the different procedures for question evaluation, we would like to mention whether the procedure is based on personal judgments or on model-based evaluations. We think that this criterion should also play a role in the choice of procedure.

Finally, we would like to suggest as a criterion the possible aspects of questions that are evaluated by the different procedures. In this context, we think about the following aspects of the quality of questions: 1. The relationship between the concept to be measured and the question specified; 2. The effects of the form of the question on the quality of the question with respect to a. the complexity of the formulation, b. the precision, c. possible method effects, d. many other characteristics; 3. The social desirability of some of the response categories. Besides that, it would be desirable if the procedure could evaluate questions with respect to the fourth criterion: the effect of respondents lack of knowledge about the topic on their answers. The use of the last criterion will lead to the suggestion to use combinations of different procedures in the evaluation of questions, because they evaluate different quality aspects of questions.

First, we will classify the different procedures with respect to the first two criteria. Thereafter, we will discuss what quality aspects the different procedures evaluate, and finally, we will describe which quality criteria can be evaluated with the different evaluation procedures. Based on this overview, we will finally draw some conclusions.

2. Two Basic Characteristics of Evaluation Procedures

In Table 1 we have classified the different procedures with respect to the amount of data needed for the evaluation (practical) and the evaluation procedures used.

It is, of course, very attractive if no new data have to be collected for the evaluation of the questionnaire. By new data we mean that one has to collect responses for the questions one would like to evaluate. There are a few procedures which satisfy this criterion. That does not mean that no new information is collected. In some cases, one has to ask experts

Table 1. The classification of 13 question evaluation procedures with respect to two procedural characteristics

Practical criterion	Evaluation procedure	
	Personal judgment	Model based
For quality prediction		
Without new data	Expert panels Focus groups Three step procedure (Saris and Gallhofer)	Quaid SQP Scaling methods
With few new data	Cognitive interviews Behavioral coding Tree step procedure (Jansen and Hak)	Scaling methods Behavioral coding
With a large pilot or full study	Debriefing of pilots	Latent variable models Quasi-simplex design/model MTMM design/model

about their judgments. In other cases, one has to code characteristics of the questions to obtain information about the quality of the questions.

There are also procedures which do not need a full study of the questionnaire, only a limited data collection for the evaluation of the questions. This is typically the case for cognitive interviewing using the think-aloud procedure, behavior coding, or some scaling procedures.

Finally, there are procedures that require a rather large data collection, such as most model-based procedures mentioned in Table 1, but also the standard procedure of debriefing interviewers after a pilot study.

It will be clear that, in principle, approaches that do not require new data are more attractive than procedures which require a new data collection before the official fieldwork. However, it should also be clear that this cannot be the only criterion.

Another very attractive criterion is whether the procedure is based on personal judgments of experts, interviewers, or respondents, or on model-based evidence collected in a special study or collected in the past. All procedures presented in the left column of Table 1 are based in some way or another on personal judgment, while the procedures on the right are model-based, collected on the spot, or evidence built up in the past. The scaling methods can be based on prior empirical studies or new empirical studies.

The model-based procedures will be more reliable if studies are well done. The results of such studies will not depend on the judgment of the researcher, and so repetition of applications of such studies will lead to approximately the same results. This is not necessarily the case when the procedure is based on personal judgments. With the change of the judges one may get different results. This is, for example, one of the problems that is mentioned in the study of Yan et al.

Combining the two criteria, one would say that the procedures on the top right side seem very attractive because they do not need the collection of new data and are based on existing evidence. This conclusion, however, would be overly hasty because the attraction of the procedures also depends on what aspects of the quality of questions are evaluated by the approach. This issue will, therefore, be discussed in the next section.

3. The Quality Aspects Evaluated by the Different Procedures

In our opinion it would be attractive if the evaluation procedures could evaluate the following aspects of the questions: 1. The relationship between the concept to be measured and the question specified; 2. The effects of the form of the question on the quality of the question with respect to: a. the complexity of the formulation, b. the precision, c. possible method effects, d. many other characteristics; 3. The social desirability of some of the response categories; 4. The lack of knowledge about the issue.

