Article

Choosing the Number of Categories in Agree–Disagree Scales

Sociological Methods & Research
2014, Vol 43(1) 73-97

DOI: 10.1177/0049124113509605
smr.sagepub.com

Melanie A. Revilla[1], Willem E. Saris[1], and Jon A. Krosnick[2]

Abstract

Although agree–disagree (AD) rating scales suffer from acquiescence response bias, entail enhanced cognitive burden, and yield data of lower quality, these scales remain popular with researchers due to practical considerations (e.g., ease of item preparation, speed of administration, and reduced administration costs). This article shows that *if* researchers want to use AD scales, they should offer 5 answer categories rather than 7 or 11, because the latter yield data of lower quality. This is shown using data from four multitrait-multimethod experiments implemented in the third round of the European Social Survey. The quality of items with different rating scale lengths were computed and compared.

Keywords

quality, MTMM, agree–disagree scales, number of response categories, measurement errors

[1] RECSM, Universitat Pompeu Fabra, Barcelona, Spain
[2] Stanford University, Stanford, CA, USA

Corresponding Author:
Melanie A. Revilla, Research and Expertise Center for Survey Methodology (RESCM), Universitat Pompeu Fabra, Edifici ESCI-Born, Passeig Pujades 1, 08003 Barcelona, Spain.
Email: melanie.revilla@hotmail.fr

Introduction

Although agree–disagree (AD) rating scales have been extremely popular in social science research questionnaires, they are susceptible to a host of biases and limitations. First, they are susceptible to acquiescence response bias (Krosnick 1991): Some respondents agree with the statement offered regardless of its content. For instance, if the statement is "Immigration is bad for the economy," acquiescence bias will lead to more negative opinions being expressed than if the statement is "Immigration is good for the economy." Some authors explain this tendency by people's natural disposition to be polite (e.g., Goldberg 1990); others believe that some respondents perceive the researchers to be experts and assume that if they make an assertion, it must be true (Lenski and Leggett 1960); still others attribute acquiescence to survey satisficing, a means of avoiding expending the effort needed to answer a question optimally by shortcutting the response process (Krosnick 1991). A recent study (Billiet and Davidov 2008) shows that acquiescence is quite stable over time, supporting the idea that acquiescence is a personality trait and not a circumstantial behavior.

Another drawback of AD scales is the imprecise mapping of the response dimension onto the underlying construct of interest which leads to a more complex cognitive response process.

This can be illustrated by breaking down the response process for AD scales into several steps. The classic decomposition comes from Tourangeau, Rips, and Rasinski (2000) who divide the question-answering process into four components: "comprehension of the item, retrieval of relevant information, use of that information to make required judgments, and selection and reporting of an answer." Other authors, however, propose a slightly different decomposition focused on AD scales specifically (Carpenter and Just 1975; Clark and Clark 1977; Trabasso, Rollins, and Shaughnessy 1971): comprehension of the item, identification of the underlying dimension, positioning oneself on that dimension, and selecting one of the AD response options to express that position. This last step is potentially the problematic one (Fowler 1995; Saris et al. 2010) since the translation of a respondent's opinion into one of the proposed response categories is not obvious. For example, if the statement is "Immigration is bad for the economy," and the respondent thinks that it is extremely bad, he or she may disagree with the statement, since the statement does not express his or her view. However, people may also disagree if they believe that immigration is good or very good for the economy or if they believe it is neither good nor bad (Saris and Gallhofer 2007). The AD scale may therefore mix people who hold very different

underlying opinions into the same response category. As a result, the relationship of the response scale to the underlying construct is not monotonic in terms of expressing beliefs about the impact of immigration on the economy.[1] More generally, with AD scales, people can do the mapping in their own way and this may create method effects (see e.g., Saris at al. 2010, for more details).

Despite this issue, AD scales are still used quite often, probably for practical reasons. The same scale can be used to measure a wide array of constructs, and visual display of the scale is easy on paper questionnaires or in web surveys. Administration of the questionnaire is also easier and quicker, since the scale needs only to be explained once to the respondent, whereas with Item-Specific (IS) scales, a new rating scale must be presented for each item. For these reasons, AD scales may entail lower costs (e.g., less paper needed, less work for the interviewers, less preparation cost), which is always tempting. Furthermore, the long tradition of using AD scales in the social sciences may inspire researchers to reuse established batteries of items using this response format, even if they yield lower quality data.

Given the popularity of this measurement approach, researchers must decide the number of points to offer on an AD rating scale. Likert (1932) proposed that these scales should offer five points, but Dawes (2008) recently argued that comparable results are obtained from 7- to 10-point scales, which may yield more information than a shorter scale would. Indeed, the theory of information states that if more response categories are proposed, more information about the variable of interest can be obtained: For instance, a 2-point scale only allows assessment of the direction of the attitude, whereas a 3-point scale with a middle category allows assessment of both the direction and the neutrality; even more categories can also allow assessment of the intensity, and so on (Garner 1960).

Some empirical results seem to support this theory. For instance, Alwin (1992) considers a set of hypotheses related to this theory of the information. Testing them with panel data, he finds that except for the 2-point scales, ''the reliability is generally higher for measures involving more response categories'' (p. 107). Many articles have been written discussing consequences of increasing the number of categories. However, only a limited number of studies compare the quality of scales of different lengths, where quality refers to the strength of the relationship between the observed variable and the underlying construct of interest (e.g., Andrews 1984; Scherpenzeel 1995; Költringer 1993; Alwin 1997; Alwin 2007).

In this article, we discuss the effect of the number of response categories on the quality of AD scales. These scales may behave in a specific way, because of the cognitive response process involved (which includes an extra step to map the underlying opinion onto one of the offered response categories). In one other study on this issue, Alwin and Krosnick (1991) compared 2-point and 5-point AD scales with respect to quality and found that the 2-point scales had better quality than the 5-point scales.

In our study, we compared 5-point AD scales with longer scales in terms of measurement quality. The study does not test the impact, for instance, of having only the end points labeled versus having all points labeled, nor does it test the impact of asking questions in battery style versus asking them separately. Another specificity of this study is that it involves data collected during the third round (2006–2007) of the European Social Survey (ESS) on large and representative samples in more than 20 countries.

We begin below by describing the analytical method used to assess quality. Then, we describe the ESS data analyzed using this method, the results obtained, and their implications.

Analytical Method

Our analysis involves two steps. The first step is to compute the reliability, validity, and quality coefficients of each item, using a Split-Ballot Multitrait-Multimethod design (SB-MTMM) as developed by Saris, Satorra, and Coenders (2004). The item-by-item results are then analyzed by a meta-analytic procedure to test the hypotheses of interest.

