UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 5

2009

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

CHOOSING THE NUMBER OF CATEGORIES IN AGREE-DISAGREE SCALES

Melanie Revilla

Willem E. Saris

Universitat Pompeu Fabra

Jon A. Krosnick

Stanford University

Abstract

Although agree-disagree (AD) rating scales suffer from acquiescence response bias, entail enhanced cognitive burden, and yield data of lower quality (Krosnick, 1991; Saris, Revilla, Krosnick, Schaeffer, forthcoming), these scales remain popular with researchers due to practical considerations (e.g., ease of item preparation, speed of administration, reduced administration costs). This paper 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 (MTMM) experiments implemented in the third round of the European Social Survey. The quality of items with different rating scale lengths were computed and compared.

1) 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 the potentially misleading one (Fowler, 1995; Saris, Revilla, Krosnick and Shaeffer, 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 in Web surveys or on paper questionnaires. Administration of the questionnaire is also easier and quicker, since the scale needs only to be explained

[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 is immigration for the economy, good or bad?" Various studies have shown that IS scales are more reliable (Scherpenzeel and Saris, 1997). Saris et al. (2010) have shown that over several topics and for many countries, item-specific scales were of 20% higher quality than AD scales.

once to the respondent, whereas with 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 re-use 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 5-points, but Dawes (2008) recently argued that comparable results are obtained from 7- and 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, etc. (Garner, 1960). Dawes (2008) compared different length scales in terms of the mean, variance, skewness and kurtosis of observed distributions. Few studies have been done on comparing 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 (Andrews, 1984; Scherpenzeel, 1995; Költringer, 1993; Alwin, 2007). Moreover, these studies do not focus on AD scales[2]. We believe that these scales behave in a very specific way, because of their specific cognitive response process (with an extra step to express the opinion into one of the response categories).

[2] They consider IS scales as well as AD scales.

We report, therefore, the findings of a study which compares 5-point AD scales with longer scales in terms of measurement quality. The study focuses on AD scales and the impact of the number of categories on measurement quality, but it 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 the method, the results obtained and their implications.

2) 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. This Structural Equation Models approach has been shown to be much more powerful than the direct examination of Campbell and Fiske and the path analysis approach of Althauser and Herberlein (Alwin, 1974). Andrews (1984) suggested applying this approach and model to evaluate the reliability and validity of single survey questions. This approach has been discussed (Browne, 1984; Cudeck, 1988; Marsh, 1989; Saris and Andrews, 1991) and used for substantive research by many researchers since then (Költringer, 1993; Scherpenzeel, 1995; Scherpenzeel and Saris, 1997; Alwin, 1997; Corten et al. 2002; Saris and Aalbers, 2003) and still gets quite some attention nowadays (e.g. Alwin, 2007; Saris and Gallhofer, 2007; Saris, Revilla, Krosnick and Schaeffer, 2010).

In the classic approach, for identification issues, each construct is usually measured for each respondent 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 measuring a construct when they answer a later question which measures that same construct. Van Meurs and Saris (1990) found that if the administration of the two questions is separated by more than 20 minutes, memory of the earlier answer is minimal. Nonetheless, considerable questionnaire administration time is required in order to apply three different methods to the same respondents. That is why it is preferable to use a Split-Ballot MTMM design.

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 scale; another receives a 7-point scale; and still another receives an 11-point scale). This reduces the number of repetitions: each respondent answers only two versions of the question instead of three (Saris et al. 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. In our study, approximately one hour separates them, so very few memory effects are expected.

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):

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \qquad (1)$$

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \qquad (2)$$

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.

****

Figure 1: Basic measurement model

****

Working with standardized variables, we have:

- r_{ij} = reliability coefficient
- r_{ij}^2 = reliability = 1 – 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: $q_{ij}^2 = (r_{ij}.v_{ij})^2$. It corresponds to the variance of the observed variable Y_{ij} explained by the variable of interest F_i.