3.1. The Relationship Between the Concept to Be Measured and the Question Specified

Although the issue of validity of questions has been mentioned in all methodology books, one of the most ignored issues in survey research is the relationship between the concept to be measured and the questions specified. In this context, Blalock (1968) and others make a distinction between concepts by postulation and concepts by intuition. For concepts by intuition, questions can be formulated for which it is obvious that they measure the concept

of interest. For example, there is no doubt that the question “How satisfied are you with your job?” measures “Job satisfaction”. However one can also measure job satisfaction by asking about the satisfaction with different aspects of a job like the salary, social contact, spare time etc. In that case, the concept “Job satisfaction” becomes a concept by postulation, because we define the concept by a combination of the satisfaction with respect to the different aspects of the job. Here, the concept by postulation is defined by the combination of different concepts by intuition.

In the case of a concept by postulation, one has to evaluate the quality of the measurement of the concept on the basis of the relationship between the indicators for the concepts by intuition and the quality of the questions for these indicators. In the case of a concept by intuition, the evaluation of the question is much simpler, because one only has to evaluate an obvious question for the concept. Nevertheless, even this simple task is often not performed well. One can very easily provide many examples of cases where people specify what they want to mention, but specify questions which do measure something different. Two examples from research follow here.

In our first example, the researchers suggested measuring the opinion about the “policy of income equality”. In order to measure this concept, the same researchers suggested using the question:

“To what extent do you agree with the statement: The government should take care that people get a job?”

This question does not measure income equality, but an opinion about a “policy concerning full employment”.

The second example comes from another study where the idea is to measure the concept “interest in work”. In that study, the researchers suggest asking:

“How frequently did you think last month that you are interested in your work?”

In this question, it is assumed that people who are more interested in their work think more often that their work is interesting. That does not have to be true. Why don’t they ask directly “how interested are you in your work”?

The problem is that the relationship between the variable to be measured and the responses to the question can be very weak, because of the effect of other variables on the responses.

We think that it would be attractive if procedures for the evaluation of questions could detect such differences in the operationalization. The problem is, however, that often the researchers do not even specify what they want to measure, but immediately specify the questions. In that case evaluation is not possible.

3.2. The Effects of the Form of the Question on the Quality of the Question

Besides the validity of a question, one should consider the consequences of the form of the question for the quality of the measure. There are many alternatives for evaluating the same question. The most common aspect evaluated by survey researchers is whether the questions are too complicated for the respondents. Besides that, one has to consider the precision of the scale and the effect of the specific method chosen. There are,

however, many more aspects of the question which have consequences for quality, such as the presence of an introduction, labeling of the scale, the nonresponse option etc. Saris and Gallhofer (2007) distinguish more than 50 form characteristics of a single question. We cannot discuss them in detail. Here we will mention only the main factors starting with the complexity of the formulation.

a. The Complexity of the Formulation

The complexity of a question has to do with the unnecessary complexity of the formulation. Typical examples are: unnecessary linguistic complications such as superfluous lengthy words and sentences, or complex sentences using of subordinate clauses or complex grammatical forms. Such complexities, if not necessary, can cause confusion in the mind of the respondent and lead to uncertainty, which can cause random fluctuation in the answers.

b. Precision of the Measurement

With respect to precision, we have to make a distinction between measures for concepts by postulation operationalized using several indicators and measures for concepts by intuition which can be operationalized by a single question. In the former case, the quality depends indirectly on the quality of several questions, while the precision of a single question depends on the precision of the scale that is used, besides other characteristics. A large variety of scales is in use. Most common are 2-, 3-, 5-, 7- and 11-point scales. However, there are also procedures available using continuous scales, like magnitude estimation or line production or so-called visual analog scales.