The idea to repeat several traits, measured with different methods (i.e., MTMM approach), has been proposed first by Campbell and Fiske (1959). They suggested summarizing the correlations between all the traits measured with all the methods into an MTMM matrix, which could be directly examined for convergent and discriminant validation. About a decade later, Werts and Linn (1970) and Jöreskog (1970, 1971) proposed to treat the MTMM matrix as a confirmatory factor analysis model, whereas Althauser, Heberlein, and Scott (1971) proposed a path analysis approach. Alwin (1974) presented different approaches to analyze the MTMM matrix. Andrews (1984) suggested applying this model to evaluate the reliability and validity of single-survey questions. Alternative models have been suggested (Browne 1984; Cudeck 1988; Marsh 1989; Saris and Andrews 1991). Corten et al. (2002) and Saris and Aalbers (2003) compared different models and concluded that the model discussed by Alwin (1974) and the equivalent model of Saris and Andrews (1991) fit best to several data sets.

These models have been used for substantive research by many researchers since then (Költringer 1993; Scherpenzeel 1995; Scherpenzeel and Saris 1997; Alwin 1997) and still get quite some attention (e.g., Alwin 2007; Saris and Gallhofer 2007; Saris et al. 2010).

In the classic approach, for identification reasons, each item is usually measured using at least three different methods (e.g., question wordings). However, this may lead to problems if respondents remember their answer to an earlier question when they answer a later question that measures the same construct. This problem has been studied by Van Meurs and Saris (1990).

In the study by Van Meurs and Saris (1990), several questions were repeated after different time intervals in the same questionnaire and after two weeks. The authors first determined how much agreement one can expect if there is no memory effect. This is defined as the level of agreement between the repeated observations that remains stable even if the time lag between the repeated questions is increased. Once this is determined, one can evaluate the minimal time interval between the repetitions necessary to reach the amount of agreement typical for the situation of no memory effect. Van Meurs and Saris found that:

1. People who expressed extreme opinions in the first interview always gave the same answer no matter the time interval between the repeated questions. So enlarging the time interval would not alter the apparent overtime consistency of these people's answers.

This is not surprising: These people presumably do not give the same answer because they remember their previous answer and repeat it. It is more likely that they do so because they have highly stable opinions and report them accurately.

2. If a person did not express an extreme opinion, and the questions intervening between the repeated questions were similar to the repeated question, then the observed relation was as follows:

$$C = 59.0 - .94T,$$

where C is the percentage matching answers and T is the time in minutes between the two repetitions. In this case, every extra minute in the time interval reduced the percentage of matching answers by approximately 1 percent. This means that after 25 minutes, the percentage of matching answers should be about 36 percent, which Van Meurs and Saris (1990) said is the percentage to be expected if people do not remember their previous answer.

3. If a person did not express an extreme opinion, and the questions intervening between the repeated questions were not similar to the repeated question, then the relationship was as follows:

$$C = 75.4 - .50T.$$

In this case, the extra minute of delay of the repeated question reduced memory by only half a percentage. Therefore, the level of 36 percent of matching answers would be reached after 80 minutes.

This result has been questioned by Alwin (2011), who studied memory effects by doing a word memory experiment wherein people were exposed to 10 words, and memory was tested immediately after exposure and again after 10 minutes. He concludes (Alwin 2011:282-84) that "if one looks at the delayed task and focuses solely on those words produced in response to the immediate recall task, the impression one gets is that within the context of the survey, people remember what they said earlier." This raises the need to do further research on the topic, to see whether MTMM results are distorted by memory.

Another way to limit the memory problem is to reduce the number of repetitions of the same measures in different forms. This approach, called split-ballot multitrait-multimethod approach (SB-MTMM), was developed by Saris, Satorra, and Coenders (2004). In such a design, respondents are randomly assigned to different groups, with each group receiving a different version of the same question. For example, the versions can vary in terms of the number of answer categories offered (e.g., one group receives a 5-point and a 7-pont scale; another receives a 7-point and a 11-point scale; and still another receives an 11-point and a 5-point scale). This reduces the number of repetitions: Each respondent answers only two versions of the question instead of three (Saris, Satorra, and Coenders, 2004). A memory effect is still possible, but with only two repetitions, it is less probable, also because the time between the first and the second form can be maximized.

Using this design and structural equation modeling techniques, the reliability, validity, and quality coefficients can be obtained for each question, as long as at least three different traits are measured and two methods are used to measure each trait in each group. Various models have been proposed; we use the true score model for MTMM experiments developed by Saris and Andrews (1991):

$$Y_{ij} = r_{ij}\ T_{ij} + e_{ij}. \tag{1}$$

$$T_{ij} = v_{ij}F_i + m_{ij}\ M_j, \tag{2}$$

where:

Y_{ij} is the observed variable for the i^{th} trait and the j^{th} method.
T_{ij} is the systematic component of the response Y_{ij}.
e_{ij} is the random error component associated with the measurement of Y_{ij} for the i^{th} trait and the j^{th} method.
F_i is the i^{th} trait. M_j represents the variation in scores due to the j^{th} method.
m_{ij} is the method effect for the i^{th} trait and the j^{th} method.

The model needs to be completed by some assumptions:

- The trait factors are correlated with each other.
- The random errors are *not* correlated with each other nor with the independent variables in the different equations.
- The method factors are *not* correlated with each other nor with the trait factors.
- The method effects for one specific method M_{j*} are equal for the different traits T_{ij*}.
- The method effects for one specific method M_{j*} are equal across the split-ballot groups; as are the correlations between the traits and the random errors.

Figure 1 illustrates the logic of this model in the case of two traits measured with a single method.

Working with standardized variables, we have:

- r_{ij} = reliability coefficient.
- r_{ij}^2 = reliability = $1 - \text{var}(e_{ij})$.
- v_{ij} = validity coefficient.
- v_{ij}^2 = validity.
- m_{ij} = method effect coefficient.
- m_{ij}^2 = method effect = $1 - v_{ij}^2$.

It follows that the total quality of a measure is: $q_{ij}^2 = (r_{ij} \times v_{ij})^2$. It corresponds to the variance of the observed variable Y_{ij} explained by the variable of interest F_i.

