Survey Research Methods (2010)
Vol.4, No.1, pp. 61-79
ISSN 1864-3361
http://www.surveymethods.org

Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options

Willem E. Saris
RECSM, Universitat Pompeu Fabra

Melanie Revilla
RECSM, Universitat Pompeu Fabra

Jon A. Krosnick
Stanford University

Eric M. Shaeffer
Ohio State University

Although Agree/Disagree (A/D) rating scales are hugely popular in the social sciences, a large body of research conducted during more than five decades has documented the bias that results from acquiescence in responses to these items. This may be a reason to prefer questions with Item Specific (IS) response options, but remarkably little research has explored whether responses to A/D rating scale questions are indeed of lower quality than responses to questions with IS response options. Using a research design that combines the advantages of a random assignment between-subjects experiment and the multitrait-multimethod approach in the context of representative sample surveys, we found that responses to A/D rating scale questions indeed had much lower quality than responses to comparable questions offering IS response options. These results attest to the superiority of questions with IS response options.
Keywords: Quality of Agree/Disagree scales, Item Specific scales, Split Ballot MTMM

Introduction

Throughout the 20^{th} Century, Agree/Disagree (A/D) questions have been and remain extremely popular in questionnaire-based research in the social sciences and psychology. For example, Rensis Likert's (1932) classic attitude measurement technique uses an A/D scale, and numerous batteries have been developed for attitude and personality measurement which do so as well (see, e.g., Shaw and Wright, 1967; Robinson and Shaver, 1973; Robinson, Shaver, and Wrightsman, 1991; Robinson and Wrightsman, 1999). Psychologists are not alone in their heavy reliance on this response format. In the National Election Study surveys (done by the University of Michigan's Center for Political Studies) and in the General Social Surveys (done by the University of Chicago's National Opinion Research Center), A/D response formats have been used in some of the most widely-studied questions, including measures of political efficacy and alienation, international isolationism, and much more (see Davis and Smith, 1996; Miller and Traugott, 1989). Leading journals in many social science fields report frequent use of these sorts of items in contemporary research projects.

One reason for the popularity of A/D response alternatives is that they seem to offer the opportunity to measure just about any construct relatively efficiently. Alternative question design approaches require that response alternatives be tailored to each item's particular construct.

Imagine that we are interested in the health of the respondents. Their health can be described for example as excellent, very good, good, fair or poor. So one could ask using the A/D form of the question:

> *To what extent do you agree strongly or disagree strongly that your health is excellent?*
>
> 1. *agree completely,*
> 2. *agree somewhat,*
> 3. *neither agree nor disagree,*
> 4. *disagree somewhat, or*
> 5. *disagree completely*

Fowler (1995) says, discussing this type of question: "However, how much simpler, direct and informative it is to ask":

> *"How would you rate your health – excellent, very good, good, fair, or bad?"*

We will call this type of question Item Specific (IS) because the categories used to express the opinion are exactly those answers we would like to obtain for this item. It seems that the IS type of scale is a much more direct way to collect an opinion from individuals than the one using the A/D response scale. Even though these two approaches aim to measure the same thing, i.e., the judgment of the respondent about his/her health, the A/D approach seems much more indirect that the IS approach.

Nevertheless this type of A/D questions is very frequently used in survey research. The reason is that this A/D response format can be used for nearly any type of question so the questionnaire needs only to present the scale once, thereby saving time and streamlining questionnaire administration. For example, a questionnaire might ask the following series of three questions:

Contact information: Willem E. Saris, ESADE, e-mail: w.saris@telefonica.net

"Next, I'm going to read you a series of statements. For each one, please tell me whether you agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly.

1. *First, 'My overall health is excellent.'*
2. *Second, 'The issue of abortion is very important to me personally.'*
3. *Third, 'I rarely feel sad.'"*

This is a typical example of a battery of A/D questions for rather different topics. The attraction of this approach is that the same scale is used for each item, independently of the dimension that one is trying to measure.

If one would use the IS scale for the same items one could formulate it as follows:

1. *"How would you rate your health overall: excellent, very good, good, fair, bad or very bad?"*
2. *"How important is the issue of abortion to you personally? Is it extremely important, very important, somewhat important, not too important, or not important at all?"*
3. *"How often do you feel sad? Constantly, very often, somewhat often, rarely, or never?"*

It should be clear that in the battery of A/D questions the scales of these questions are not "Item Specific." Where the first item aims at evaluation of the health of the respondent, the response scale is not from "very good" to "very bad" but in that item one possible response is given and the people are asked to indicate to what extent they agree or disagree with this chosen option. For the other items the same procedure is used and therefore the same A/D scale can be used for all items.

However one should realize that the chosen statement is arbitrary. One could also have chosen:

To what extent do you agree or disagree that your health is very good?

or

To what extent do you agree or disagree that your health is good?

or

To what extent do you agree or disagree that your health is fair?

or

To what extent do you agree or disagree that your health is bad?

or

To what extent do you agree or disagree that your health is very bad?

In principle these questions represent alternative forms of A/D questions and they could be seen as equivalent. However the three questions in the middle have the problem that people could disagree because their health is better but also because their health is worse than indicated. This is a rather unattractive character of these questions (Fowler, 1995). In addition, it has been found that the positive and negative formulated questions are not each others complement (Schuman and Presser, 1981).

Although these two types of questions as formulated above seem to aim at measuring the same three opinions, it is not clear whether the questions are equally good. So our goal in this paper is to explore the quality of these different types of questions by comparing questions with A/D response options to questions with IS response options. We begin by offering a theory of the cognitive response process to A/D questions that suggests questions with IS response options may yield higher quality data. Then we test this hypothesis using data from several experiments in a large number of European countries using data of the European Social Survey.

Cognitive Response Processes and Their Consequences

The goal of A/D questions is usually to place respondents on a continuum. For example, a question stem saying "I am usually happy" is intended to gauge frequency of happiness: how often the respondent is happy, on a scale from "never" to "always". A question stem saying "I like hot dogs a lot" is intended to gauge quantity of liking/disliking: how much the respondent likes hot dogs, on a scale from "dislike a lot" to "like a lot". And a question stem saying, "Ronald Reagan was a superb President" is intended to gauge respondents' evaluations of Reagan's performance, on a scale ranging from "superb" to "awful".

Tourangeau, Rips and Rasinski (2000) identify four components in the process of answering questions, which are: "comprehension of the item, retrieval of relevant information, use of that information to make required judgments and selection and reporting of an answer". Focusing more on A/D items, other authors (see, e.g., Carpenter and Just, 1975; Clark and Clark, 1977; Trabasso, Rollins, and Shaughnessy, 1971) also state that answering such questions requires respondents to execute four cognitive steps, but they are a bit different. First, respondents must read the stem and understand its literal meaning. Then, they must look deeper into the question to discern the underlying dimension of interest to the researcher. This is presumably done by identifying the *variable quantity* in the question stem. In the first example above, the variable is identified by the word "usually" – it is the frequency of happiness. In the second example above, the variable is quantity, identified by the phrase "a lot". And in the third example, the variable is quality, identified by the word "superb". Having identified this dimension, respondents must then place themselves on the scale of interest.

For example, the stem, "I am usually happy", asks respondents first to decide how often they are happy. Then, they must translate this judgment onto the A/D response options appropriately, depending upon the valence of the stem. Obviously, it would be simpler to skip this latter step altogether and simply ask respondents directly for their judgments of how often they are happy. This was also noted many years ago by Fowler (1995) as we have indicated above.

Doing this has another benefit as well, in that it avoids a unique potential problem with A/D questions that we have not yet considered. Researchers often presume that if a question stem is worded "positively", as all three examples are above (indicating high frequency of happiness, liking of hot dogs, and a positive evaluation of Reagan's performance, respectively), then people who answer "agree" are indicating more happiness, liking, and positive evaluation, respectively, than people who answer "disagree". However, "disagree", "false", and "no", responses can be offered for various reasons, some of which violate the presumed monotonic relation between answers and respondent placement on the underlying scale of interest.

For example, consider a person who is asked whether he or she agrees or disagrees with the statement: "I am generally a happy person". A person who disagrees may believe (1) he or she is generally an unhappy person, (2) he or she is generally neither happy nor unhappy, and instead is usually affectless, (3) he or she is happy 55 percent of the time and unhappy 45 percent of the time, and 55 percent of the time is not frequent enough to merit the adjective "generally", or (4) he or she is *always* happy, and "generally" does not represent this universality adequately.