c. The Effect of the Method Used

A lot of attention has been paid in psychological literature to the problem of “common method variance”. This CMV is a consequence of the fact that people may react in a specific way to a specific method consistently across questions. In that case, a correlation will occur between these variables. This correlation, caused by the reaction of the respondents to the method used, has no substantive meaning. In this context, the method can be the mode of data collection but it also can be a type of scale or another characteristic. If such a systematic effect exists, this may not only cause CMV but also invalidity in the responses, because the responses are not only affected by the opinion or attitude to be measured, but also by the reaction to the method used.

d. Other Form Characteristics

Besides these basic form characteristics, there are many other aspects of the form of a question which can have an effect. To mention some: presence of an introduction, or an instruction, or a show card, the labeling of the response alternatives, direction of the alternatives, etc. There are many specific studies that evaluate some of these characteristics (Schuman and Presser 1981, Andrews 1984, Scherpenzeel 1995, Tourangeau et al. 2000, Alwin 2007, Saris and Gallhofer 2007).

3.3. The Social Desirability of Some of the Response Categories

Social desirability also is a common concern of survey researchers. If respondents are affected in their choice of an answer category by the social desirability of the categories, this will lead to lack of validity because a different variable has an effect on the responses than the variable one would like to measure.

3.4. Lack of Knowledge of Respondents About the Topic

In many cases, questions are asked about topics which the respondent has never thought about. This means that the respondent creates an answer on the spot (Zaller 1992, Tourangeau et al. 2000). The respondent can do so on the basis of related information that is available in his/her mind. This automatic process will be based on the information which is most salient at that moment for the respondent. Therefore Zaller suggests that the responses of the same person can vary from one moment to the other. This expresses itself in a large random variation in the responses (see also Converse 1964).

4. Evaluation of the Different Procedures

In this section we want to describe the different procedures and the kind of results one can obtain with them.

4.1. Expert Panels

It is very common in survey research to ask colleagues to evaluate questions or even whole questionnaires. The researcher can ask the expert to give the evaluations without any structure, but he/she can also provide a formal appraisal system. In case of an evaluation without an appraisal system, the experts may make comments about the validity of the question, some form effects like complexity, the precision of the scale, and possible social desirability problems and knowledge problems, but they most likely will not give a detailed discussion of many possible characteristics of the questions and their consequences. In general, different people will provide comments on different aspects. This can be seen as an advantage of this procedure because in this way the information becomes more complete. On the other hand, one can also wonder about the significance of the remarks if some experts detect some problems while others do not see these problems.

The use of a formal appraisal system can avoid both problems, and one can get as detailed information as one would like. However, it is unlikely that an expert has sufficient knowledge of the consequences of the different choices to also give an evaluation of the effects on the quality of the question, let alone with respect to the effects of the combination of all these choices.

4.2. Focus Groups

In general, focus groups are used to determine how potential respondents interpret specific concepts which are used in a questionnaire. In this way one tries to check the validity of the questions for the concepts they want to measure. In focus groups, one can also detect that some questions are too complex or that the people have no knowledge of the topic in

question. What this procedure cannot provide is information about the positive or negative effects of specific choices with respect to the form of the questions.

4.3. *The Three-step Procedure of Saris and Gallhofer*

Saris and Gallhofer (2007) developed a procedure to design survey questions of which they claim that it guarantees that the question measures what the researcher wants to measure. So this procedure is completely directed at the validity of the measures.

The first step in the process is the decision whether the variable one wants to measure is a concept by intuition or a concept by postulation. If it is the former, one can immediately proceed to the next step. If it is a concept by postulation, one has first to define the concept in concepts by intuition. This is, of course, a theoretical step which can only be evaluated by the researcher and the research community.

The second step is the specification of a statement for the chosen concepts by intuition. For this step, Saris and Gallhofer have specified production rules. One first has to decide what the concept is that one wants to measure: an evaluation, a feeling, a norm, a policy, a preference, or another concept, and what the object is. Having done so, the production rules can be used to generate assertions for the concept of interest. These production rules are based on linguistic knowledge (Koning and van der Voort 1997, Harris 1978, Givon 1984, Weber 1993, Graesser et al. 1994, Huddleston 1994, Ginzburg 1996, and Groenendijk and Stokhof 1997).