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

****

Figure 2: The MTMM model

****

We used the LISREL multi-group approach to estimate the model's parameters (Jöreskog and Sörbom, 1991). The input instructions are shown in Appendix A. The initial model was estimated for all countries and all experiments, but some adaptations for particular countries were made in order to avoid the effects of misspecifications 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. Since they were non significant, we do not expect a huge effect on the results when fixing them, and thus these adjustments should have little, if any, effect on comparisons across countries. 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[3].

[3] http://docs.google.com/Doc?id=dd72mt34_164fzsc8qhr See also footnote 5 for the list of countries' names and their abbreviations

In order to determine what modifications were necessary for each model, we tested for misspecifications using the JRule software (Van der Veld, 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 (for more details about the statistical justification of our approach, see Saris, Satorra and Van der Veld, 2009). 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.

****

Table 1 over here

****

We tried, as much as possible, to find a model which 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.

3) Data

The ESS round 3 MTMM experiments

The European Social Survey (ESS) is a biannual cross-national project designed to measure social attitudes and values throughout Europe[4]. Third-round interviewing, with probability samples in 25 European countries[5], 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% to 73% between countries (c.f. Round 3 Final Activity Report[6]). 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, gp C). All groups received the same main questionnaire, but each group received a different supplementary questionnaire, which included four experiments with a total of twelve questions (4 experiments * 3 traits = 12 repetitions). The four experiments were:

- dngval: deals with respondents' feelings about life and relationships

[4] http://www.europeansocialsurvey.org/

[5] 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

[6] Available on the ESS website: http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=101&Itemid=139

- 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[7]. 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 (M1): 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 (M2 or M3 or M4). It is important to notice, however, that the methods vary from one experiment to another: that 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).

****

Table 2 over here

****

[7] 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

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 (see below). 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% (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 M1 and one of the three other methods) and between Split-Ballot observations (M2 and M3, or M2 and M4, or M3 and M4). 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 repetitions of 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.

Since much more than 20 minutes separate the first form of the question in the main questionnaire and its repetition in the supplementary questionnaire there should not be any reason to expect memory effects (Van Meurs and Saris 1990). Besides that, memory effect 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 non exact 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, 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 this 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 twelve rows and twelve 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 non-zero. 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, 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, p.141). This is exactly what we want to do, so following the authors' advice, Pearson correlations have been used.

Besides that, the person correlations or ordinary regressions are still the most commonly used statistical procedures used. Therefore it is important that the users of these statistical procedures know what the size of the combination of measurement errors and categorization errors on the relationships between the observed and the latent variables is. This may lead them to use latent variable models because in this way it is possible in most cases to correct for both errors, as was shown in the same paper.

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).

4) Results

We computed the reliabilities, validities and qualities for each method (four methods each time: M1 to M4), for each experiment (four experiments: "dngval", 'imbgeco", "imsmetn'," 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 only kept the observations for the AD scales when an experiment mixed methods with AD scales and methods with Item Specific (IS) scales (cf. footnote 1 for a definition). Third, because of the possible time effect mentioned above, 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 over here

****

Table 3 shows that in only a minority of cases (17 out of 92 = 18%) the mean quality does not decrease when the number of points on the scale increases. In other words, the main trend (in 82% of the cases) is: 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 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-points 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 over here

****

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% (a t-test for differences in means gives a p-value of 0.000). Moving from seven to eleven categories also leads to a decrease of mean quality, but here it is very small (0.011) and not significant at 5% (p-value = 0.500). 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 versus 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 0.93 and 0.66 respectively for the test between 5- and 7- points and 7- and 11-points). 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%, 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: $v_{ij}^2 = 1\text{-}m_{ij}^2$. This means that the method effects increase as the number of categories increases, causing the observed quality loss.