In fact, one tends to assume that individuals who are not at all happy will disagree strongly with the statement, individuals who are not happy will disagree, individuals who are neither happy nor unhappy will respond neither agree nor disagree, individuals who are happy will agree and individuals who are completely happy will agree strongly. But this is not the case. It is not at all clear where individuals should place themselves on the A/D scale if they are "not usually happy." They may disagree but where should they place themselves? To solve this problem they can use different solutions for different items which would lead to lower reliability. They can also use a similar solution for different items and if this solution is different for different individuals this will lead to a method effect and consequently to lower validity because the score is not only influenced by their opinion but also by a response pattern. For example, it is possible that some individuals interpret the "strongly" as a way to say that individuals are more at the extreme on the considered scale: they are always happy (or unhappy). However, as Fowler (1995) suggested, others could interpret the "strongly" as the intensity of their opinion: how sure he/she is about the fact that he/she is happy or unhappy? Thus, even an individual who is generally neither happy nor unhappy can end up not only expressing disagreement with the statement about happiness above, but also expressing it *strongly*. Offering "neither agree nor disagree" as a response option would not necessarily prevent this sort of problem, since an individual who is confident that he/she is generally neither happy nor unhappy might well be inclined to strongly disagree in this case. When this sort of mismatch of the response dimension to the latent construct of interest occurs, it will compromise the validity of responses.

If these arguments are true, then responses to questions with IS response options may contain less measurement error than A/D questions and probably also less method effects and therefore are more valid. But there is another reason why method effects are expected, namely because A/D scales are normally used for batteries, and in batteries of questions with the same A/D scale, acquiescence may occur.

Acquiescence response bias

A great deal of research points to a potential danger inherent in this response format: acquiescence response bias in A/D batteries. More than one hundred studies using a wide variety of methods have demonstrated that some respondents are inclined to agree with just about any assertion, regardless of its content (for a review, see Krosnick and Fabrigar, forthcoming). Three different theoretical accounts for this phenomenon have been proposed, and all of them enjoy some empirical support. The first argues that acquiescence results from a personality disposition some individuals have to be polite and to avoid social friction, leading them to be especially agreeable (Costa and McCrae, 1988; Goldberg, 1990; Leech, 1983). The second explanation argues that A/D formats unintentionally suggest to some respondents that interviewers and/or researchers believe the statements offered in such items, and some respondents who perceive themselves to be of lower social status than the interviewer or researcher may choose to defer to their apparent expertise and endorse their apparent beliefs (Carr, 1971; Lenski and Leggett, 1960; Richardson, Dohrenwend, and Klein, 1965). Finally, the theory of survey satisficing argues that a general bias in hypothesis testing toward confirmation rather than disconfirmation inclines some respondents who shortcut the response process toward agreeing with assertions presented to them in A/D questions (Krosnick, 1991).

When portrayed in this fashion, it might seem obvious that acquiescence would compromise the quality of data obtained. According to all of these explanations, respondents susceptible to acquiescence are inclined to answer an A/D question by saying "agree", regardless of whether that answer accurately represents their opinion or not. Therefore, regardless of whether a question stem says "I rarely feel sad" or "I often feel sad", these individuals would answer "agree", yet these "agree" answers cannot both be correct. If this question were to be presented instead with IS response options, perhaps these individuals would report their true opinions more accurately.

This behavior may not occur for all people and under all conditions. Possibly it occurs for people who lack an opinion on an issue and have an inclination to acquiesce. However if that is the case this will again lead to correlations related to the A/D response scale which may not occur for IS scales because there the items are normally measured with a dif-

ferent scale. So for A/D scales we expect, again because of acquiescence, more method effect and therefore less validity of the responses.

Some past studies have compared the reliability and validity of measurements made with A/D questions and questions with IS response options, but their results have been mixed. For example, Counte (1979) and Scherpenzeel and Saris (1997) found questions with IS response options to have greater reliability than A/D questions, but Berkowitz and Wolken (1964) found a slight trend in the opposite direction. Ray (1979), Ross, Steward, and Sinacore (1995), and Schuman and Presser (1981) reported findings suggesting that questions with IS response options had greater correlational validity than A/D questions, though Berkowitz and Wolken (1964), Counte (1979) and Ray (1980) found trends in the opposite direction. In light of this small body of evidence and the fact that many of the reported differences were not subjected to tests of statistical significance, it seems worthwhile to investigate this issue further, which is what the research reported here was designed to do.

The Present Investigation: Research Design

In the next sections we will report on a large number of experiments to compare the quality of a battery of items using a standard A/D scale or true/not true scale with the quality of IS scales. These experiments have been done as part of the European Social Survey (ESS). The ESS surveys are carried out in all European countries that are willing to participate following the rules[1] specified by the Central Coordinating Team (CCT) of the ESS. In this cross national study a major effort is made to draw samples from each country that are as comparable as possible, being all probability samples (Häder and Lynn, 2007). Strict rules are also specified for the translation of the questions (Harkness, 2007) and the fieldwork (Billiet, Koch and Philippens, 2007).

In addition, the ESS has the unique feature that experiments are added to the main questionnaire in order to evaluate the quality and comparability of the different questions in the different countries. The main questionnaire is always completed in face-to-face interviews in the homes of the respondents using show cards. The supplementary questionnaires, given randomly to the different split-ballot groups, are also administered in a face-to-face mode in most countries, though in some they are self-administered. In several experiments a comparison has been made of the A/D scales with IS scales. We will concentrate on these experiments.

All studies reported here used a relatively new data collection design and analytic method. The two most commonly used approaches for the evaluation of the quality of questions in survey research are the split-ballot experiment (e.g., Billiet, Loosveldt, and Waterplas, 1985; Schuman and Presser, 1981) and the cross-sectional multitrait-multimethod (or MTMM) approach (Andrews, 1984; Saris and Andrews, 1991; Saris and Münnich, 1995; Scherpenzeel and Saris, 1997). Both of these approaches have advantages and disadvantages, and the approach we employed combines the two to yield a stronger technique than either approach alone.

In split-ballot experiments, respondents are randomly assigned to be asked different versions of the same question. For example, in the first experiment described below the respondents of two randomly selected groups were asked to indicate their opinions on an issue specified in two alternative statements about the frequency of an event using an A/D type of question.

> *Please indicate how much you agree or disagree with the statement: Before doctors decide on a treatment, they usually discuss it with their patient.*
>
> - *agree strongly*
> - *agree*
> - *neither disagree not agree*
> - *disagree*
> - *disagree strongly*

And the alternative form used in the other subgroup was:

> *Please indicate how much you agree or disagree with the statement: Before doctors decide on a treatment, they rarely discuss it with their patient*
> With the same answer categories

If no acquiescence occurs the proportion of respondents in the first group agreeing with the first assertion should be the same as the proportion of respondents in the second group disagreeing with the second assertion. But if acquiescence does occur, the proportion of people agreeing with the first assertion will exceed the proportion of people disagreeing with the second assertion.

In the classical multitrait-multimethod (MTMM) studies each respondent is asked at least three questions measuring each of various opinions (also called "traits") using at least three different methods, leading to a correlation matrix with nine observed variables.

In questionnaire studies, a method is an item format characteristic. Here we are especially interested in the difference in quality of A/D and IS type questions about the same topic. So, for instance, next to the two forms of an item mentioned above asked using the A/D format, a third form can be specified with the IS format:

> *Please indicate how often you think the following applies to doctors in general: Before doctors decide on a treatment, they discuss it with their patient*
>
> - *never or almost never*
> - *some of the time*
> - *about half of the time*
> - *most of the time*

[1] For more information we refer to the ESS website: http://www.europeansocialsurvey.org/

- *always or almost always*

Ideally, methods are completely crossed with traits, meaning that every opinion is measured by every method. With data collected via this sort of design, it is possible to employ structural equation modeling techniques to estimate the reliability and validity of items in each format, as well as the amount of correlated method-induced error variance for items in each format (see Alwin, 1974; Andrews, 1984; Browne, 1984; Coenders and Saris, 2000; Marsh and Bailey, 1991; Saris and Andrews, 1991, Saris and Gallhofer 2007). Consequently, a researcher can compare the quality of data collected by various methods.

A potential drawback to this approach is the fact that each respondent must be asked multiple questions assessing the same opinion, and early questions might influence answers to later questions, thus distorting their apparent quality. For example, having just reported my opinion on abortion on a 7+point rating scale, I may use my answer to that question as a basis for deriving a report of that same attitude on a 101-point scale later. The precise meanings of the 101 scale points may not be especially clear, but having reported my opinion on a much clearer 7 point scale first, I may be able to simply translate that report onto the 101-point-scale, thereby bypassing the need to interpret all the scale points. A response of 6 on the 7 point scale, for instance, corresponds proportionally to a response of about 80 on the 101-point scale. Therefore, if respondents are first asked a question involving a format that is easy to use and yields relatively error-free reports, this may help respondents to provide apparently reliable and valid reports of the same opinions on more difficult-to-use rating scales later. As a result, this approach may under-estimate differences between question formats in terms of reliability and validity.

