Comparing Questions with Agree/Disagree Response Options

to Questions with Construct-Specific Response Options

Willem E. Saris

Faculty of Political Social Cultural Sciences
University of Amsterdam

Jon A. Krosnick

Departments of Psychology and Political Science
Stanford University

Eric M. Shaeffer

Department of Psychology
Ohio State University

November, 2005

Willem Saris is a Fellow at the Netherlands Institute for Advanced Study in the Humanities and Social Sciences. Jon Krosnick is University Fellow at Resources for the Future. Address correspondence to Willem E. Saris, University of Amsterdam, Methodology Department, OZ Achterburgwal 237, 1012 DL Amsterdam, the Netherlands (email: wsaris@planet.nl), or Jon A. Krosnick, Stanford University, 432 McClatchy Hall, 450 Serra Mall, Stanford, California 94305 (email: krosnick@stanford.edu). We are very grateful to Albert Satorra and Germà Coenders for their helpful comments on specifying some of tests done in this paper.

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

Abstract

A huge body of research conducted during more than five decades has documented the role that acquiescence response bias plays in distorting answers to agree/disagree questions, which might be taken as a reason to prefer questions with construct-specific response options. But remarkably little research has explored whether responses to agree/disagree questions are indeed of lower quality than responses to questions with construct-specific response options. Using two research designs that combine the advantages of a random-assignment split-ballot experiment and the multitrait-multimethod approach in the context of representative sample surveys, we found that responses to agree/disagree questions were indeed less reliable than responses to comparable questions offering construct-specific response options. Furthermore, agree/disagree questions with "negation" stems (including the word "not") were found to have lower reliability than agree/disagree questions without negation. These results attest both to the superiority of questions with construct-specific response options and to the value of the new analytic method proposed here for efficiently comparing the reliability and validity of different question forms.

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

Introduction

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

One reason for the popularity of agree/disagree response alternatives is that they seem to offer the opportunity to measure just about any construct relatively efficiently. Alternative question design approaches require that response alternatives be tailored to each item's particular construct. For example, a questionnaire might ask the following series of three questions:

1. "How would you rate your health overall: excellent, very good, good, fair, or poor?"

2. “How important is the issue of abortion to you personally? Is it extremely important, very important, somewhat important, not too important, or not important at all?”

3. “How often do you feel sad? Constantly, very often, somewhat often, rarely, or never?”

These might be called questions with construct-specific response options, because each question offers a unique set of response options specifically addressing points along the continuum of the construct of interest.

However, the same questions can also be asked in an agree/disagree format:

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

1. First, ‘My overall health is excellent.’

2. Second, ‘The issue of abortion is very important to me personally.’

3. Third, ‘I rarely feel sad.’”

Because the response scale is the same for each question, the questionnaire can present the scale only once, thereby saving some time and streamlining questionnaire administration.

However, a great deal of research points to a potential danger inherent in this response format: acquiescence response bias. More than one hundred studies using a wide variety of methods have demonstrated that some respondents are inclined to agree with just about any assertion, regardless of its content (for a review, see Krosnick and Fabrigar, forthcoming). Three different theoretical accounts for this phenomenon have been proposed, and all of them enjoy some empirical support. The first argues that acquiescence results from a personality disposition some people have to be polite and to avoid social friction, leading them to be especially

agreeable (Costa and McCrae 1988; Goldberg 1990; Leech 1983). The second explanation argues that agree/disagree formats unintentionally suggest to some respondents that interviewers and/or researchers believe the statements offered in such items, and some respondents who perceive themselves to be of lower social status than the interviewer or researcher may choose to defer to their apparent expertise and endorse their apparent beliefs (Carr 1971; Lenski and Leggett 1960; Richardson, Dohrenwend, and Klein 1965). Finally, the theory of survey satisficing argues that a general bias in hypothesis testing toward confirmation rather than disconfirmation inclines some respondents who shortcut the response process toward agreeing with assertions presented to them in agree/disagree questions (Krosnick 1991).

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

However, this sort of reasoning is not necessarily sensible. Perhaps people are inclined to acquiesce most when they lack a true opinion on the topic of a question. Therefore, if the item were presented in a different format, these respondents would still answer in a meaningless way. For example, perhaps people would flip a mental coin and answer randomly (Converse 1964), which would scatter these individuals across the set of offered response alternatives. The agree/disagree format might therefore offer the advantage of inclining these people toward a

particular response (i.e., “agree”) rather than scattering them randomly. As a result, balancing a battery with oppositely worded items would permit researchers to identify acquiescers and adjust research conclusions correcting for their impact.