In the third step, the assertions can be transformed into requests for answers as they call it, because not all so-called questions in survey research are real questions. One can also use imperatives or assertions. Characteristic of all forms is that they require an answer.

The guarantee of validity in this approach comes from the procedures developed for steps two and three. Step one is a theoretical step.

While this three-step procedure is a production system, one can also use it to evaluate the quality of questions by comparing the existing question with the results expected when the three-step procedure was used, or by looking to see if the question specified has the characteristics that were expected for the concept of interest.

The limitation of this procedure is that it concentrates completely on the validity of the measures and no other aspect. So for more complete evaluations of questions, this procedure has to be combined with other methods.

4.4. *Cognitive Interviews*

The most common procedure of cognitive interviewing is that one asks potential respondents to think aloud while answering the questions. An alternative is that one asks the respondent to tell how he/she came to his/her answer after the answer was given. Whatever procedure is chosen, this procedure aims at detecting whether the respondent interprets the concepts in the question in the correct way, and therefore this procedure aims again at the evaluation of the validity of the questions. However, like in the focus group approach, one can also see whether the respondents have the knowledge to answer the question or whether the question is formulated in too difficult a manner. Furthermore, in this case one will not get much information about the form effects.

4.5. *Behavioral Coding*

Behavioral coding is another way to achieve the same information. In this case, the communication between the respondent and the interviewer is recorded and later checked for indications of misunderstandings by the respondent to a question, which should show themselves in discussion with the interviewer about the meaning of the question. This procedure can also be used to detect wrong behavior of the interviewer, but that is less relevant here.

4.6. *Three-step Procedure of Jansen and Hak*

This is a combination of different forms of cognitive interviewing, starting with a think-aloud step, followed by probing to clarify the understanding of the process and later a normal debriefing. Given that the basis is cognitive interviewing, we expect that this procedure also mainly provides information about the validity of questions and possibly also about lack of knowledge and the complexity of the formulation.

4.7. *Standard Pretests With Debriefing*

In large and important surveys, it is rather common to pretest the questionnaire before the official data collection in order to check whether there are any problems. The check on problems is mostly done by asking the interviewers about the problems they have encountered while interviewing. Because the interviewer is mainly concerned about the communication with the respondent, the information one gets from the interviewers is similar to that obtained by behavioral coding, i.e., the misunderstandings about the meaning of questions, complexity of the questions, and lack of knowledge about the issue at stake.

4.8. *Quaid*

Quaid is a computer program that can analyze questions with respect to several aspects of questions namely: unfamiliar technical terms, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax and working memory overload. These judgments are based on long term research with respect to readability of texts (Graesser et al. 1994, Graesser et al. 2000a, Graesser et al. 2000b). Most of these checks are directed at problems of the form of the question, especially, at the complexity of the question and answer formulation with exception of the checks on vague or ambiguous noun phrases and vague or imprecise relative terms which are directed at the precision of the formulation. The attraction of the program is that one has to introduce the text of the questions and after a limited time one gets the results of the analysis. A disadvantage is that the program can only analyze questions in English and that the number of checks are limited. Suggestions for extension of the program are made for example by Faaß et al. (2008).

4.9. *Latent Variable Models Like Test-retest, Factor Analysis and LCA*

All latent variable models evaluate the quality of different questions for measurement of a latent variable. The quality of the question is based on the strength of the relationship

between this latent variable and the observed variable. The difference between the models arises from the type of data, continuous or discrete, and the assumptions made about the latent variables and the relationship between the observed and the latent variable. The latent variable is a variable which all observed variables have in common. Whether this variable is what the researcher was supposed to measure cannot be determined by this method. So the validity is difficult to determine. If the observed variables measure the same variable, the models can evaluate which form of the question provides more information about the concept measured by the latent variable. If the observed variables contain unique components, the latent variable is a concept by postulation defined by different observed concepts by intuition. In that case the strength of the relationship between the observed variables and the latent variable is a combination of the quality of the question and the strength of the relationship between the concept by intuition and the concept by postulation.