In order to minimize the likelihood that an initial question will contaminate answers to later questions measuring the same construct, it is desirable to maximize the time period between administering the two questions. Work by Van Meurs and Saris (1990) suggests that at least 20 minutes are required between the administrations of related items in order to eliminate a respondent's recollection of his or her first answer when answering the second question. And an impressive set of laboratory studies show that people cannot remember attitude reports they made just one hour previous (Aderman and Brehm, 1976; Bem and McConnell, 1970; Goethals and Reckman, 1973; Ross and Shulman, 1973; Shaffer, 1975a, 1975a; Wixon and Laird, 1976). In the ESS studies the two repetitions of the same question were separated by approximately an hour of other survey questions.

In order to avoid problems in estimation of the parameters of an ordinary MTMM structural equation model, it is necessary to have at least three measures for each construct (Saris, 1990). This means that in one survey all respondents have to reply to the same question using a different method three times. In order to avoid this problem Saris, Satorra and Coenders (2004) have developed the so called Split-Ballot MTMM (or SB-MTMM) experiment where each respondent has to answer all questions only twice. If all questions have been evaluated by one method by all respondents while the other two methods are only used in two randomized subgroups of the original sample, all reliability, validity coefficients and method effects of this model can be estimated. In order to indicate what information will be obtained from these experiments we will discuss the MTMM model that will be used in these studies.

The Quality criteria: Reliability, Validity, and Quality

In MTMM experiments commonly conducted to estimate reliability and validity (internal validity), a minimum of three traits and three methods are used, yielding a correlation matrix for nine variables. In the current split-ballot MTMM design, we measured three traits (opinions, in this case), but in each sample, only two methods were implemented (see the correlation matrices in the LISREL input displayed in the Appendix). Each zero correlation occurred because the two questions were not asked of the same respondents.

To analyze conventional MTMM correlation matrices, various analytic models have been suggested (Alwin, 1974; Andrews, 1984; Browne, 1984; Marsh and Bailey, 1991; Saris and Andrews, 1991; Coenders and Saris, 2000; Eid, 2000; Saris and Aalberts, 2003). Corten et al. (2002) showed, analyzing many data sets, that the additive model of Saris and Andrews (1991) should be preferred above the multiplicative model originally suggested by Browne (1984) and reformulated by Coenders and Saris (2000). Saris and Aalberts (2003) compared several alternative explanations for method specific correlations and concluded that the MTMM model of Saris and Andrews gave the best explanation. Therefore we employed their model in this paper (see Figure 1), which has the further advantage of making an explicit distinction between reliability and validity:

$$Y_{ij} = h_{ij}T_{ij} + e_{ij} \quad (1)$$

$$T_{ij} = v_{ij}F_j + m_{ij}M_i \quad (2)$$

Where Y_{ij} is the observed variable for the jth trait (attitude or belief in this case) and the ith method, T_{ij} is the systematic component of the response Y_{ij}, F_j is the jth trait, and M_i represents the variation in scores due to the ith method. This model posits that the observed variable is the sum of the systematic component plus random error. And the systematic component of a measure is the sum of the trait and the effect of the method used to assess it.

As usual, it is assumed that the random errors are uncorrelated with each other and with the independent variables in the different equations. The trait factors were permitted to correlate with one another. The method factors were assumed to be uncorrelated with one another and with the trait factors, which is a standard approach in specifying such models (e.g., Jarvis and Petty, 1996; Krosnick and Alwin, 1988; Saris and Andrews, 1991). If the method factors are allowed to correlate with each other this will mostly lead to

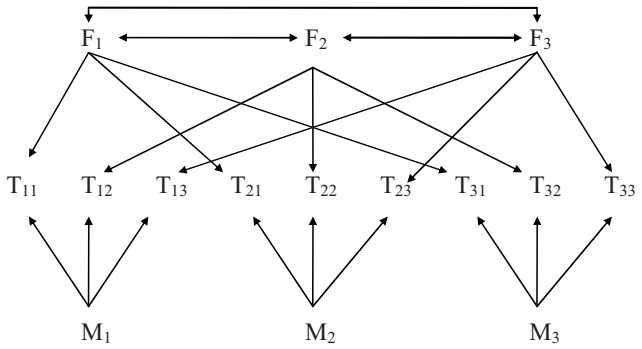

Figure 1. Path Diagram of the Relations Between the Traits (F), Methods (M), and the True Scores (T)

empirical identification problems[2] as described by Rindskopf (1984), Saris (1990) and Kenny and Kashy (1992). Restricting these correlations to zero seems to be too strong an assumption; however Scherpenzeel (1995) showed that the effect of these correlations on the quality coefficients is, usually, very minimal even if these correlations are close to 1. Therefore we have followed the policy of introducing these correlations only if the analysis shows that they are absolutely necessary for the fit of the model.[3] Even then the introduction will be done step by step and not in combination in order to avoid the above mentioned identification problems.

If all variables other than the error term (e_{ij}) are standardized, the parameters can be interpreted as follows:

- h_{ij} is a measure's reliability coefficient.
- h_{ij}^2 is the measure's reliability, i.e., $1 - var(e_{ij})$.
- v_{ij} is the measure's validity coefficient.
- v_{ij}^2 is the measure's validity.
- m_{ij} is the method effect coefficient, where $m_{ij}^2 = 1 - v_{ij}^2$, meaning that the method effect is equal to the systematic invalidity of the measure.

Furthermore, it follows from this model that $q_{ij}^2 = h_{ij}^2 * v_{ij}^2$ is the explained variance of the observed variable by the variable of interest F_j. We will denote this coefficient as the quality of the indicator.

According to this model, the correlations between observed variables decrease if random error increases (i.e., reliability decreases). Method effects can make correlations between variables observed with the same method more positive or less negative. Consequently, observed correlations do not simply provide valid estimates of the correlation between the variables of interest, because a correlation can be inflated by method effects and attenuated by unreliability and invalidity. Therefore, we compare data quality across measures focusing on estimates of the amount of random error variance and the reliability and validity coefficients.

Although MTMM analyses usually require at least three measures of at least three traits in a single sample of respondents, our design provides only two measures of three traits in a single sample. This might seem to leave the model in Equations (1) and (2) under-identified. However, because we have data from two independent samples, both of which provide estimates of some of the same correlations and underlying relations among latent variables, it is in fact possible to estimate the model's parameters via the multisample approach in structural equation modeling (e.g., LISREL, Jöreskog and Sörbom, 1991). The statistical justification for our approach has been discussed extensively by Saris, Satorra and Coenders (2004) and is based on the earlier papers of Allison (1987) and Satorra (1990, 1992).

For each experiment the above specified model was the starting point. The program Jrule (Van der Veld, Saris and Satorra 2009) based on the work of Saris, Satorra and Van der Veld (2009) has been used to detect misspecifications in the starting model. If such misspecifications were detected the model was adjusted accordingly.

The specification and results of the SB-MTMM experiments

The experiments we will report about have been performed in the second and third rounds of the ESS. In all cases the SB-MTMM design was used while the randomly selected subgroups contained a third of the total sample which was, in most countries, except the smallest ones, between 1500 – 2000 cases. So the subsamples are in general larger than 500 cases. With respect to the sampling, response rates, translation and the fieldwork procedures we refer to the above literature.

ESS experiments in Round 2

In the second round two experiments were done where A/D scales were compared with IS scales. In the first experiment the IS scale was used before the A/D scales. In the second experiment the order was the other way around.

Experiment 1: The social distance between doctors and patients

In the main questionnaire of the first experiment IS questions have been asked. In this case the same scale has been used for all questions (see table 1). It is essential to note that if we are interested in assessing the same dimension for different questions, then, the response categories might be the same for these questions. But we will still call them "Item Specific" if they correspond to the dimension we are interested in. This is typically the case in this experiment because we want to know the frequency of an event in each question.

[2] In split ballot MTMM designs with two groups there are not even data to estimate some of these correlations because of data missing by design.

[3] Eid (2000) suggests omitting one method factor but allows for correlations between the method factors left. However this approach does not facilitate the estimation of the method effects for all three methods and is therefore rejected. The solution of Marsh and Baily (1991) to introduce more than 3 traits in the design is a good alternative that is applied when possible but every extra trait will require a more complex design and additional costs.