As the model in Figure 1 is not identified, it is necessary to estimate the parameters of a slightly more complicated model (one model with more traits and more methods). Figure 2 presents a simplified version of the model,

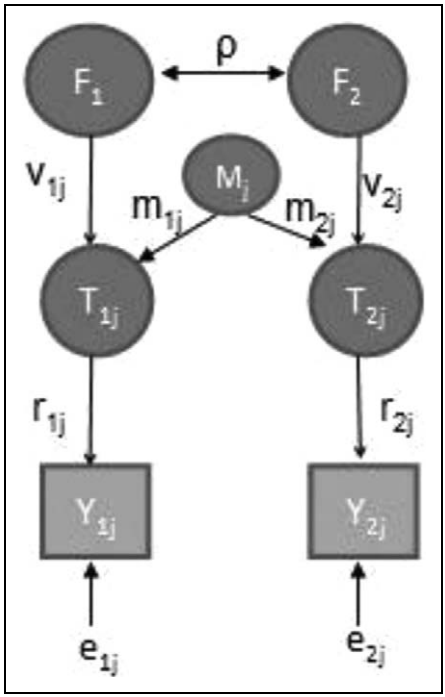

Figure 1. Illustration of the true score model.

omitting, for the sake of clarity, the observed variables, and the random errors associated with each true score.

We used the LISREL multigroup approach to estimate the model's parameters (Jöreskog and Sörbom 1991). The input instructions are shown in the Appendix (which can be found at http://smr.sagepub.com/supplemental/). The initial model was estimated for all countries and all experiments, but some adaptations for particular countries were made when misspecifications were present in the models. The main adaptations were the freeing of some of the method effects (i.e., allowing a method factor to have different impacts on different traits), and fixing a method variance at zero when its unconstrained variance was not significant and negative. All the adaptations of the initial model in the different countries and for the four different experiments (each column corresponds to an experiment) are available on the Internet.[2]

In order to determine what modifications were necessary for each model, we tested for misspecifications using the JRule software (Van der Veld,

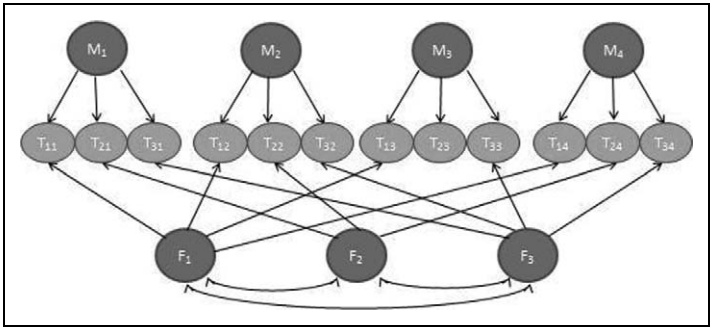

Figure 2. Illustration of an MTMM model. MTMM = multitrait-multimethod.

Saris, and Satorra 2008). This testing procedure developed by Saris, Satorra, and Van der Veld (2009) is based on an evaluation of the expected parameter changes (EPC), the modification indices (MI), and the power. The procedure thus takes into account both type I and type II errors as shown in Table 1, unlike the chi-square test, which only considers type I errors. Another advantage is that the test is done at the parameter level and not at the level of the complete model, which is helpful for making corrections (for more details about the statistical justification of our approach, see Saris, Satorra, and Van der Veld 2009).

We tried, as much as possible, to find a model that fits in the different countries (i.e., to make the same changes for one experiment in the different countries, for instance, to fix the same method effect to zero each time). Nevertheless, this was not always possible, resulting in several models specific to certain countries or groups of countries. However, the differences between the models are often limited.

Data

The ESS Round 3 MTMM Experiments

The ESS is a biannual cross-national project designed to measure social attitudes and values throughout Europe.[3] Third-round interviewing, with probability samples in 25 European countries,[4] was completed between September 2006 and April 2007. The one-hour questionnaire was administered by an interviewer in the respondent's home using show cards for most of the questions. The response rates varied from 46 percent to 73 percent

Table 1. Testing.

	Low Power	High Power
Insignificant MI	Inconclusive	No misspecification
Significant MI	Misspecification	Inspect EPC

Note. EPC = expected parameter changes; MI = modification indices.

between countries (cf. Round 3 Final Activity Report[5]). Around 50,000 individuals were interviewed.

The survey administration involved a main questionnaire and a supplementary questionnaire, in which items from the main questionnaire were repeated using different methods. Four MTMM experiments, each involving four methods and three traits, were included in the third round of the ESS. Because of the Split-Ballot design, the respondents were randomly assigned into three groups (gp A, gp B, and gp C). All groups received the same main questionnaire, but each group received a different supplementary questionnaire, which included 4 experiments with a total of 12 questions (4 experiments $\times$ 3 traits = 12 repetitions). The four experiments were:

- dngval: deals with respondents' feelings about life and relationships,
- imbgeco: deals with respondents' position toward immigration and its impact on the country,
- imsmetn: deals with respondents' opinion about immigration policies (should the government allow more immigrants to come and live in the country?),
- lrnnew: deals with respondents' openness to the future.

Table 2 gives a summary of the variables and methods used in the different Split-Ballot groups. The column "meaning" gives the statement for each variable proposed to the respondents in the AD questions. The statement may vary slightly in IS questions. The complete questionnaires are available on the ESS website.[6] The four last columns provide information about the methods used in each experiment. The column "main" refers to the method used in the main questionnaire of the ESS (M_1): It is therefore a method that all respondents receive. The next three columns indicate the second method that each Split-Ballot group received. Respondents were randomly assigned to one of these Split-Ballot groups (A, B, or C) and therefore, each person answered only one of these methods (M_2 or M_3, or M_4). It is important to notice, however, that the methods vary from one experiment to another: That

Table 2. The Split-Ballot Multitrait-Multimethod Experiments.

Experiment	Variable		Meaning	Main = M1	gpA = M2	gpB = M3	gpC = M4
Imbgeco 1	*Imbgeco*	–	It is generally bad for [country's] economy that people come to live here from other countries	11S end	5AD full	11AD end	7AD end
	imueclt	–	[Country's] cultural life is generally undermined by people coming to live here from other countries				
	imwbcnt	–	[Country] is made a worse place to live by people coming to live here from other countries				
Imsmetn 2	*Imsmet*	–	[Country] should allow more people of the same race or ethnic group as most [country's] people to come and live here.	4IS full	5AD full	4IS full	7AD end
	imdfctn	–	[Country] should allow more people of a different race or ethnic group from most [country's] people to come and live here				
	impcntr	–	[Country] should allow more people from the poorer countries outside Europe to come and live here				
Lrnnew 3	*Lrnnew*	–	I love learning new things	5AD full	5AD full	11S end	11AD end
	accdng	–	Most days I feel a sense of accomplishment from what I do				
	plprftr	–	I like planning and preparing for the future				
Dngval 4	*Dngval*	–	I generally feel that what I do in my life is valuable and worthwhile	5AD full	5AD full	5AD full	7AD end
		–	There are people in my life who really care about me				
	ppllfcr flclpla	–	I feel close to the people in my local area				

Note. "End" = only the end points of the scale are labeled; "full" = scale is fully labeled.

is why in each of the four experiments (which correspond to different rows in Table 2) we can see four distinct methods (each method corresponding to a specific scale: a 5-point AD scale, an 11-point AD scale, etc.).