Our goal in this paper is to explore these issues by comparing questions with agree/disagree response options to questions with construct-specific response options. In addition to the potentially damaging effects of acquiescence, there are other reasons to suspect that answers to agree/disagree questions provided by non-acquiescing respondents may be of lower psychometric value, whereas questions offering construct-specific response options may not suffer from these problems. If that is true, then the quality of data collected by agree/disagree questions may be compromised relative to that generated by questions with construct-specific response options. To evaluate this hypothesis, we focus in particular on two standard social science indicators of data quality: reliability and validity.

We begin by offering a theory of the cognitive response process to agree/disagree questions that suggests questions with construct-specific response options may yield higher quality data. Then we test this hypothesis using data from two general public sample surveys done in the Netherlands.

Cognitive Response Processes and Their Consequences

The goal of agree/disagree questions is usually to place respondents on a continuum. For example, a question stem saying “I am usually happy” is intended to gauge frequency of happiness: how often the respondent is happy, on a dimension from “never” to “always.” A question stem saying “I like hot dogs a lot” is intended to gauge quantity of liking/disliking: how much the respondent likes hot dogs, on a dimension from “dislike a lot” to “like a lot.” And a

question stem saying, “Ronald Reagan was a superb President” is intended to gauge respondents’ evaluations of Reagan’s performance, on a dimension ranging from “superb” to “awful.”

To answer such questions requires respondents to execute four cognitive steps (see, e.g., Carpenter and Just 1975; Clark and Clark 1977; Trabasso, Rollins, and Shaughnessy 1971). First, respondents must read the stem and understand its literal meaning. Then, they must look deeper into the question to discern the underlying dimension of interest to the researcher. This is presumably done by identifying the variable quantity in the question stem. In the first example above, the variable is identified by the word “usually” – it is frequency of happiness. In the second example above, the variable is quantity, identified by the phrase “a lot.” And in the third example, the variable is quality, identified by the word “superb.” Having identified this dimension, respondents must then place themselves on the dimension of interest. For example, the stem, “I am usually happy,” asks respondents first to decide how often they are happy. Then, they must translate this judgment onto the agree/disagree response options appropriately, depending upon the valence of the stem. Obviously, it would be simpler to skip this latter step altogether and simply ask respondents directly for their judgments of how often they are happy.

Doing this has another benefit as well, in that it avoids a unique potential problem with agree/disagree questions that we have not yet considered. Researchers often presume that if a question stem is worded “positively,” as all three examples are above (indicating high frequency of happiness, liking of hot dogs, and a positive evaluation of Reagan’s performance, respectively), then people who answer “agree” are indicating more happiness, liking, and positive evaluation, respectively, than people who answer “disagree.” However, “disagree,” “false,” and “no” responses can be offered for various different reasons, some of which violate

the presumed monotonic relation between answers and respondent placement on the underlying dimension of interest.

For example, consider a person who is asked whether he or she agrees or disagrees with the statement: "I am generally a happy person." A person who disagrees may believe (1) he or she is generally an unhappy person, (2) he or she is generally neither happy nor unhappy, and instead is usually affectless, (3) he or she is happy 55% of the time and unhappy 45% of the time, and 55% of the time is not frequent enough to merit the adjective "generally", or (4) he or she is always happy, and "generally" does not represent this universality adequately. In an even more startling example, a respondent asked to agree or disagree with the statement "I like my wife" may well disagree, indignantly announcing "I love my wife."

Even a person who is generally neither happy nor unhappy can end up not only expressing disagreement with the statement about happiness above, but also expressing it strongly. Offering "neither agree nor disagree" as a response option would not necessarily prevent this sort of problem, because a person who is confident that he or she is generally neither happy nor unhappy might well be inclined to strongly disagree in this case. When this sort of mismatch of the response dimension to the latent construct of interest occurs, it will compromise the validity of responses.