Table 1: Experiment 1 of round 2

	Introduction	Statements	Answer categories
Main questionnaire	Please indicate how often you think the following applies to doctors in general	- Doctors keep the whole truth from their patients - GPs treat their patients as their equals - Before doctors decide on a treatment, they discuss it with their patient	- never or almost never - some of the time - about half of the time - most of the time - always or almost always
IS			
SC group 1 A/D	Please indicate how much you agree or disagree with each of the following statements about doctors in general	- Doctors rarely keep the whole truth from their patients - GPs rarely treat their patients as their equals - Before doctors decide on a treatment, they rarely discuss it with their patient	- agree strongly - agree - neither disagree not agree - disagree - disagree strongly
SC group 2 A/D	Please indicate how much you agree or disagree with each of the following statements about doctors in general.	- Doctors usually keep the whole truth from their patients - GPs usually treat their patients as their equals - Before doctors decide on a treatment, they usually discuss it with their patient.	- agree strongly - agree - neither disagree not agree - disagree - disagree strongly

Table 2: Means and standard deviations (in brackets) of the reliability, validity and quality estimates across 14 countries for experiment 1 of round 2 of the ESS for the different methods

	Reliability r^2			Validity v^2			Quality q^2		
Method	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
IS	.21	.76	.74	1	1	1	.21	.76	.74
	(.13)	(.12)	(.07)	(0)	(0)	(0)	(.13)	(.12)	(.07)
A/D(rarely)	.20	.45	.46	.91	.37	.41	.20	.18	.20
	(.24)	(.11)	(.09)	(.25)	(.17)	(.13)	(.25)	(.12)	(.10)
A/D(usually)	.51	.59	.59	.94	.68	.62	.48	.40	.37
	(.23)	(.09)	(.13)	(.08)	(.11)	(.14)	(.23)	(.09)	(.10)

The supplementary questionnaires are presented to the respondents after the main questionnaire. Two A/D forms of the same questions were presented to two random subgroups of the sample. The difference between the two A/D forms is the position of the statement on the frequency scale for the events. In the first sub-group the frequency term used was "rarely" whereas in the second sub-group the frequency term used was "usually". In both cases the same 5 answer categories are proposed. More details can be found in table 1.

The results. Table 2 summarizes the results of this experiment across countries. The right part of this table very clearly shows the big difference in quality between the IS scale (around .75) and the A/D scales (between .18 and .40) with the exception of the first item ("Doctors keep the whole truth from their patients"). This item turned out to be an item which the respondents could not answer very well. The quality of the measures was between .20 and .48 depending on the method, with very large standard deviations. This is most likely the case because individuals do not know what the doctor does. This is not the case for the other two items ("GPs treat their patients as their equals" and "Before doctors decide on a treatment, they discuss it with their patient"). This may explain why the difference in quality between question 1 and questions 2 and 3 is so large.

Looking at the left part of the table, it appears that, for the second and the third questions, both quality indicators (i.e. reliability and validity) are much lower for the A/D scales than for the IS scales, but the biggest difference is found for the validity. Since the method effect $m_{ij}^2 = 1 - v_{ij}^2$, we conclude that we observe more method effects when A/D scales are used than when IS scales are used. This can be an indication of the acquiescence bias mentioned at the beginning of this paper but can also be due to other method related response behavior.

In order to get an idea of the results per country, the results from all 14 countries are presented in table 3; as can be seen, the results are quite comparable. However, to limit the size of the table the decomposition into reliability and validity is not included.

This table shows that in all countries the IS scale was of better quality for the second and third questions while the quality of the first question was, in general, very bad even

Table 3: The quality of the three questions of experiment 1 in Round 2 of the ESS for the different methods

Country	Question 1	Question2	Question 3	Country	Question 1	Question2	Question 3
Austria				Greece			
IS	.38	.74	.67	IS	.30	.81	.81
A/D	.10	.18	.13	A/D	.31	.05	.09
A/D	.64	.46	.47	A/D	.50	.52	.49
Belgium				Luxembourg			
IS	.18	.71	.74	IS	.10	.98	.83
A/D	.00	.12	.18	A/D	.09	.03	.09
A/D	.64	.31	.30	A/D	.10	.21	.20
Czech Republic				Poland			
IS	.06	.81	.83	IS	.15	.66	.77
A/D	.27	.10	.19	A/D	.15	.11	.12
A/D	.36	.48	.19	A/D	.16	.40	.33
Denmark				Portugal			
IS	.05	.74	.77	IS	.26	.98	.66
A/D	.04	.32	.19	A/D	.98	.34	.32
A/D	.66	.39	.48	A/D	.94	.37	.39
Estonia				Slovenia			
IS	.42	.85	.83	IS	.29	.67	.71
A/D	.14	.20	.21	A/D	.29	.48	.43
A/D	.66	.48	.46	A/D	.30	.45	.47
Finland				Sweden			
IS	.01	.61	.64	IS	.19	.72	.76
A/D	.06	.12	.13	A/D	.19	.13	.16
A/D	.60	.42	.32	A/D	.30	.37	.27
Germany				United Kingdom			
IS	.31	.56	.71	IS	.18	.81	.67
A/D	.04	.20	.22	A/D	.06	.16	.30
A/D	.46	.51	.43	A/D	.52	.40	.32

for the IS method. In many countries for this question the A/D format was of better quality. We can also see that the scale with the high frequency word (usually) was better than the items with the low frequency word (rarely).

In this experiment the IS scale was presented first so the difference cannot be due to memory effects. In the case of Luxembourg, the differences are really huge: the quality of the second item is .98 using the first method (IS scale), and only .03 using the second method. For the third trait, it is respectively .83 and .09, so the differences can be very extreme, but they are only present for the second and third items. For some reason the IS scale is not working very well for the first item. For the moment we think that this is due to respondents' lack of knowledge regarding the behavior of doctors. This seems to be indicated by the very low reliability of this measure (cf. table 2).

Experiment 2: Opinions about work

In the second experiment of round 2, a battery of items using *not* the standard A/D scale but a scale with truth categories has been compared with IS scales: one IS scale is a 4 point scale, the other is an 11 point scale. The battery of A/D questions was presented in the main questionnaire; the IS scales were presented in the supplementary questionnaire to two random subgroups of the original sample. The first formulation was a bit different because it did not use the A/D format but the categories "not at all true" till "very true". Nevertheless, the important difference with the other two formulations is that statements have been used in which a value for the specific variable content was specified. In the IS forms of the second and the third formulation the question was directed to the evaluation of this characteristic directly. For more details see table 4.

The results. In order to get a general picture of the results table 5 presents the quality estimates across countries.

This experiment also very clearly shows the much higher quality of the IS scales over the battery questions with a fixed scale. We can see that in this case there is no difference in quality between the 4 and 11 point IS scale. This may be due to the fact that the respondents confronted with the 11 point scale use different maximum values for their responses. For more details on this issue see Saris (1986). Finally, we must mention that the lower quality in this experiment is not due to a lower validity but to a lower reliability. The validity is very high (close to 1) for all three methods.

In order to see the differences in quality in the different countries we also present the estimates for the different scales and questions for each country in table 6.

With the exception of Belgium, in all other countries the IS scale with 4 categories was of better quality than the battery

Table 4: Experiment 2 of round 2

	Introduction	Statements	Answer categories
Main questionnaire "A/D"	Using this card, please tell me how true each of the following statements is about your current job.	- There is a lot of variety in my work - My job is secure - My health or safety is at risk because of my work	- not at all true - a little true - quite true - very true
SC group 1 IS	The next 3 questions are about your current job.	- Please choose one of the following to describe how varied your work is. - Please choose one of the following to describe how secure your job is - Please choose one of the following to say how much, if at all, your work puts your health and safety at risk.	- not at all varied - a little varied - quite varied - very varied (same type of response scale using terms secure and safe instead of varied)
SC group 2 IS		- Please indicate, on a scale of 0 to 10, how varied your work is, where 0 is not at all varied and 10 is very varied. - Now please indicate, on a scale of 0 to 10, how secure your job is, where 0 is not at all secure and 10 is very secure. - Please indicate, on a scale of 0 to 10, how much your health and safety is at risk from your work, where 0 is not at all at risk and 10 is very much at risk.	Horizontal 11 point scale only labelled at the end points

Table 5: The means reliability, validity and quality of the three questions of experiment 2 in Round 2 of the ESS across 10 countries for the different methods (standard deviations in brackets)

	Reliability r^2			Validity v^2			Quality q^2		
Method	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
A/D(4)	.65	.59	.61	.99	.98	.99	.64	.58	.60
	(.09)	(.18)	(.15)	(.02)	(.03)	(.03)	(.10)	(.18)	(.15)
IS(4)	.80	.80	.80	1	1	1	.80	.80	.80
	(.14)	(.13)	(.14)	(0)	(0)	(0)	(.14)	(.13)	(.14)
IS(11)	.81	.83	.77	.98	.98	.98	.80	.82	.76
	(.09)	(.11)	(.12)	(.03)	(.03)	(.04)	(.10)	(.12)	(.14)

using a truth scale with the same number of categories for all three questions (around .7 to .9 versus .5 to .6). The position of the IS scale in the supplementary questionnaire is not an issue as the better quality of the IS scale is also observed both when it comes first and when it comes later.