In all experiments, the 5-point AD scales propose the same categories: "Agree strongly," "Agree," "Neither agree nor disagree," "Disagree," "Disagree strongly." All 5-point AD scales are fully labeled scales with the categories presented vertically, except for one case. On the contrary, all 7- and 11-point AD scales are presented as horizontal rating scales and have only the end points labeled by: "Agree strongly" and "Disagree strongly."

The ESS questionnaire never offers the option "Don't Know" as a response. The interviewer will only code an answer as "Don't Know" if a respondent independently gives this response. Therefore, there are very few such answers: usually less than 2 percent (insignificant enough to be ignored in the analysis).

This design allows comparisons to be made between both repetitions of the questions for the same respondents (e.g., using M_1 and one of the three other methods) and between Split-Ballot observations (M_2 and M_3, or M_2 and M_4, or M_3 and M_4). Since the supplementary questions are asked at the end of the interview, some time effect could play a role (positive impact on the quality if respondents learn, or negative if they become less attentive and lose motivation) and explain differences in qualities between the different measures. Nevertheless, Table 2 shows that for two of the experiments (imbgeco and imsmetn) the variations in the lengths of the scales are present only in the supplementary experiments, therefore, timing is not an issue. In the two others (dngval and lrnnew), the 5-point AD scale in the main questionnaire is repeated in one of the groups in the supplementary questionnaires, so once again, we can and will focus in the analysis only on Split-Ballot comparisons and, so, no order or time effect can explain the quality variations.

The first form of the question is presented in the beginning of the main questionnaire and its repetition is presented in the supplementary questionnaire. The main questionnaire contained approximately 240 questions. The repeated question is separated by at least 200 questions. If we assume that people answer three to four questions per minute, the time between the questions is 50 and 70 minutes. Given that many of the questions in between are rather similar and the repeated question is in general not the same in form as the first question, a memory effect seems unlikely.

Besides that, memory effects cannot explain the differences found in the measures in the supplementary questionnaires, since all groups receive the same form in the main questionnaire. Therefore, *if* a memory effect is present, it should be the same for all groups. The only possible difference that can

be anticipated is between the groups with an exact repetition and groups getting a different method the second time. In the case of the exact repetitions of the same questions in the main and the supplementary questionnaire, the quality may be higher the second time than with nonexact repetitions. This possibility would need to be kept in mind when interpreting our results.

Finally, it is noticeable that in the experiment called "dngval," a 5-point AD scale is used both in groups A and B. However, these two scales correspond to two distinct methods, because they differ at some other levels: In group A, a battery is used, whereas in group B, each question is separated from the others; in group A, the response categories are presented horizontally, whereas in group B, they are presented vertically. These differences may lead to different quality estimates.

Adaptation of the Data for Our Study

First, we had to select only the observations that could be used for our study. Hungary did not complete the supplementary questionnaire, so we could not include it. Moreover, in some countries, the supplementary questionnaire was self-completed instead of being administered by an interviewer. In that case, some people answered it on the same day as the main questionnaire, but others waited one, two, or many more days. A time effect may intervene in these circumstances, because the opinion of the respondent can change, so we did not take the individuals who answered on different days into consideration (Oberski, Saris, and Hagenaars 2007). This led us to exclude Sweden from the data, due to the fact that no one there completed both parts of the questionnaire on the same day. In the other countries, the number of ignored observations (due to completion of the supplementary questionnaire on another day) was not very high, and we still had more than 45,000 observations for our study.

We then converted these data into the correlation or covariance matrices and means needed for each group and experiment. Because we had four methods and three traits, the matrices contain 12 rows and 12 columns. However, these matrices are incomplete, due to Split-Ballot design: Only the blocs (i.e., correlations or covariances) for the specific methods that each group receives are nonzero. These matrices were obtained using ordinary Pearson correlations and the pairwise deletion option of R for missing and "Don't Know" values. Results would be different if we had corrected the categorical character of questions in the correlations calculation as indicated in Saris, van Wijk, and Scherpenzeel (1998). However, as demonstrated by Coenders and Saris (1995), the measurement quality estimates would then have meant something different. Indeed, when polychoric correlations are

used,[7] it is the measurement of the continuous underlying variable y^* that is assessed, whereas when covariances or Pearson correlations are used, it is the measurement quality of the observed ordinal variable y which is assessed. Therefore, "if the researcher is interested in measurement-quality altogether (including the effects of categorization), or in assessing the effects of categorization on measurement quality, the Pearson correlations should be used" (Coenders and Saris 1995:141). This is exactly what we want to do, so following the authors' advice, Pearson correlations have been used.

The matrices for the different experiments and countries were analyzed in LISREL in order to obtain estimates for the coefficients of interest. For details on this approach, we refer to Saris, Satorra, and Coenders (2004). The number of 12×12 matrices was 276 (for 23 countries, four experimental conditions, and three split-ballot groups).

Results

We computed the reliabilities, validities, and qualities for each method (four methods each time: M_1 to M_4), for each experiment (four experiments: "dngval," "imbgeco," "imsmetn," and "lrnnew"), each trait (three traits), and in each country (23 countries). This provided 1,104 reliability coefficients, 1,104 validity coefficients, and 1,104 quality coefficients. In order to obtain an overview, it was therefore necessary to reduce and summarize this huge amount of data.

First, we focused on the quality and not on the validity and reliability separately. Second, since we were interested in the AD scales, we kept only the observations for the AD scales when an experiment mixed methods with AD scales and methods with IS scales (cf. note 1 for a definition). Third, because of the possible time effect mentioned previously, and in order to isolate the effect of the length of the scale, we decided to focus only on comparison of the qualities of the Split-Ballot groups. Finally, we did not consider each trait separately, but computed the mean quality of the three traits. Table 3 presents the results obtained from this process.