5) 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 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 above, 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:

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

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:

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

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:

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

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:

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

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 instable, 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 was 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. 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 be particularly interesting. Indeed, perhaps the tendency is not the same when there are very few categories. For instance, is a 2-point scale ("Disagree" versus "Agree") better than the 5-point scale used in the ESS round 3? Such a dichotomous scale, lacking a middle category, may lead to a higher non-response rate. Having too few categories is perhaps not an optimal situation either. Since we had no data to test this, we must qualify our statement with more precision: an AD 5-point scale is better than an AD 7- or 11-point scale, thus, 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 is always completely labeled, whereas only the end points of the 7- and 11-point scales are labeled. The comparison of 7- and 11-point scales can therefore be made *ceteris paribus*, and as mentioned above, 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.

Some previous work suggests that completely labeled scales have higher reliability. Alwin and Krosnick (1991) report that the mean reliability for 7-point scales with only the end points labeled is lower than the mean reliability for 7-point fully labeled scales (table 4, p. 167). This study is based on panel data analyzed with a quasi simplex model.

Their results and others based on the same studies (Krosnick and Fabrigar, 1997; Alwin, 2007) suggest that the higher difference in quality that we find in our analyses between on the one hand 5- and on the other hand 7- and 11-point scales could come from a combined effect of different numbers of categories and a different use of labels. All the difference may even comes from the fact that the 5-point scales are fully labeled whereas the 7- and 11-points scales have only their endpoints labeled.

However, Alwin and Krosnick's result is not based on AD scales, which we argue behave differently. Besides, the previously mentioned studies can only provide estimates of the reliability and cannot estimate the validity. When, in 1997, Alwin made this distinction between reliability and validity using the MTMM approach, he does not find this effect of labeling: instead, he reports no differences between fully and partially labeled 7-point scales.

Andrews (1984), using again an MTMM approach and model, finds even a negative impact of labeling: the reliability is lower for fully labeled scales compared to partially labeled ones.

Saris and Gallhofer (2007) also use an MTMM approach. In their meta-analysis, they detect a positive impact of labels, but their result is that 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.

Focusing on studies based on MTMM analyses, Saris and Gallhofer's result is the one that could the most mitigate our results about the effect of the number of response categories on the quality. Therefore, we use their estimates 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: $r_{7pts,all\ labels}$ = (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:

$$q^2_{7pts,all\ labels} = (\sqrt{0.716} + 0.033)^2 * (\sqrt{0.555} - 0.0045)^2 = 0.424.$$

This is only slightly higher than the quality of the same scale before the correction ($q^2_{7pts,\ only\ end\ pts\ labels} = 0.394$), 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.

Saris and Gallhofer also highlight many other aspects which may have an impact on quality which we have not considered in this paper. However, most of them were so similar for all scales that they do not have to be considered.

In conclusion, we must emphasize once again that this paper only focuses on AD scales. This is very important, because the difference between our findings and evidence which can be found 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 which expresses their position. IS scales behave differently and yield data of higher quality regardless of the number of points (Saris et al, 2010). Moreover, we believe that the quality of IS scales usually increases when the number of categories increases: previous analyses (e.g. Alwin, 1997 or Saris and Gallhofer, 2007) confirm this tendency, even without differentiating between AD and IS scales. Since in our study longer AD scales show lower quality, it suggests that the positive impact of having more response categories in IS scales would even be 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 which varied the lengths of IS scales in the first ESS round, but not enough of them to draw conclusions. Future round, 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 paper. 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.

REFERENCES

Althauser, Robert P., Heberlein, Thomas A., Scott, Robert A. (1971). "A causal assessment of validity: the augmented multitrait-multimethod matrix". In *Causal Models in the Social Sciences,* ed. H. M. Blalock Jr., pp. 374-99. Chicago: Aldine

Alwin, Duane F. (1974). "Approaches to the interpretation of relationships in the mutlitrait-multimethod matrix." In H.L. Costner (ed.), *Sociological Methodology* 1973-74. San Francisco: Jossey-Bass.

Alwin, Duane F. (1997). "Feeling Thermometers Versus 7-Point Scales: Which are Better?" *Sociological Methods and Research* 25 (3):318.