Possibly the order of the observations with the different scale types has an impact on the size of the differences since we see fewer differences in this second experiment than in the first, but this may also be linked to the subject matter of the experiments or to other characteristics of the methods used (such as the number of points). More research is needed to determine this, however the important point here is that in different combinations, the superiority of the IS in terms of quality holds.

In addition, the IS scale with 11 categories scale was of better quality in all countries and for all three questions with the exception of Belgium. In 6 out of the 10 countries the IS scale with 11 categories was also better than the IS scale with 4 categories. So, not only might the kind of scale (IS versus A/D) impact the total quality of a measure, but so might the length of the scale (number of response categories). However, it seems that this effect varies across countries.

Experiments in Round 3 of the ESS

In round 3 of the ESS again two SB-MTMM experiments have been done which allow the comparison of the IS scales with A/D scales. The attraction of these experiments is that by now the CCT of the ESS was convinced of the better quality of the IS scales so that they were used in the main questionnaire. This means that the A/D scales are now second in order. So, if the former are better this cannot be due to memory effects.

Table 6: The quality of the three questions of experiment 2 in Round 2 of the ESS for the different methods

Country	Question 1	Question 2	Question 3
Austria			
A/D(4)	.58	.59	.58
IS(4)	.90	.86	.90
IS(11)	.79	.85	.74
Belgium			
A/D(4)	.88	.88	.92
IS(4)	.49	.52	.52
IS(11)	.56	.56	.56
Czech Republic			
A/D(4)	.59	.44	.52
IS(4)	.85	.88	.81
IS(11)	.81	.83	.86
Denmark			
A/D(4)	.69	.67	.74
IS(4)	.74	.69	.83
IS(11)	.81	.80	.46
Finland			
A/D(4)	.59	.62	.41
IS(4)	.88	.90	.85
IS(11)	.74	.76	.81
Germany			
A/D(4)	.67	.74	.67
IS(4)	.76	.86	.77
IS(11)	.85	.85	.81
Luxembourg			
A/D(4)	.62	.52	.46
IS(4)	.86	.85	.94
IS(11)	.94	.98	.85
Slovenia			
A/D(4)	.61	.21	.55
IS(4)	.67	.69	.66
IS(11)	.74	.71	.74
Spain			
A/D(4)	.53	.53	.56
IS(4)	.98	.92	.98
IS(11)	.86	.98	.98
Sweden			
A/D(4)	.61	.62	.56
IS(4)	.83	.83	.77
IS(11)	.85	.94	.81

Experiment 1: Opinions about immigration

In this experiment the IS scale with 4 categories, presented in the main questionnaire, was compared with a standard 5 point A/D scale presented in the supplementary questionnaire. The second form was an IS question, identical to the first one. The last procedure used in a third random sample was an A/D scale with 7 categories. More details can be found in table 7.

The results. In order to give an impression of the differences in quality across the different countries the mean qualities over all countries are presented in table 8, as well as the mean reliabilities and validities.

This table shows that the IS scale with 4 categories is of better quality on average than the 5 point A/D scale and for all three questions even though the A/D question was the second in order and had more categories. In fact the A/D scale with 7 categories also is of less quality than 4 point IS scale for all three questions. This holds not only for the IS scale which was presented before the others in the main questionnaire but also for the identical IS scale which was presented in one of the random subgroups of the sample at the same time as the A/D scales. In this experiment, the decomposition into reliability and validity indicates that the lower quality is not due to lower reliability but comes from a lower validity and so comes from higher method effects.

In order to give an impression of the size of the difference in quality in the different countries, table 9 presents the quality for all questions and scale types.

The table shows that there is a big difference in quality in all countries between the IS scale and the A/D scales even if the latter have more categories than the IS scale and if the IS

Table 7: Experiment 1 of round 3

	Introduction	Statements	Answer categories
Main questionnaire + SC group 2		- Now, using this card, to what extent do you think [country] should allow people of the *same race or ethnic group* as most [country's] people to come and live here?	- allow many to come and live here - allow some - allow a few - allow none
IS		- How about people of a different race or ethnic group from most [country] people? Still use this card - How about people from the poorer countries outside Europe?	
SC group 1 A/D	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	- [Country] should allow more people of the *same race or ethnic group* as most [country's] people to come and live here - [Country] should allow more people of a *different* race or ethnic group from most [country's] people to come and live here - [Country] should allow more people from the *poorer countries outside Europe* to come and live here	- agree strongly - agree - neither agree nor disagree - disagree - disagree strongly
SC group 3 A/D	Same as group 1	Same as group 1	7 point A/D scale only labelled at the end points (from "disagree strongly" to "agree strongly")

Table 8: The means reliability, validity and quality of the three questions of experiment 1 in Round 3 of the ESS across 23 countries for the different methods (standard deviations in brackets)

	Reliability r^2			Validity v^2			Quality q^2		
Method	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
IS(4)	.85	.89	.88	1	1	1	.85	.89	.88
	(.07)	(.04)	(.04)	(0)	(0)	(0)	(.07)	(.04)	(.04)
A/D(5)	.74	.92	.80	.65	.55	.64	.48	.51	.51
	(.10)	(.04)	(.05)	(.12)	(.09)	(.09)	(.10)	(.08)	(.08)
IS(4)	.86	.94	.87	.88	.82	.87	.76	.77	.76
	(.06)	(.04)	(.05)	(.06)	(.07)	(.07)	(.09)	(.07)	(.08)
A/D(7)	.72	.96	.81	.61	.50	.57	.44	.47	.46
	(.12)	(.04)	(.06)	(.11)	(.08)	(.08)	(.10)	(.07)	(.07)

scale is used before or at the same moment as the A/D scales.

Experiment 2: Opinions about consequences of immigration

In the main questionnaire an 11 point IS scale was used for all three items. This scale was compared with a 5 point and 7 point A/D scale and an 11 point A/D scale. The latter three scales were presented to three random subgroups of the sample. Table 10 gives more details.

Results. In order to give an impression of the differences in quality across the different countries the quality in average across all countries is presented in table 11.

This table shows again that the IS scale is much better than any of the other measures for all questions, and that this is due to a higher validity, so less method effects. For the third trait, for instance, the difference in quality between the 11 point IS scale and the 11 point A/D scale is .44 (=.76-.32); the difference in reliability is .05 only (=.76-.71), and the difference in validity is .54 (=1-.46). It is also clear that for all three questions the quality of the 5 points A/D scale is better than the quality of the A/D scales with more categories but still much lower than for the IS scale. This issue deserves further research (Revilla, Saris and Krosnick, 2009).