Table 3 shows that in only a minority of cases (17 of the $92 = 18$ percent) the mean quality does not decrease when the number of points on the scale increases. In other words, the main trend (in 82 percent of the cases) is as follows: the more categories an AD scale contains, the worse its mean quality is.

In order to have a more general view of the number of points' effect on quality, we also considered the mean quality depending on the number of categories across countries. The last row of Table 3 reflects this information. The decline across countries is quite clear. For example, in the experiment

Table 3. Mean Quality for the Different Traits, Countries, and Experiments.

	imbgeco			imsmetn		lrnnew		dngval		
cntry	5AD	7AD	11AD	5AD	7AD	5AD	11AD	5AD	5AD	7AD
AT	0.51	0.33	0.39	0.54	0.44	0.64	0.46	0.59	0.63	0.40
BE	0.54	0.38	0.33	0.45	0.46	0.72	0.66	0.60	0.59	0.56
BG	0.31	0.28	0.17	0.66	0.53	0.67	0.36	0.54	0.41	0.30
CH	0.56	0.54	0.34	0.47	0.41	0.57	0.53	0.73	0.56	0.50
CY	0.50	0.40	0.50	0.52	0.54	0.68	0.58	0.61	0.50	0.35
DE	0.49	0.48	0.41	0.53	0.49	0.57	0.47	0.53	0.62	0.54
DK	0.60	0.45	0.49	0.59	0.47	0.61	0.47	0.67	0.66	0.36
EE	0.38	0.26	0.21	0.44	0.48	0.64	0.52	0.62	0.66	0.50
ES	0.51	0.31	0.23	0.55	0.51	0.68	0.66	0.64	0.59	0.41
FI	0.58	0.29	0.42	0.51	0.41	0.48	0.49	0.80	0.78	0.61
FR	0.60	0.37	0.44	0.48	0.44	0.57	0.49	0.67	0.73	0.53
GB	0.50	0.36	0.37	0.51	0.37	0.64	0.59	0.41	0.32	0.34
IE	0.37	0.18	0.08	0.35	0.40	0.56	0.33	0.40	0.33	0.27
LV	0.25	0.11	0.07	0.53	0.42	0.51	0.41	0.58	0.47	0.35
NL	0.40	0.28	0.26	0.28	0.27	0.67	0.63	0.56	0.45	0.36
NO	0.61	0.39	0.28	0.47	0.40	0.71	0.59	0.60	0.49	0.40
PL	0.34	0.19	0.14	0.47	0.50	0.67	0.54	0.62	0.52	0.52
PT	0.43	0.40	0.22	0.46	0.58	0.61	0.50	0.53	0.42	0.34
RO	0.37	0.19	0.15	0.63	0.60	0.57	0.30	0.49	0.53	0.41
RU	0.44	0.30	0.34	0.53	0.49	0.42	0.36	0.48	0.42	0.43
SI	0.37	0.18	0.11	0.50	0.41	0.66	0.57	0.46	0.41	0.28
SK	0.30	0.17	0.14	0.50	0.42	0.53	0.46	0.45	0.61	0.39
UA	0.46	0.22	0.21	0.54	0.50	0.37	0.33	0.69	0.70	0.48
All	0.45	0.31	0.27	0.50	0.46	0.60	0.49	0.58	0.54	0.42

called "imbgeco," the 5-point scale results in a 0.45 mean quality across countries, whereas with the 7-point scale it is only 0.31, and with an 11-point scale only 0.27. The same trend appears in the other three experiments.

To come back to the question of potential memory effects, studying this table, one can notice that the highest quality is found for the 5-point AD scales in the two experiments ("lrnnew" and "dngval") with exact repetitions, which is what one would expect if memory effects lead to reduced errors. However, the general trend is similar in the experiments using a 5-point AD scale in the main questionnaire and those using IS scales. The same order of quality is found for all four topics, it does not matter if there is an exact repetition or not.

In order to aggregate our findings further, we considered the mean quality across countries, experiments, and methods. This allowed us to make a distinction between reliability and validity while maintaining a clear overview.

Table 4 confirms the trend noted above and also shows that when a 7-point AD scale is chosen instead of a 5-point AD scale, the mean quality declines by 0.139. This is quite an important reduction in quality significant at 5 percent (a t test for differences in means gives a p value of .000). Moving from 7 to 11 categories also leads to a decrease of mean quality, but here it is very small (.011) and not significant at 5 percent (p value = .500). Interestingly, the difference between the 5- and 7-point scales is much larger than the difference between 7- and 11-point scales (not significant) although the difference in number of categories is smaller (two vs. four). It seems that seven response categories are already too many, and adding more does not produce any noticeable changes.

Looking at reliability and validity separately, one can see the robustness of reliability in terms of variations in the number of categories (t tests show that there are no significant differences between the three means, with p values of .93 and .66, respectively, for the test between 5- and 7-point and 7- and 11-point scales). However, validity is quite sensitive, as is quality, to the number of categories and changes: The difference in means between a 5- and a 7-point scale is quite high (0.198) and significant at 5 percent, whereas the difference between a 7- and an 11-point scale is very small (0.024) and not significant. The reduction in total quality is clearly due to the decrease in the validity. The validity is $v_{ij}^2 = 1 - m_{ij}^2$. This means that the method effects increase, as the number of categories increases, causing the observed quality loss.

Discussion and Further Research

The quality coefficients computed above show the same trends clearly appear at different levels of aggregation: On an AD scale, the quality decreases as the

Table 4. Mean Quality, Reliability, and Validity by Number of Response Categories.

No. of Points	Mean q^2	Mean r^2	Mean v^2
5	0.533	0.717	0.753
7	0.394	0.716	0.555
11	0.383	0.709	0.531

number of categories increases, so that the best AD scale is a 5-point one. This contradicts the main statement of the theory of information, which as mentioned previously, argues that more categories mean more information about the variable of interest. In terms of quality of measurement, 5-point scales yield better quality data. Our suggestion is, therefore, to use 5- and not 7-point scales.