Alwin, Duane F. (2007). *Margins of errors: a study of reliability in survey measurement*. Wiley-Interscience

Alwin, Duane. F., and Jon A. Krosnick (1991). "The Reliability of Survey Attitude Measurement". *Sociological Methods and Research* 20 (1):139-181.

Andrews, Frank. (1984). "Construct validity and error components of survey measures: A structural modeling approach." *Public Opinion Quarterly*, 46, 409-42. Reprinted in W.E. Saris & A. van Meurs. (1990). *Evaluation of measurement instruments by metaanalysis of multitrait multimethod studies*. Amsterdam: North-Holland

Billiet, Jaak and Davidov, Eldad (2008). "Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design". *Sociological Methods and Research*, 36(4), 542-562.

Browne, Michael W. (1984). "The decomposition of multitraitmultimethod matrices". *British Journal of Mathematical and Statistical Psychology*, 37, 1-21.

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

Carpenter, Patricia A. and Just, Marcel A. (1975). "Sentence comprehension: A psycholinguistic processing model of verification." *Psychological Review, 82,* 45-73.

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

Coenders, Germà, and Saris, Willem, E. (1995). "Categorization and measurement quality. The choice between Pearson and Polychoric correlations". In W.E. Saris, *The MTMM approach to evaluate measurement instruments* (1995), Chapter 7, 125-144.

Corten, Irmgard W., Saris, Willem E., Coenders, Germà M., van der Veld, William, Aalberts, Chris E., and Kornelis, Charles (2002). "Fit of different models for multitrait-multimethod experiments". *Structural Equation Modeling*, 9(2), 213-232.

Cudeck, Robert. (1988). "Multiplicative Models and MTMM Matrices." *Journal of Educational Statistics* 13 (2):131-147.

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*, Vol. 38, p56-57.

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,* 1216-1229.

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, 109-133.

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

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

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

Krosnick, Jon A., and Leandre R. Fabrigar (1997). "Designing rating scales for effective measurement in surveys". In *Survey Measurement and Process Quality*, eds. L. Lyberg, P.P. Biemer, M. Collins, C. Dippo, E. de Leeuw, N. Schwarz, and D. Trewin, New York : John Wiley & Sons, Inc.

Marsh, Herbert W. (1989). "Confirmatory factor analyses of multitrait-multimethod data: many problems and a few solutions." *Applied Psychological Measurement*, 13, 335-361.

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

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

Oberski, Daniel, Saris, Willem E., and Hagenaars, Jacques (2007). "Why are there differences in the quality of questions across countries?" In: Geert Loosveldt, Marc Swyngedouw and Bart Cambre (eds.), "Measuring meaningful data in social research". Leuven: acco.

Saris, Willem E., and Chris Aalberts (2003). "Different Explanations for Correlated Disturbance Terms in MTMM Studies." *Structural Equation Modeling: A Multidisciplinary Journal* 10 (2):193-213.

Saris, Willem E. and Andrews, Frank M. (1991). "Evaluation of measurement instruments using a structural modeling approach". In Paul P. Biemer, Robert M. Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman (Eds.), *Measurement errors in surveys* (pp. 575-597). New York: Wiley.

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

Saris, Willem E., Revilla, Melanie, Krosnick, Jon A., Shaeffer, Eric M. (2010). "Comparing Questions with Agree/Disagree Response Options to Questions with Construct-Specific Response Options" *Survey Research Methods* Vol.4, No.1, pp. 61-79

Saris, Willem E., Satorra, Albert and Coenders, Germa (2004). "A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design." *Sociological Methodology 2004*

Saris, Willem E, Satorra, Albert, Van der Veld, William M. (2009). "Testing Structural Equation Models or Detection of Misspecifications?" *Structural equation modeling: A multidisciplinary Journal*, Vol. 16, Issue 4, pp 561-582

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, 173-199

Scherpenzeel, Annette C. (1995). "A question of quality: Evaluating survey questions by multitrait-multimethod studies". Amsterdam, Nimmo.