Table 9: The quality of the different scales for the different questions

Country	Question 1	Question 2	Question 3
Austria			
IS(4)	.83	.92	.90
A/D(5)	.52	.55	.54
IS(4)	.75	.80	.80
A/D(7)	.47	.40	.45
Belgium			
IS(4)	.79	.85	.81
A/D(5)	.41	.48	.47
IS(4)	.83	.75	.78
A/D(7)	.38	.49	.51
Bulgaria			
IS(4)	.90	.90	.90
A/D(5)	.61	.65	.72
IS(4)	.81	.78	.83
A/D(7)	.50	.55	.53
Switzerland			
IS(4)	.77	.92	.86
A/D(5)	.45	.47	.48
IS(4)	.89	.83	.83
A/D(7)	.42	.42	.40
Cyprus			
IS(4)	.94	.92	.92
A/D(5)	.57	.52	.49
IS(4)	.76	.76	.75
A/D(7)	.67	.54	.41
Germany			
IS(4)	.85	.86	.90
A/D(5)	.53	.52	.54
IS(4)	.85	.80	.80
A/D(7)	.46	.48	.51
Denmark			
IS(4)	.66	.86	.88
A/D(5)	.59	.61	.59
IS(4)	.75	.76	.75
A/D(7)	.44	.49	.50
Estonia			
IS(4)	.83	.81	.85
A/D(5)	.43	.45	.43
IS(4)	.64	.72	.67
A/D(7)	.43	.56	.45
Spain			
IS(4)	.94	.94	.94
A/D(5)	.56	.53	.55
IS(4)	.89	.87	.90
A/D(7)	.51	.49	.52
Finland			
IS(4)	.88	.90	.79
A/D(5)	.48	.52	.53
IS(4)	.80	.76	.83
A/D(7)	.37	.48	.40
France			
IS(4)	.83	.85	.83
A/D(5)	.41	.47	.55
IS(4)	.80	.85	.87
A/D(7)	.39	.45	.47
United Kingdom			
IS(4)	.81	.88	.88
A/D(5)	.47	.55	.51
UK continued			
IS(4)	.75	.78	.78
A/D(7)	.31	.42	.39
Ireland			
IS(4)	.85	.90	.88
A/D(5)	.33	.34	.37
IS(4)	.57	.59	.53
A/D(7)	.33	.43	.43
Latvia			
IS(4)	.94	.92	.85
A/D(5)	.56	.52	.51
IS(4)	.73	.74	.71
A/D(7)	.42	.41	.42
Netherlands			
IS(4)	.88	.92	.88
A/D(5)	.19	.34	.33
IS(4)	.68	.63	.65
A/D(7)	.23	.30	.29
Norway			
IS(4)	.85	.90	.88
A/D(5)	.39	.49	.53
IS(4)	.63	.76	.70
A/D(7)	.34	.40	.45
Poland			
IS(4)	.86	.88	.90
A/D(5)	.47	.52	.43
IS(4)	.83	.88	.83
A/D(7)	.47	.54	.48
Portugal			
IS(4)	.94	.92	.96
A/D(5)	.46	.45	.47
IS(4)	.73	.75	.70
A/D(7)	.59	.60	.56
Romania			
IS(4)	.94	.94	.94
A/D(5)	.60	.67	.61
IS(4)	.90	.90	.74
A/D(7)	.57	.63	.61
Russia			
IS(4)	.88	.90	.90
A/D(5)	.57	.50	.53
IS(4)	.67	.74	.75
A/D(7)	.54	.46	.46
Slovenia			
IS(4)	.83	.83	.85
A/D(5)	.50	.49	.50
IS(4)	.76	.78	.80
A/D(7)	.39	.41	.41
Slovakia			
IS(4)	.79	.90	.86
A/D(5)	.46	.54	.51
IS(4)	.78	.79	.75
A/D(7)	.40	.45	.40
Ukraine			
IS(4)	.83	.90	.94
A/D(5)	.49	.55	.57
IS(4)	.73	.73	.78
A/D(7)	.46	.49	.53

Table 10: Experiment 2 of round 3

	Introduction	Statements	Answer categories
Main questionnaire IS	- Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries? - 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? - Is [country] made a worse or a better place to live by people coming to live here from other countries?	11 point scale from "bad for the economy" to "good for the economy" only labelled at the end points (same with "cultural life undermined/enriched" and "worse place to live/better place to live")	
SC group 1 A/D	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.	- It is generally bad for [country's] economy that people come to live here from other countries - [Country's] cultural life is generally undermined by people coming to live here from other countries - [Country] is made a worse place to live by people coming to live here from other countries	standard 5 point A/D scale
SC group 2 A/D		- How much do you agree or disagree that it is generally bad for [country's] economy that people come to live here from other countries? - 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 to live here from other countries?	11 point A/D scale only labelled at the end points ("disagree strongly" to "agree strongly")
SC group 3 A/D	Same as group 1	Same as group 1	7 point A/D scale only labelled at the end points ("disagree strongly" to "agree strongly")

Table 11: The means 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)

	Reliability r^2			Validity v^2			Quality q^2		
Method	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
IS(11)	.76	.81	.76	1	1	1	.76	.81	.76
	(.07)	(.05)	(.08)	(0)	(0)	(0)	(.07)	(.05)	(.08)
A/D(5)	.66	.77	.74	.66	.59	.65	.44	.45	.47
	(.08)	(.08)	(.07)	(.15)	(.14)	(.16)	(.11)	(.11)	(.12)
A/D(11)	.56	.74	.71	.37	.38	.46	.21	.28	.32
	(.12)	(.11)	(.10)	(.20)	(.20)	(.17)	(.13)	(.15)	(.13)
A/D(7)	.56	.73	.73	.48	.42	.48	.27	.31	.35
	(.10)	(.12)	(.09)	(.19)	(.14)	(.17)	(.11)	(.12)	(.13)

In order to give an impression of the size of the differences in quality of the different scales in the different countries table 12 presents the quality for the three questions for the four types of scales in each country.

This table again shows that for this topic the differences in quality are also too big in all countries to be ignored. The IS scales lead to a much better quality of the measure.

Discussion

Many prior studies have documented that some people answer A/D questions by agreeing with any assertion, regardless of its content (Krosnick and Fabrigar, forthcoming). Furthermore, we outlined earlier how the cognitive processes entailed in answering an A/D question are likely to be more burdensome and complex than the cognitive processes entailed in answering a comparable IS response option question. And we outlined why responses to A/D items do not necessarily have monotonic relations with the underlying constructs. Presumably, because questions with IS response options avoid acquiescence, minimize cognitive burden, and do indeed produce answers with monotonic relations with the underlying constructs of interest, this format may yield better quality self-reports, where the quality, as mentioned earlier, is defined as the product of the reliability and the validity.

Few previous studies have compared the amount of measurement error in responses to A/D questions with the amount of such error in responses to comparable questions asked with IS response options. We have reported 4 studies using the SB-MTMM design in many different countries. The evidence from all these studies is consistent with the conclusion that data quality is indeed much higher for questions offering IS response options. Very few exceptions have been found, and the differences in quality were in general quite large.

Many aspects of the design have been manipulated, but still the same conclusion is drawn. It did not matter whether the IS scale was asked before or after the A/D scale or at the same time. Even if the A/D scale had more answer categories the IS scale with fewer categories was still of higher quality. The mode in which the questions were asked (face-to-face interview or selfcompletion) also did not change this general tendency. So the better quality of IS scales is a quite general and robust result, which holds across different topics, countries, modes and ordering of the questions in the experiments. More detailed analyses would be needed in order to determine more specifically the exact impact of these different choices on the quality of the A/D scale instead of an IS scale. But the main tendency appears very clearly in the analyses to be, that the IS scales are of a much higher quality than the A/D scales.

Lord and Novick (1968) and others have shown that lack of quality will reduce the correlations between variables. Therefore these results suggest that researchers should seriously consider changing the standard procedure of using batteries of A/D items because the difference in quality with item specific scales is too large to be ignored.

Looking at the reliability and validity separately we can observe that in all experiments and for all questions the IS scales were better than the A/D scales with few exceptions with respect to validity. This is not always true for reliability. For the question for which people lacked information (round 2 experiment 1 question 1) it was not true. In this case one A/D format had greater reliability than the IS questions. There is also one question (round 3 experiment 1 question 2) for which the reliabilities were approximately equal. In all other cases the reliability of the IS questions was better and the validity was always better. In all ESS experiments the validity was even considerably better for the IS scales.

These results suggest that individuals are less certain when choosing a position on the A/D scale than on an IS scale, leading to lower reliability. In addition, and more importantly, they develop a response pattern to A/D questions which varies from individual to individual creating correlations between items which have nothing to do with the substantive variables asked about. This phenomenon leads to method effects. In the introduction we have given two explanations for this: one is that individuals solve their uncertainty about the way to use the A/D scales in different but stable ways across questions; the second is that they give arbitrary answers to questions no matter their formulation. This is called acquiescence. Both explanations are possible. Further research is needed to determine what happens in this case. However what is certain is that the A/D scales perform more poorly than the IS scales.

The first experiment in Round 2 also indicated that A/D items using the term “rarely” had lower reliabilities than items using the term “usually”. This finding is consistent with past studies that used very different methods to demonstrate that people make more cognitive errors when they have to disagree with a negative statement than they make when they have to agree with affirmative statements. We documented this general phenomenon here using a new methodological approach and a new indicator of data quality.