This result is noteworthy because the choice of the number of response categories is consequently related to correlations between variables. For example, if we focus on two factors (e.g., the two first traits of the "imbgeco" experiment), as shown in Figure 1, the correlation between the observed variables is given by:

$$\rho(Y_{1j},Y_{2j}) = r_{1j}\, v_{1j}\, \rho(F_1, F_2)\, v_{2j}\, r_{2j} + r_{1j}\, m_{1j}\, m_{2j}\, r_{2j}.$$

If we assume that $r_{1j} = r_{2j}, v_{1j} = v_{2j}$ and $m_{1j} = m_{2j}$, and that the true correlation is $\rho(F_1,F_2) = 0.4$, then:

$$\rho(Y_{1j},Y_{2j}) = 0.4q^2 + r^2(1 - v^2).$$

If a survey uses a 5-point AD scale, using that scale's mean quality given in Table 4, it is expected that the correlation between the observed variables will be:

$$\rho(Y_{1,5\text{AD}},Y_{2,5\text{AD}}) = 0.4 \times 0.533 + 0.717 \times (1 - 0.753) = 0.213 + 0.177 = 0.39.$$

The first term of the sum illustrates the decrease in the observed correlation due to the relatively low quality. The second term shows the increase in observed correlation due to high method effects. However, if another survey asks the same questions but uses a 7-point AD scale, the observed correlation becomes:

$$\rho(Y_{1,7\text{AD}},Y_{2,7\text{AD}}) = 0.4 \times 0.394 + 0.716 \times (1 - 0.555) = 0.157 + 0.318 = 0.48.$$

Now the first term is even lower, since the quality is lower, whereas the second term is higher, since the method effects are higher overall, this leads to a higher observed correlation. For the 5-point scale, 0.177 of the observed

correlation is due to the method and has no substantive relevance. For the 7-point scale, this is even 0.318 which is due to the method.

This example is simplistic because only the mean quality is used. Of course, depending on the specific traits of interest and depending on the country studied, the effects might be less, or more, than those computed. However, it gives an idea of the chosen scale's importance and its possible consequences on the analysis: Depending on the method, even if the true correlation is the same, the observed correlations may be different; they might also be different from the true correlation. The decomposition of the observed correlation also demonstrates that this correlation is really unstable, because it depends on a combination of quality and method effects.

Because decrease in total quality is mainly due to decrease in validity, method effects are greater when the number of response categories is higher. This can be explained by a systematic but individual interpretation and use of AD scales: Each person uses the scales in a different way from other persons, but the same person uses the scale in the same way when answering different items. Because more variations in a personal interpretation of the scale are possible with more categories, providing a scale with more categories leads to more method effects, and hence to lower validity and lower quality.

The results are quite robust in different countries, for different experiments, and for different traits. It is therefore possible to give some general advice: Regardless of the country, regardless of the topic, and despite what the information theory states, there is no gain in information when an AD scale with more than five categories is used. There is, instead, a loss of quality. That is why if AD scales must be used, we recommend that they contain no more than five response categories.

However, this study has some limits. Even if the amount of data used is huge, the specific design of the available experiments still limits the possible analyses. There are two specific points (impossible to test in our study because the necessary data were unavailable) that we think should be examined: the first is the interest in having other numbers of categories. In the third round of the ESS, only 5-, 7-, and 11-point scales were present in the MTMM experiments. This is too limited. The 8- or 9-point scales may confirm the tendency that using more response categories does not improve the quality, but this should, nonetheless, be tested. A test of scales containing fewer categories would also be particularly interesting. Perhaps the tendency is not the same when there are very few categories. For instance, is a 2-point scale ("Disagree" vs. "Agree") better than the 5-point scale used in the ESS round 3? As we have mentioned previously, such a comparison was done by Alwin and Krosnick (1991), and they found that the 2-point scale had better quality than the 5-point scale. However, in this case, one

should consider as well that such a dichotomous scale, lacking a middle category, may lead to higher nonresponse rate. We do not know what happens if 3- or 4-point scales are used. So, further research is required for AD scales to discern what the optimal number of categories is. Since we had no data to test this, we must qualify our statement with more precision: An AD 5-point scale appears to be better than an AD 7- or 11-point scale, so employing more than five categories in an AD scale is not recommended, although, perhaps, scales with even fewer categories might result in better quality and validity.

Furthermore, in round 3 of the ESS, the 5-point scale was always completely labeled, whereas only the end points of the 7- and 11-point scales were labeled. The comparison of 7- and 11-point scales can therefore be made ceteris paribus, and as mentioned previously, shows no significant difference in the measurement's total quality. However, we cannot distinguish between the effect of the number of categories and the effect of labels in the comparison between the 5-point scale, on one hand, and the 7- and 11-point scales, on the other.

Previous research nevertheless gives us some information about the potential effect of labeling on the quality. Andrews (1984), using an MTMM approach and model, finds a negative impact of labeling: The reliability is lower for fully labeled scales compared to partially labeled ones. Alwin's (2007:87-88) MTMM studies comparing fully and partially labeled scales showed that the effect of full labeling on the quality (b_t) was negative. But Alwin (2007:200-2) also reports analyses of panel studies data using a quasi-simplex model for the estimation: There the effect of labeling is positive. Also, these analyses do not control for other elements of question design.

Saris and Gallhofer (2007) in their meta-analysis control for many other characteristics and found a positive impact of labels. When a completely labeled scale is used instead of a partially labeled scale, the reliability coefficient in general increases by 0.033, whereas the validity coefficient decreases by 0.0045. This result is in line with findings reported by Krosnick and Berent (1993).

We used Saris and Gallhofer's MTMM results and the reliability and validity found in our study for a partially labeled 7-point AD scale (cf. Table 4) in order to compute the anticipated quality for a completely labeled 7-point AD scale. The expected value of the reliability coefficient is indeed: $r_{7\text{pts, all labels}}$ = (mean reliability coefficient found in our study for a 7-point scale with only the end point labeled + increase of the reliability coefficient expected if the scale would have all points labeled, based on Saris and Gallhofer's estimate). A similar formula can be obtained for the validity coefficient. Finally, we have:

$$q^2_{7\text{pts, all labels}} = (\sqrt{0}.716 + 0.033)^2 \times (\sqrt{0}.555 - 0.0045)^2 = 0.424.$$

This is only slightly higher than the quality of the same scale before the correction ($q^2_{7\text{pts, only end pts labels}} = 0.394$), and the difference in quality from a 5-point scale remains quite large. If the estimates of the impact of labeling are correct, the difference in labels seems to explain only a minimal difference in quality. We do believe that this is the case, but to be more exact, we should qualify our statement with even more precision: A fully labeled 5-point AD scale is better than a 7- or 11-point AD scale with only the end points labeled, thus, employing more than five categories with only end points labeled in an AD scale is not recommended.