Given this description of these evaluation procedures it follows that these procedures mainly provide information about the effect of the form of the question, because these approaches cannot provide information about the validity of the measure nor about the social desirability or lack of knowledge about the topic.

A limitation of these procedures is that for each set of questions a separate study has to be done. This means that the results cannot be generalized across topics.

Another limitation is that these methods are difficult to apply as well on background variables. This design requires variations of the question for the same concept. These variations are rather difficult for background variables and simple behavioral questions. Therefore these questions should be evaluated in a different way as mentioned below (quasi simplex models).

An extra limitation is that these procedures are normally applied in such a way that method variance cannot be estimated. To detect method effects, one needs a special design: the MTMM design.

4.10. Quasi-simplex Design and Model

A procedure that can be used for evaluation of background variables and simple behavioral questions is the quasi-simplex design and model. In this design, the same question is repeated at least three times in a panel study. If these data are available, the so-called quasi-simplex model, allowing for change through time and measurement error at each point in time, can be used to estimate the quality of the question. This model has been used intensively by Alwin (2007) to evaluate many different questions. The quality of the question is in this case the explained variance in the observed variable by the latent variable. In Alwin (2007), valuable information about the quality of many questions tested in this way can be found.

Given the form of these experiments, we would say that this approach provides information about the quality of the form of the specific question. The procedure does not provide information about validity, the social desirability of some categories, or lack of knowledge.

The limitation of this approach is that its application to more subjective variables leads to problems for two reasons. The first is the assumption that the latent variable may change

but only with a lag of one time point. This means that an opinion that plays a role in the first moment, not in the second moment but again in the third moment cannot be specified in this model. This leads to identification problems (Coenders et al. 1999). The second problem is that all random changes in the latent variables are included in the error term. That means, for example, that in a measure of happiness the mood of a person, which is part of the happiness, will be included in the error and not in the latent variable. This characteristic of the model leads to serious problems with respect to the estimation of the quality of the questions, as was discussed by van der Veld (2006).

Another limitation of this approach is that method effects are ignored, while in general the same method is used at all points in time. The model does not allow the estimation of this effect. For background variables that may not be a serious problem, but for opinion questions it may cause a problem.

4.11. MTMM Design and Model

The multitrait multimethod (MTMM) design for evaluation of measurement instruments requires that for at least three different latent variables, at least three different forms that are however the same across latent variables are presented to the respondents (Campbell and Fiske 1959). On the basis of this design, a correlation matrix of nine variables is obtained. Different MTMM models have been developed for this matrix, which are special cases of latent variable models. Corten et al. (2002) and Saris and Aalberts (2003) showed that the classical MTMM model (Andrews 1984) and the equivalent true score model (Saris and Andrews 1991) fit the best to these matrices. This approach allows the estimation of reliability (the complement of random error variance) and internal validity (the complement of method variance). For details of this approach and for experiments to evaluate single questions, we refer to Saris and Gallhofer (2007). For evaluation of measures of concepts by postulation, we refer to Cote and Buckley (1987) and Lance et al. (2010).

The major advantage compared with the latent variable models discussed above is that with this design, besides the quality of the questions, the common method variance can also be estimated due to the use of the same method across questions. This is relevant because in survey research, batteries with the same form of questions are frequently used.

This approach provides estimates of the quality related with the different form of questions for the same latent variables. This procedure cannot say whether the specific latent variable is a good indicator for the concept of interest. Neither can social desirability and lack of knowledge be evaluated in this manner.

A limitation of this approach is that only a limited set of alternative forms for a specific latent variable are evaluated. The obtained results cannot be generalized. If meta-analyses across the existing MTMM experiments are conducted, a more general picture will arise. This was the basis for the SQP approach.

Another limitation is that the models used presently are based on the assumption of continuous observed variables. Whether this is a serious problem has yet to be studied in more detail. Some results suggest that it is not so serious an issue (Coenders et al. 1997). Only a start has been made with MTMM models for categorical variables (Oberski 2011).

This design has also problems with background variables and simple behavioral questions, because variations of these questions are difficult to formulate and to study.