This finding also suggests, for several reasons, caution before presuming that battery balancing is an effective and wise solution to the acquiescence problem. First, negation items bring with them an inherent cost: lower data quality due to reduced reliability. And the greater cognitive difficulty entailed in generating answers to these items is likely to enhance respondent fatigue, which may compromise the quality of individuals’ responses to items later in a questionnaire. Furthermore, the “balancing” approach simply places all acquiescing respondents at or near the middle of the response dimension, regardless of the fact that there is no reason to believe that these individuals belong there. This relatively arbitrary placement of those individuals may hurt data quality as well. Therefore, solving the acquiescence problem seems to be accomplished more effectively by using questions with IS response options instead of by balancing large batteries of A/D questions.

One may also think that using more categories may help to improve the quality of the A/D questions. However, the two last experiments indicate that this solution seems doubtful because in these experiments we see that the quality of the 5 point A/D scale is better than the quality of the 7 and 11 points scale. This issue has received more attention (Revilla

Table 12: The quality of the different scales for three different questions in each country

Country	Question 1	Question 2	Question 3
Austria			
IS(11)	.81	.83	.79
A/D(5)	.46	.51	.56
A/D(11)	.32	.37	.46
A/D(7)	.32	.33	.32
Belgium			
IS(11)	.72	.79	.64
A/D(5)	.51	.48	.63
A/D(11)	.24	.35	.41
A/D(7)	.29	.38	.47
Bulgaria			
IS(11)	.71	.81	.85
A/D(5)	.30	.31	.33
A/D(11)	.13	.18	.22
A/D(7)	.22	.29	.32
Switzerland			
IS(11)	.71	.85	.67
A/D(5)	.50	.60	.60
A/D(11)	.20	.46	.36
A/D(7)	.49	.57	.57
Cyprus			
IS(11)	.81	.86	.83
A/D(5)	.47	.55	.47
A/D(11)	.53	.55	.41
A/D(7)	.36	.43	.42
Germany			
IS(11)	.77	.79	.79
A/D(5)	.43	.49	.56
A/D(11)	.32	.41	.51
A/D(7)	.38	.48	.59
Denmark			
IS(11)	.74	.83	.79
A/D(5)	.61	.59	.60
A/D(11)	.40	.53	.55
A/D(7)	.41	.44	.50
Estonia			
IS(11)	.55	.77	.81
A/D(5)	.41	.37	.35
A/D(11)	.17	.22	.25
A/D(7)	.22	.24	.31
Spain			
IS(11)	.83	.77	.69
A/D(5)	.46	.56	.51
A/D(11)	.24	.17	.27
A/D(7)	.21	.28	.43
Finland			
IS(11)	.71	.76	.74
A/D(5)	.60	.52	.63
A/D(11)	.38	.36	.51
A/D(7)	.37	.14	.36
France			
IS(11)	.79	.85	.77
A/D(5)	.55	.64	.61
A/D(11)	.31	.52	.48
A/D(7)	.25	.44	.43
United Kingdom			
IS(11)	.81	.83	.83
UK continued			
A/D(5)	.41	.49	.59
A/D(11)	.28	.38	.44
A/D(7)	.31	.36	.42
Ireland			
IS(11)	.77	.77	.81
A/D(5)	.37	.33	.39
A/D(11)	.02	.09	.14
A/D(7)	.16	.12	.27
Latvia			
IS(11)	.81	.90	.86
A/D(5)	.24	.28	.24
A/D(11)	.05	.07	.08
A/D(7)	.10	.11	.13
Netherlands			
IS(11)	.72	.69	.62
A/D(5)	.38	.35	.47
A/D(11)	.23	.24	.30
A/D(7)	.29	.23	.32
Norway			
IS(11)	.72	.79	.77
A/D(5)	.67	.57	.58
A/D(11)	.09	.32	.43
A/D(7)	.36	.42	.38
Poland			
IS(11)	.69	.81	.67
A/D(5)	.33	.31	.39
A/D(11)	.10	.13	.18
A/D(7)	.19	.20	.18
Portugal			
IS(11)	.83	.81	.86
A/D(5)	.47	.39	.43
A/D(11)	.18	.22	.27
A/D(7)	.40	.35	.45
Romania			
IS(11)	.88	.85	.79
A/D(5)	.29	.39	.44
A/D(11)	.08	.14	.22
A/D(7)	.17	.19	.20
Russia			
IS(11)	.77	.83	.83
A/D(5)	.42	.46	.44
A/D(11)	.36	.33	.34
A/D(7)	.27	.33	.29
Slovenia			
IS(11)	.81	.79	.74
A/D(5)	.37	.36	.38
A/D(11)	.01	.10	.22
A/D(7)	.13	.20	.22
Slovakia			
IS(11)	.67	.69	.56
A/D(5)	.32	.31	.26
A/D(11)	.12	.14	.15
A/D(7)	.14	.22	.16
Ukraine			
IS(11)	.81	.88	.83
A/D(5)	.44	.49	.46
A/D(11)	.17	.20	.25
A/D(7)	.12	.26	.27

et al., 2009) and the result is that increasing the number of categories of A/D scales reduces the quality instead of increasing it. This is, thus, also not a solution to the problem.

A/D scales are especially attractive for researchers because they save time for the researcher and the interviewer. However, this advantage is obtained at the cost of the effort of respondents. We have shown here that respondents make many more errors using these scales than IS scales. Therefore the quality of A/D scales is lower than the quality of IS scales. For this reason we advise the use of IS scales whenever possible.

This is, of course, a problem if a specific survey has used the A/D format in past years. What should be done for the next rounds in that case? In a time series perspective, persisting in the use of A/D scales seems the only option, since introduction of IS scales would break the continuity. But if one has to deal with A/D scales, great care should be taken in interpreting the results, always keeping in mind the drawbacks of such scales. In addition, one should introduce a method factor in doing multivariate analysis. Another possibility would be to introduce in future rounds the same kinds of SB-MTMM experiments used in this paper, in order to be able to evaluate the reliability, validity and quality of the different scales and in subsequent rounds to be able to correct for measurement errors.

Acknowledgements

This work has been made possible by a grant from the European Commission to the European Social Survey. We are very grateful to Albert Satorra, Germ Coenders and two anonymous reviewers for their helpful comments.

References

Aderman, D., & Brehm, S. S. (1976). On the recall of initial attitudes following counterattitudinal advocacy: An experimental reexamination. *Personality and Social Psychology Bulletin*, *2*, 59-62.

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. C. Clogg (Ed.), *Sociological methodology 1987* (p. 71-102). Washington, DC: American Sociological Association.

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

Andrews, F. M. (1984). Construct validity and error components of survey measures: A Converse structural modeling approach. *Public Opinion Quarterly*, *48*, 409-422.

Bem, D. J., & McConnell, H. K. (1970). Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes. *Journal of Personality and Social Psychology*, *14*, 23-31.

Berkowitz, N. H., & Wolken, G. H. (1964). A forced choice form of the F scale-Free of acquiescent response set. *Sociometry*, *27*, 54-65.

Billiet, J., Koch, A., & Philippens, M. (2007). Understanding and improving response rates. In R. F. R. Jowell C. Roberts & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the european social survey* (p. 113-139). London: Sage.

Billiet, J., Loosveldt, G., & Waterplas, L. (1985). *Het Survey-Interview Onderzocht: Effecten Van Het Ontwerp En Gebruik Van Vragenlijsten Op De Kwaliteit Van De Antwoorden.* [Research on Surveys: Effects of the Design and Use of Questionnaires on the Quality of the Response]. Leuven, Belgium: Sociologisch Onderzoeksinstituut KU Leuven.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology*, *37*, 1-21.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45-73.

Carr, L. G. (1971). The Srole items and acquiescence. *American Sociological Review*, *36*, 287-293.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language*. New York: Harcourt Brace.

Coenders, G., & Saris, W. E. (2000). Testing additive and multiplicative MTMM models. *Structural Equation Modeling*, *7*, 219-251.

Corten, I. W., Saris, W. E., Coenders, G., M.van der Veld, W., Aalberts, C. E., & Kornelis, C. (2002). Fit of different models for multitrait-multimethod experiments. *Structural Equation Modeling*, *9*(2), 213-232.

Costa, P. T., & McCrae, R. R. (1988). From catalog to classification: Murray's needs and the five-factor model. *Journal of Personality and Social Psychology*, *55*, 258-265.

Counte, M. A. (1979). An examination of the convergent validity of three measures of patient satisfaction in an outpatient treatment center. *Journal of Chronic Diseases*, *32*, 583-588.

Davis, J. A., & Smith, T. M. (1996). *General social surveys, 1972-1996: Cumulative codebook*. Chicago: National Opinion Research Center.

Eid, M. (2000). A Multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*, 242-261.