Differences between our findings and evidence elsewhere in literature about the length of the scales may be explained by our focus on AD scales. Indeed, the answering process is more complex with AD scales, because of the extra step involved in translating the position on the requested judgment in the AD categories. This last step is tricky: People can interpret the meaning of each AD category in very different ways, and when the number of categories increases, so do the possibilities of differences in interpretation. By contrast, with IS scales, it is easier for respondents to choose a response category that expresses their position. IS scales behave differently and yield data of higher quality regardless of the number of points (Saris et al. 2010). Moreover, the quality of IS scales may increase when the number of categories increases: Previous analyses (e.g., Alwin 1997 or Saris and Gallhofer 2007) documented this tendency even without differentiating between AD and IS scales. Since in our study, longer AD scales showed lower quality, the positive impact of having more response categories in IS format may be even higher than what has been found in the literature so far if a distinction was made between AD and IS scales.

The third round of the ESS focused on AD experiments and did not allow for testing of this hypothesis about IS scales. We were only able to find some experiments that varied the lengths of IS scales in the first ESS round, but not enough of them to draw conclusions. Future rounds, however, should contain such experiments, enabling a similar study of IS scales in the near future. In that case, determining how many categories are necessary to obtain the best total quality will be an interesting complement to this article. Moreover, if improved quality is substantiated by such experiments, their results will only reinforce our belief that the difference between our findings and previous research is explained by the fact that previous researchers did not control the kinds of scales they employed (AD or IS), inasmuch as these scales can generate quite different results.

Acknowledgment

We are very grateful to three anonymous reviewers for their very helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. For these and other reasons, AD scales are expected to yield more measurement error than do Item-Specific (IS) rating scales. By IS scale, we mean, following Saris et al. (2010), a scale where "the categories used to express the opinion are exactly those answers we would like to obtain for this item." For instance, we can propose the statement "immigration is good for the economy" with an AD scale: "Agree–Disagree." Alternatively, we can ask this question using an IS scale as follows: "how good or bad is immigration for the economy, very good, good, neither good nor bad, bad or very bad?" Various studies have shown that IS scales are more reliable (Scherpenzeel and Saris 1997). Saris et al. (2010) have shown that the quality of IS scales over several topics and for many countries is 20 percent higher than the quality of AD scales.
2. http://docs.google.com/Doc?id=dd72mt34_164fzsc8qhr. See also note 4 for the list of countries' names and their abbreviations.
3. http://www.europeansocialsurvey.org/
4. Austria = AT, Belgium = BE, Bulgaria = BG, Switzerland = CH, Cyprus = CY, Germany = DE, Denmark = DK, Estonia = EE, Spain = ES, Finland = FI, France = FR, United Kingdom = GB, Hungary = HU, Ireland = IE, Latvia = LV, Netherlands = NL, Norway = NO, Poland = PL, Portugal = PT, Romania = RO, Russia = RU, Sweden = SE, Slovenia = SI, Slovakia = SK, Ukraine = UA
5. Available on the ESS website: http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=101&Itemid=139
6. http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=63&Itemid=98 for the main questionnaire and for the supplementary questionnaires: http://www.europeansocialsurvey.org/index.php? option=com_content&view=article&id=65&Itemid=107
7. The use of the polychoric correlations also assumes that the latent variables behind the observed variables have a multivariate normal distribution which seems rather unlikely for many social sciences variables, while the power of the test for this assumption is extremely low (Quiroga 1992). Winship and Mare (1984) suggest an alternative test but do not indicate the power of this test.

References

Althauser, Robert P., Thomas A. Heberlein, and Robert A. Scott. 1971"A Causal Assessment of Validity: The Augmented Multitrait-Multimethod Matrix." Pp. 374-99 in *Causal Models in the Social Sciences*, edited by H. M. Blalock Jr. Chicago, IL: Aldine.

Alwin, Duane F. 1974. "Approaches to the Interpretation of Relationships in the Mutlitrait-Multimethod Matrix." Pp. 79-105 in *Sociological Methodology 1973-74*, edited by H. L. Costner. San Francisco, CA: Jossey-Bass.

Alwin, Duane F. 1992. "Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement." Pp. 83-118 in *Sociological Methodology*, Vol. 22, edited by Peter V. Marsden. Washington, DC: American Sociological Association.

Alwin, Duane F. 1997. "Feeling Thermometers versus 7-point Scales: Which Are Better?" *Sociological Methods and Research* 25:318.

Alwin, Duane F. 2007. *Margins of Errors: A Study of Reliability in Survey Measurement*. Wiley-Interscience. Hoboken, New Jersey: Wiley and Sons, Inc.

Alwin, Duane. F. 2011 "Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach." Pp. 265-95 in *Question Evaluation Methods*, edited by Jennifer Madans, Kristen Miller, Aaron Maitland, and Gordon Willis. John Wiley. Hoboken, New Jersey: Wiley and Sons, Inc.

Alwin, Duane. F. and Jon A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement." *Sociological Methods and Research* 20:139-81.

Andrews, Frank. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 46:409-42 Reprinted inW. E. Saris and A. van Meurs. 1990. *Evaluation of Measurement Instruments by Metaanalysis of Multitrait Multimethod Studies*. Amsterdam, the Netherland: North-Holland.

Billiet, Jaak B. and Eldad Davidov. 2008. "Testing the Stability of an Acquiescence Style Factor Behind Two Interrelated Substantive Variables in a Panel Design." *Sociological Methods and Research* 36:542-62.

Browne, Michael W. 1984. "The Decomposition of Multitraitmultimethod Matrices." *British Journal of Mathematical and Statistical Psychology* 37:1-21.

Campbell, Donald T. and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 6: 81-105.

Carpenter, Patricia A. and Marcel A. Just. 1975. "Sentence Comprehension: A Psycholinguistic Processing Model of Verification." *Psychological Review* 82:45-73.

Clark, Herbert H. and Eve V. Clark. 1977. *Psychology and Language*. New York: Harcourt Brace.

Coenders, Germà and Willem E. Saris. 1995. "Categorization and Measurement Quality. The Choice between Pearson and Polychoric Correlations." Pp. 125-144 in *The MTMM Approach to Evaluate Measurement Instruments*, Chapter 7, edited by W. E. Saris. Budapest: Eötvös University Press.

Corten, Irmgard W., Willem E. Saris, Germà M. Coenders, William M. van der Veld, Chris E. Aalberts, and Charles Kornelis. 2002. "Fit of Different Models for Multitrait-Multimethod Experiments." *Structural Equation Modeling* 9: 213-32.