4.12. Survey Quality Prediction: SQP

The computer program SQP 2.0 has been developed to generate predictions of the quality of questions, based at the moment on a data set of 4000 questions which have been involved in MTMM experiments. The quality is defined as the product of the reliability and validity of a question. The reliability and validity of a question are estimated in MTMM experiments. The program SQP 2.0 provides these estimates for all questions which have been involved in an MTMM experiment. But the program does more. Based on coding of the question characteristics of these 4,000 questions, a prediction procedure has been developed for the quality of the questions. The prediction of the quality of these 4,000 questions is rather good (close to .9), therefore, the program also offers the possibility to use this prediction procedure for predicting the quality of new questions. In order to do so, the user has to code the characteristics of the question, including some research characteristics, and the program then generates the prediction. It also provides suggestions for the improvement of the question, if necessary. For details of the procedure we refer to Saris and Gallhofer (2007) and a more recent publication by Saris et al. 2012.

Given that the predictions are based on coding of around 50 question characteristics and some research design characteristics, quality evaluation is mainly directed at the effects of the form of the questions, although the domain and the concept of the question and the social desirability and knowledge of the respondents of the issue are also taken into account in the prediction. An attractive feature is that form characteristics can be coded in all languages, and so the program can make predictions of the quality of questions in all languages that have been involved in the MTMM experiments, which are more than 20.

A limitation of the program is that it is concentrated on the form of single questions, keeping the concept by intuition the same. Whether this concept by intuition is a good indicator for the concept the respondent wants to measure is outside the scope of this program. So the validity coefficient predicted is the validity for a concept by intuition. The quality can be defined as the explained variance in the observed responses by the concept by intuition studied.

A second limitation of the program SQP is that it is based on MTMM experiments. These experiments are rather difficult for background variables and simple behavioral questions, as was mentioned above. So SQP cannot predict the quality of these questions.

4.13. Scaling Procedures

Most scaling procedures analyze the data of several questions simultaneously to test an expected structure between them. Typical examples are the Thurstone scale, Likert scale, etc. (Torgerson 1958), Rasch scale and item response theory (Hambleton et al. 1991), Guttman scale, Mokken scale and the unfolding scale (van Schuur 1997), to mention some of them. These scales are based on different models, but all aim at ultimately deriving a score for a respondent on one or perhaps more scales. So these procedures claim to determine a score for the respondents on a scale for the variable of interest. However, the scaling procedure itself cannot guarantee that the score obtained really represents the variable of interest. In fact, like all model based methods mentioned, the procedure can only provide an estimate of the quality of the obtained score for whatever the latent variable may be.

The limitation of these approaches is therefore that they provide only an estimate of the quality of the observed scores, but not of the validity, the social desirability or the lack of knowledge of the respondents with respect to the issue.

Besides this, no attention is paid in these procedures to the problem of common method variance.

5. Conclusions

Looking at the given criteria, some obvious results can be observed:

1. All procedures based on personal judgments provide information about the validity, social desirability, and knowledge of the respondents about the issue of the question and much less about the effects of the form of the questions.
2. The model-based procedures provide rather precise information about the effect of the form of the question on the quality, and the quality can even be expressed in a number between 0 and 1. However, these procedures cannot provide information about the validity of the question for the concept of interest.
3. It is quite obvious that it makes no sense to start with the evaluation of the form of a question before the validity of the measure for a concept has been determined. This means that the personal judgment procedures, at the left side of Table 1, should play an important role in the first phase of questionnaire design.

Based on our experience with questionnaire design, we have decided to spend extra time on the development of a procedure that can guarantee with more certainty that researchers measure what they are supposed to measure. This has become the three-step procedure of Saris and Gallhofer (2007). We are still convinced that this procedure requires more attention because it can prevent a lot of problems with respect to validity.