Fowler, F. J. (1995). Improving Survey Questions: Design and Evaluation. *Applied Social Research Methods Series*, *38*, 56-57.

Goethals, G. R., & Reckman, R. F. (1973). The perception of consistency in attitudes. *Journal of Experimental Social Psychology*, *9*, 491-501.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216-1229.

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

Harkness, J. (2007). Improving the comparability of translations. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (p. 79-95). London: Sage.

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology*, *70*, 172-194.

Jöreskog, K. G., & Sörbom, D. (1991). *LISREL VII: A guide to the program and applications*. Chicago, IL: SPSS.

Kenny, D. A., & Kashy, D. A. (1992). The analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*, 165-172.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant

correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*, 526-538.

Krosnick, J. A., & Fabrigar, L. R. (Forthcoming). *Designing great questionnaires: Insights from psychology*. New York: Oxford University Press.

Leech, G. N. (1983). *Principles of pragmatics*. New York: Longman.

Lenski, G. E., & Leggett, J. C. (1960). Caste, class, and deference in the research interview. *American Journal of Sociology*, *65*, 463-467.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1-55.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental scores*. MA: Addison-Wesley.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, *15*, 47-70.

Miller, W. E., & Traugott, S. (1989). *American national election studies data sourcebook, 1952-1986*. Cambridge, MA: Harvard University Press.

Ray, J. J. (1979). A quick measure of achievement motivation – validated in Australia and reliable in Britain and South Africa. *Australian Psychologist*, *14*, 337-344.

Ray, J. J. (1980). The comparative validity of Likert, projective, and forced-choice indices of achievement motivation. *Journal of Social Psychology*, *111*, 63-72.

Revilla, M., Saris, W. E., & Krosnick, J. A. (2009). *Choosing the number of categories in agree/disagree scales.* RECSM working paper number 5.

Richardson, S. A., Dohrenwend, B. S., & Klein, D. (1965). Expectations and premises: The socalled "leading question". In S. A. Richardson, B. S. Dohrenwend, & D. Klein (Eds.), *Interviewing: Its forms and functions.* New York: Basic Books.

Rindskopf, D. (1984). Structural equation models: empirical identification, Heywood cases, and related problems. *Sociological Methods & Research*, *13*, 109-119.

Robinson, J. P., & Shaver, P. R. (1973). *Measures of social psychological attitudes*. Ann Arbor, Michigan: Institute for Social Research.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes*. San Diego, CA: Academic Press, Inc.

Robinson, J. P., & Wrightsman, L. S. (1999). *Measures of political attitudes*. San Diego, CA: Academic Press.

Ross, C. K., Steward, C. A., & Sinacore, J. M. (1995). A comparative study of seven measures of patient satisfaction. *Medical Care*, *33*, 392-406.

Ross, M., & Shulman, R. F. (1973). Increasing the salience of initial attitudes: Dissonance versus self-perception theory. *Journal of Personality and Social Psychology*, *28*, 138-144.

Saris, W. E. (1986). *Variation in response functions: a source of measurement error in attitude research.* Amsterdam Sociometric Research Foundation.

Saris, W. E. (1990). The choice of a research design for MTMM studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitraitmultimethod studies.* Amsterdam: North Holland.

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

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (p. 575-597). New York: Wiley.

Saris, W. E., & Münnich, A. (1995). *The multitrait-multimethod approach to evaluate measurement instruments*. Budapest, Hungary: Eötvös University Press.

Saris, W. E., & N.Gallhofer, I. (2007). *Design, evaluation and analysis of questionnaires for survey research*. Hoboken: Wiley.

Saris, W. E., Satorra, A., & Coenders, G. (2004). *A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design.* Sociological Methodology 2004.

Saris, W. E., Satorra, A., & Van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561-582.

Satorra, A. (1990). Robustness issues in the analysis of MTMM and RMM models. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies.* Amsterdam: North Holland.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, *22*, 249-278.

Scherpenzeel, A. C. (1995). *A question of quality. Evaluating survey questions by multitraitmultimethod studies*. Leidschendam: PTT Nederland NV, KPN Research.

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

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.

Shaffer, D. R. (1975a). Another look at the phenomenological equivalence of pre- and postmanipulation attitudes in the forced-compliance experiment. *Personality and Social Psychology Bulletin*, *1*, 497-500.

Shaffer, D. R. (1975b). Some effects of consonant and dissonant attitudinal advocacy on initial attitude salience and attitude change. *Journal of Personality and Social Psychology*, *32*, 160-168.

Shaw, M. E., & Wright, J. M. (1967). *Scales for the measurement of attitudes*. New York: McGraw Hill.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response :*. Cambridge: Cambridge University Press.

Trabasso, T., Rollins, H., & Shaughnessey, E. (1971). Storage and verification stages in processing concepts. *Cognitive Psychology*, *2*, 239-289.

Van der Veld, W., Saris, W. E., & Satorra, A. (2008). *JRule 2.0: User manual.* Unpublished document.

van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies.* Amsterdam: North Holland.

Wixon, D. R., & Laird, J. D. (1976). Awareness and attitude change in the forced-compliance paradigm: The importance of when. *Journal of Personality and Social Psychology*, *34*, 376-384.

Appendix: Input Commands for LISREL to Run the Model Shown in Figure 1

```
Analysis of SB-MTMM experiments group 1
! 9 variables, 2 groups, 270 observations, covariance matrix analysed
da ni=9 ng=2 no=270 ma=cm
! correlation matrix and standard deviations obtained from the data for group 1
km
1.0
.467 1.0
.006 -.086 1.0
.671 .368 -.026 1.0
.356 .585 -.076 .313 1.0
-.134 -.078 .399 -.103 -.160 1.0
0 0 0 0 0 0 1.0
0 0 0 0 0 0 0 1.0
0 0 0 0 0 0 0 0 1.0
! standard deviations
sd
.797 1.603 1.110 .890 1.877 1.081 1.00 1.00 1.00

!definition of the structure of the model
model ny=9 ne=9 nk=6 ly=fu,fi be=fu,fi ps=sy,fi te=sy,fi ga=fu,fi ph=sy,fi

!selection of variables for the first order factor model
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
free te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 0 ly 7 7 ly 8 8 ly 9 9
value 1 te 7 7 te 8 8 te 9 9

!second order factor model
free ga 1 1 ga 2 2 ga 3 3 ga 4 1 ga 5 2 ga 6 3 ga 7 1 ga 8 2 ga 9 3
value 1 ph 1 1 ph 2 2 ph 3 3

!traits are correlated
free ph 2 1 ph 3 1 ph 3 2

!3 method effects
free ph 4 4 ph 5 5 ph 6 6

!same impact of the method on the different traits
value 1 ga 1 4 ga 2 4 ga 3 4 ga 4 5 ga 5 5 ga 6 5 ga 7 6 ga 8 6 ga 9 6

out mi iter= 300 adm=off sc

!idem for the second group
analysis of SB-MTMM experiments group 2
da ni=9 no=240 ma=cm
km
*
1.0
.401 1.0
-.092 -.186 1.0
0 0 0 1.0
0 0 0 0 1.0
0 0 0 0 0 1.0
.523 .207 -.021 0 0 0 1.0
```

```
.304 .697 -.143 0 0 0 .174 1.0
-.053 -.165 .477 0 0 0 -.121 -.148 1.0
sd
*
.646 1.608 1.084 1.00 1.00 1.00 .874 1.813 1.327

!betas and psis specified to be invariant
model ny=9 ne=9 nk=6 ly=fu,fi te=in ps=in be=in ga=in ph=in

!first order factor model different by selection of different variables
value 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9
free te 7 7 te 8 8 te 9 9
value 0 ly 4 4 ly 5 5 ly 6 6
value 1 te 4 4 te 5 5 te 6 6

!method 1 similar in both group so we assume equality of the errors
eq te 1 1 1 te 1 1
eq te 1 2 2 te 2 2
eq te 1 3 3 te 3 3

out mi iter= 300 adm=off sc
```

Note: In the input, the effect of the true score on the observed variables involving questions not measured was fixed at 0, and the error variances of the questions not asked of a subsample were fixed at 1.0. The model, then, automatically yields correlations of zero and variances of 1.0 for the not observed "measured" variables. LISREL considers the input correlations of zero to be observed data points even though they were in fact not, so we subtracted a total of 48 degrees of freedom from the number of degrees of freedom given by LISREL to compensate for these 48 illusory correlations. More details about this approach can be found in Saris, Satorra and Coenders (2004) and in Saris and Gallhofer (2007).