Cudeck, Robert. 1988. "Multiplicative Models and MTMM Matrices." *Journal of Educational Statistics* 13:131-47.

Dawes, John. 2008. "Do Data Characteristics Change According to the Number of Points Used? An Experiment Using 5-point, 7-point and 10-point Scales." *International Journal of Market Research* 50:61-77.

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

Garner, Wendell R. 1960. "Rating Scales, Discriminability, and Information Transmission." *Psychological Review* 67:343-52.

Goldberg, Lewis R. 1990. "An Alternative 'Description of Personality': The Big-Five Factor Structure." *Journal of Personality and Social Psychology* 59:1216-29.

Jöreskog, Karl G.. 1970. "A General Method for the Analysis of Covariance Structures." *Biometrika* 57:239-51.

Jöreskog, Karl G. 1971. "Statistical Analysis of Sets of Congeneric Tests." *Psychometrika* 36:109-33.

Jöreskog, Karl G. and Dag Sörbom. 1991. *LISREL VII: A Guide to the Program and Applications*. Chicago: SPSS.

Költringer, Richard. 1993. Messqualität in der sozialwissenschaftlichen Umfrageforschung. Endbericht Project P8690-SOZ des Fonds zur Förderung der wissenschaftlichen Forschung (FWF), Wien, Austria.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213-36.

Krosnick, Jon A. and Matthew K. Berent. 1993. "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format." *American Journal of Political Science* 37:941-64.

Lenski, Gerhard E. and John C. Leggett. 1960. "Caste, Class, and Deference in the Research Interview." *American Journal of Sociology* 65:463-67.

Likert, Rensis. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 140:1-55.

Marsh, Herbert W. 1989. "Confirmatory Factor Analyses of Multitrait-Multimethod Data: Many Problems and a Few Solutions." *Applied Psychological Measurement* 13:335-61.

Oberski, Daniel, Willem E. Saris, and Jacques Hagenaars. 2007. "Why Are There Differences in the Quality of Questions across Countries?" Pp. 281-299 in *Measuring Meaningful Data in Social Research*, edited by Geert Loosveldt, Marc Swyngedouw, and Bart Cambre. Leuven, Belgium: Acco.

Quiroga, Ana M. 1992. *Studies of the Polychoric Correlation and Other Correlation Measures for Ordinal Variables*. PhD thesis, Uppsala, Sweden.

Saris, Willem E. and Aalberts Chris. 2003. "Different Explanations for Correlated Disturbance Terms in MTMM Studies." *Structural Equation Modeling: A Multidisciplinary Journal* 10:193-213.

Saris, Willem E. and Frank M. Andrews. 1991. "Evaluation of Measurement Instruments Using a Structural Modeling Approach." Pp. 575-97 in *Measurement Errors in Surveys*, edited by Paul P. Biemer, Robert M. Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman. New York: John Wiley.

Saris, Willem E. and Irmtraud Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: John Wiley.

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

Saris, Willem E., Albert Satorra, and Germa Coenders. 2004. "A New Approach to Evaluating the Quality of Measurement Instruments: The Split-ballot MTMM Design." *Sociological Methodology.*

Saris, Willem E., Albert Satorra, and William M. Van der Veld. 2009. "Testing Structural Equation Models or Detection of Misspecifications?" *Structural Equation Modeling: A Multidisciplinary Journal* 34:311-347.

Saris, Willem E., Theresia van Wijk, and Annette C. Scherpenzeel. 1998. "Validity and Reliability of Subjective Social Indicators: The Effect of Different Measures of Association." *Social Indicators Research* 45:173-99.

Scherpenzeel, Annette C. 1995. *A Question of Quality: Evaluating Survey Questions by Multitrait-Multimethod Studies*. Amsterdam, the Netherlands: Nimmo.

Scherpenzeel, Annette C. and Willem E. Saris. 1997. "The Validity and Reliability of Survey Questions. A Meta-analysis of MTMM Studies." *Sociological Methods & Research* 25:341-83.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.

Trabasso, Tom, Howard Rollins, and Edward Shaughnessey. 1971. "Storage and Verification Stages in Processing Concepts." *Cognitive Psychology* 2:239-89.

Van der, Veld, William M., Willem E. Saris, and Albert Satorra. 2008. *Judgment Aid Rule Software*. Jrule 2.0: User manual (Unpublished Manuscript, Internal Report). Radboud University Nijmegen, the Netherlands.

Van Meurs, Lex and Willem E. Saris. 1990. "Memory Effects in MTMM Studies." Pp. 134-146 in *Evaluation of Measurement Instruments by Meta-analysis of Multitrait-Multimethod Studies*, edited by E. Saris Willem and Lex van Meurs. Amsterdam, the Netherlands: North Holland.

Werts, Charles E. and Robert L. Linn. 1970. "Path Analysis: Psychological Examples." *Psychological Bulletin* 74:194-212.

Winship, Christopher and Robert D. Mare. 1984. "Regressions Models with Ordinal Variables." *American Sociological Review* 49:512-25.

Author Biographies

Melanie A. Revilla is a postdoctoral researcher at the Research and Expertise Centre for Survey Methodology (RECSM) and an associate professor at Universitat Pompeu Fabra (UPF, Barcelona, Spain). She received her PhD from Universitat Pompeu Fabra in 2012, in the areas of statistics and survey methodology, under the supervision of professors Willem Saris (UPF) and Peter Lynn (Essex University). Her dissertation dealt with the effects of different modes of data collection on the quality of survey questions. She is interested in all aspects of survey methodology.

Willem E. Saris is Professor and researcher at the Research and Expertise Centre for Survey Methodology (RECSM) since 2009. In 2005, he was laureate of the Descartes Research Prize for the best scientific collaborative research. In 2009, he received the Helen Dinerman award from the World Association of Public Opinion Research (WAPOR), in recognition to his lifelong contributions to the methodology of public opinion research. In 2011 he received the degree of Doctor Honoris Causa from the University of Debrecen in Hungary. More recently, he was awarded the "2013 Outstanding Service Prize" by the European Survey Research Association.

Jon A. Krosnick conducts research in three primary areas: (1) attitude formation, change, and effects, (2) the psychology of political behavior, and (3) the optimal design of questionnaires used for laboratory experiments and surveys, and survey research methodology more generally. He is the Frederic O. Glover Professor in Humanities and Social Sciences, Professor of Communication, Political Science, and (by courtesy) Psychology. At Stanford, in addition to his professorships, he directs the Political Psychology Research Group and the Summer Institute in Political Psychology. He is the author of four books and more than 140 articles and chapters.