4. Evaluating the form of the questions, the model-based procedures, at the right side of Table 1, will be very helpful. In this context, a distinction should be made between evaluation procedures that can only evaluate single questions like SQP, the standard MTMM approach in survey research, and the quasi-simplex approach on the one side, and on the other side procedures that can evaluate measures for concepts by postulation like latent variable models and scaling procedures. In this respect the latter procedures have an advantage. However, they have also the disadvantage that they ignore method effects. In Saris and Gallhofer (2007, ch. 14) we have shown that this may lead to very different conclusions. In psychology, the MTMM approach has also been used for the evaluation of measures for concepts by postulation (Cote et al. 1987 and Lance et al. 2010).
5. There is a fundamental difference between the quality predictions of SQP, which are based on a multivariate prediction approach, and predictions of the quality of the empirical studies, such as latent variable models and also MTMM studies. In SQP, both results are available for all MTMM questions of the ESS. Most of the time the estimates are rather similar, but sometimes they are different. This can occur because the specific question is quite different from the other questions in the database, or in the study of this specific question something was different from normal. This is something one has to decide when looking at these results.

6. The procedures that do not need new data are obviously more attractive than procedures which require new data. On the personal judgment side, it would mean that asking experts for comments is a very attractive procedure before one starts to collect data. On the model-based side, Quaid and SQP seem to be attractive approaches to use before data collection.

6. References

- 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.
- Blalock, H.M. Jr, (1968). The Measurement Problem: A Gap Between Languages of Theory and Research. In *Methodology in the Social Sciences*, H.M. Blalock and A.B. Blalock (eds). London: Sage, 5–27.
- Campbell, D.T. and Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrices. *Psychological Bulletin*, 56, 81–105.
- Coenders, G., Saris, W.E., Batista-Foguet, J.M., and Andreenkova, A. (1999). Stability of Three-Wave Simplex Estimates of Reliability. *Structural Equation Modeling*, 6, 135–157.
- Coenders, A.S. and Saris, W.E. (1997). Alternative Approaches to Structural Modeling of Ordinal Data: A Monte Carlo Study. *Structural Equation Modeling*, 4, 261–282.
- Converse, P. (1964). The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, D.A. Apter (ed.). New York: Free Press, 206–261.
- Corten, I., Saris, W.E., Coenders, G., van der Veld, W., Albers, C., and Cornelis, C. (2002). The Fit of Different Models for Multitrait-Multimethod Experiments. *Structural Equation Modeling*, 9, 213–232.
- Cote, J.A. and Buckley, M.R. (1987). Estimating Trait, Method and Error Variance; Generalizing Across 70 Construct Validity Studies. *Journal of Marketing Research*, 11, 535–559.
- DeMaio, T. and Landreth, A. (2004). Cognitive Interviews: Do Different Methods Produce Different Results? In *Methods for Testing and Evaluating Survey Questionnaires*, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). Hoboken, NJ: John Wiley and Sons.
- Faaß, T., Kaczmirek, L., and Lenzner, A. (2008). Psycholinguistic Determinants of Question Difficulty: A Web Experiment. *Proceedings of the Seventh International Conference on Social Science Methodology (RC33)* [cd-rom], University of Naples “Federico II”, Italy.
- Forsyth, B., Rothgeb, J., and Willis, G. (2004). Does Question Pretesting Make a Difference? An Experimental Test. In *Methods for Testing and Evaluating Survey Questionnaires*, Presser et al. (eds). Hoboken, NJ: Wiley.
- Fowler, F.J. and Roman, A.M. (1992). *A Study of Approaches to Survey Question Evaluation*, Final Report for U.S. Bureau of the Census. Boston: Center for Survey Research.