Scherpenzeel, Annette C., Saris, Willem E. (1997). "The Validity and Reliability of Survey Questions. A Meta-Analysis of MTMM Studies." *Sociological Methods & Research*, Vol. 25 No. 3, February 1997 341-383

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

Trabasso, Tom, Rollins, Howard and Shaughnessey, Edward (1971). Storage and verification stages in processing concepts. *Cognitive Psychology, 2,* 239-289.

Van der Veld, William M., Saris, Willem E., Satorra, Albert (2008) Judgment Aid Rule Software

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

Werts, Charles E., and Linn, Robert L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, 74, 194-212.

APPENDIX

Appendix A: The basic Lisrel input for analysis of the SB-MTMM

```
Analysis of [country] round 3 [experiment] group 1
Data ng=3  ni=12 no=884 ma=cm
cm file=sb-group-1.cov
mean file=sb-group-1.mean

model ny=12 ne=12 nk=7  ly=fu,fi te=sy,fi ps=sy,fi be=fu,fi ga=fu,fi ph=sy,fi
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 1 te 7 7 te 8 8 te 9 9 te 10 10 te 11 11 te 12 12
value 0 ly 7 7 ly 8 8 ly 9 9 ly 10 10 ly 11 11 ly 12 12
fr 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 ga 10 1 ga 11 2 ga 12 3
va 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 ga 10 7 ga 11 7 ga 12 7
fr ph 2 1 ph 3 1 ph 3 2  ph 5 5 ph 6 6 ph 7 7
va 1 ph 1 1 ph 2 2 ph 3 3
start .5 all
out mi iter= 200 adm=off sc

Analysis of group 2
Data  ni=12 no=744 ma=cm
cm file=sb-group-2.cov
mean file=sb-group-2.mean

model ny=12 ne=12 nk=7 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 1 1 te 2 2 te 3 3 te 7 7 te 8 8 te 9 9
value 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9
equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
value 1 te 4 4 te 5 5 te 6 6 te 10 10 te 11 11 te 12 12
value 0 ly 4 4 ly 5 5 ly 6 6 ly 10 10 ly 11 11 ly 12 12
out iter= 200 adm=off sc

Analysis of group 3
Data ni=12 no=777 ma=cm
cm file=sb-group-3.cov
mean file=sb-group-3.mean

model ny=12 ne=12 nk=7 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 1 1 te 2 2 te 3 3 te 10 10 te 11 11 te 12 12
value 1 ly 1 1 ly 2 2 ly 3 3 ly 10 10 ly 11 11 ly 12 12
equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
value 1 te 4 4 te 5 5 te 6 6 te 7 7 te 8 8 te 9 9
value 0 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9
pd
out mi iter= 200 adm=off sc
```

Tables

Table 1: Testing

Table 1. Testing	Low Power	High Power
Insignificant MI	Inconclusive	No misspecification
Significant MI	Misspecification	Inspect EPC

Table 2: the Split-ballot Multitrait-Multimethod experiments

Table 2. The SB-MTMM experiments

Expt.	Var.	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.	11IS 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	11IS 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
	ppllfcr	- There are people in my life who really care about me.				
	flclpla	- I feel close to the people in my local area.				

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

Table 3: mean quality for the different traits, countries, experiments

Table 3. Mean quality for the 3 traits in each country for each experiment										
cntry	imbgeco			imsmetn		lrnnew		dngval		
	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

Table 4: mean quality, reliability, validity by number of response categories

Table 4. Mean quality, reliability, validity by number of points			
No 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

Figures

Figure 1: illustration of the True Score model

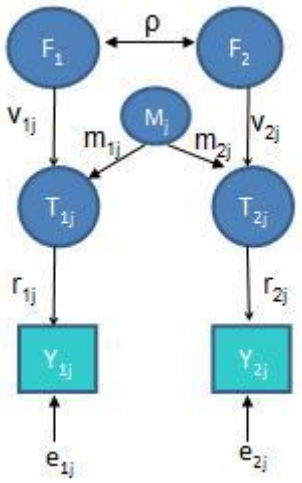

Figure 2: illustration of a MTMM model