- Ginzburg, J. (1996). Interrogatives: Questions, Facts and Dialogue. In *The Handbook of Contemporary Semantic Theory*, S. Lappin (ed.). Cambridge, MA: Blackwell, 385–421.
- Givon, T. (1984). *Syntax. A Functional-Typological Introduction Vol. I–II*. Amsterdam: J. Benjamin.
- Graesser, A.C., McMahan, C.L., and Johnson, B.K. (1994). Question Asking and Answering. In *Handbook of Psycholinguistics*, M. Gernsbacher (ed.). San Diego, CA: Academic Press, 517–538.
- Graesser, A.C., Wiemer-Hastings, P.K., Kreuz, R., and Wiemer-Hastings, P. (2000a). QUAID: A Questionnaire Evaluation Aid for Survey Methodologists. *Behavior Research Methods, Instruments, and Computers*, 32, 254–262.
- Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., and Kreuz, R. (2000b). The Gold Standard of Question Quality on Surveys: Experts, Computer Tools, Versus Statistical Indices. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. Washington, DC: American Statistical Association, 459–464.
- Groenendijk, J. and Stokhof, M. (1997). Questions. In *Handbook of Logic and Language*, J. van Benthem and A. ter Meulen (eds). Amsterdam: Elsevier, 1055–1124.
- Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. London: Sage.
- Harris, Z. (1978). The Interrogative in a Syntactic Framework. In *Questions*, H. Hiz (ed.). Dordrecht: Reidel, 37–89.
- Huddleston, R. (1994). The Contrast Between Interrogatives and Questions. *Journal of Linguistics*, 30, 411–439.
- Jansen, H. and Hak, T. (2005). The Productivity of the Three-Step Test-Interview (TSTI) Compared to an Expert Review of a Self-Administered Questionnaire on Alcohol Consumption. *Journal of Official Statistics*, 21, 103–120.
- Koning, P.L. and van der Voort, P.J. (1997). *Sentence Analysis*. Groningen: Wolters-Noordhoff.
- Lance, C.E., Dawson, B., Birkelbach, D., and Hoffman, B.J. (2010). Method Effects, Measurement Error and Substantive Conclusions. *Organizational Research Methods*, 13, 435–455.
- Oberski, D. (2011). *Latent Class Multitrait- Multimethod models*. In *Measurement error in comparative research*, D. Oberski (ed.). Unpublished PhD dissertation of the University of Tilburg.
- Presser, S. and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? *Sociological Methodology*, 24, 73–104.
- Rothgeb, J., Willis, G., and Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results. *Proceedings of the Section on Survey Methods*. Alexandria, VA: American Statistical Association.
- Saris, W.E. and Andrews, F.M. (1991). Evaluation of Measurement Instruments Using a Structural Modeling Approach. In *Measurement Errors in Surveys*, P.P. Biemer, R.M. Groves, L.E. Lyberg, N. Mathiowetz and S. Sudman (eds). New York: Wiley, 575–599.

- Saris, W.E. and Aalberts, C. (2003). Different Explanations for Correlated Errors in MTMM Studies. *Structural Equation Modeling*, 10, 193–214.
- Saris, W.E., Oberski, D., Revilla, M., Zavalla, D., Lilleoja, L., Gallhofer, I., and Grüner, T. (2012). Final Report About the Project JRA3 as Part of ESS Infrastructure. RECSM Working paper, 24.
- Saris, W.E. and Gallhofer, I.N. (2007). *Design, Evaluation and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: Wiley.
- Scherpenzeel, A.C. (1995). *A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies*. Leidschendam: KPN Research.
- Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. New York: Academic Press.
- Torgerson, W.S. (1958). *Theory and Methods of Scaling*. London: Wiley.
- Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, MA: Cambridge University Press.
- Weber E.G. (1993). *Varieties of Questions in English Conversations*. Amsterdam: J. Benjamins Publ. Co.
- Willis, G.B., Schechter, S., and Whitaker, K. (2000). A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What do They Tell Us? *Proceedings of the Section on Survey Research Methods*. American Statistical Association.
- van der Veld, W. (2006). Judging Different Models to Estimate Survey Question Quality. In *The Survey Response Dissected: A New Theory About the Survey Response Process*, W. van der Veld (ed.). Unpublished PhD dissertation of the University of Amsterdam, Chapter 5.
- van Schuur, W.H. (1997). Nonparametric IRT Models for Dominance and Proximity Data. In *Objective Measurement: Theory into Practice*, M. Wilson, G. Engelhard, Jr, and K. Draney (eds). Volume 4. Greenwich (Cn)/London: Ablex Publishing Corporation, 313–331.
- Zaller, J.R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.