If these arguments are true, then responses to questions with construct-specific response options may contain less measurement error than agree/disagree questions. Fortunately, some past studies have compared the reliability and validity of measurements made with agree/disagree questions and questions with construct-specific response options, but their results are mixed. For example, Counte (1979) and Scherpenzeel and Saris (1997) found questions with construct-specific response options to have greater reliability than agree/disagree questions, but

Berkowitz and Wolken (1964) found a slight trend in the opposite direction. Ray (1979), Ross, Steward, and Sinacore (1995), and Schuman and Presser (1981) reported findings suggesting that questions with construct-specific response options had greater correlational validity than agree/disagree questions, though Berkowitz and Wolken (1964), Counte (1979), Ray (1980), and Scherpenzeel and Saris (1997) found trends in the opposite direction. In light of this small body of evidence and the fact that many of the reported differences were not subjected to tests of statistical significance, it seems worthwhile to investigate this issue further, which is what the research reported here was designed to do.

Agree/Disagree Questions with Stems Involving Negation

We also investigated a second issue of relevance to agree/disagree questions: the relative reliability of questions that offer positively phrased stems and those with stems including negations. Many researchers presume that the distorting impact of acquiescence on agree/disagree item responses can be eliminated by building a large “balanced” battery of questions all measuring the same construct. “Balanced” means that about half of the item stems are worded such that agreement indicates a high level of the construct of interest, and the other half of the items are worded so that agreement indicates a low level of the construct. To try to assure that item reversals in a balanced battery are truly mutually contradictory and are equal in extremity and social desirability implications, many researchers have begun with affirmatively-phrased statements (e.g., "I am generally a happy person") and then complemented them with an equal number of negatively-phrased statements, often including negations (e.g., "I am not generally a happy person"). Then, when combining a person’s answers to all the items in the battery, acquiescence in response to the positively-phrased items is thought to be cancelled out by acquiescence in response to the negatively-worded items.

This approach only makes sense if people who do not acquiesce respond equivalently reliably and validly to both positively and negatively phrased items. But much research in psychology suggests a reason for caution before making this assumption; people generally make more cognitive errors when processing negative statements than they make when processing affirmative statements. For example, Eifermann (1961) found that when asked to make true/false judgments about statements regarding numbers (e.g., "6 is not an even number"), people made more errors when responding to statements that included the word "not" (35% errors) as compared to affirmative statements (2%). Dudycha and Carpenter (1973) compared multiple-choice tests that asked "Which of the following is ..." or "Which of the following is not ..." People made more errors in answers to the latter, negative questions, than the former, positive questions. Likewise, Mehler (1963) read sentences aloud to respondents and found that people were better able to accurately recall affirmative sentences than negative ones.

Complementing this evidence are studies attesting to the greater cognitive burden imposed by negative statements. Although Slobin (1966) found that people were slower in processing affirmative sentences than negative sentences, many other studies found the opposite (Gough 1965; Wason 1959, 1961, 1962, 1965, 1972; Wason and Johnson-Laird 1972; Wason and Jones 1963; Wembridge and Means 1918). For example, Wembridge and Means (1918) found that it took people, on average, 1.7 seconds to answer an affirmatively-phrased question, 3.4 seconds to answer a question with a single negative in it, and 7.6 seconds to answer a question including a double negative. Likewise, Gough (1965) found that it took people 1.05 seconds on average to make true/false judgments about affirmative sentences, and 1.32 seconds to verify sentences containing a negative. Just and Carpenter (1976) showed that not only does it

take longer to evaluate negative assertions, but it also takes people longer simply to read and understand these assertions.

Taken together, this evidence suggests that when respondents are confronted with agree/disagree question stems with negations, some may answer less reliably and validly than when they respond to positively-phrased stems. If this is so, then any comparisons of agree/disagree questions with questions offering construct-specific response options must be careful to note the impact of negations on responses to the agree/disagree items. We did so and tested the proposition that negations yield more measurement error.

The Present Investigation: Research Design

The first of our two studies used a new data collection design and analytic method. The three quantitative approaches most commonly used for the evaluation of the quality of questions in survey research are the split-ballot experiment (e.g., Billiet, Loosveldt, and Waterplas 1985; Schuman and Presser 1981), the cross-sectional multitrait-multimethod (or MTMM) approach (Andrews 1984; Saris and Andrews 1991, Saris and Münnich 1995; Scherpenzeel and Saris 1997), and longitudinal analysis of panel data (Alwin and Krosnick 1991; Forsman and Schreiner 1991). Each of these approaches has advantages and disadvantages, and the approach we employed combines all three of these approaches to yield a stronger technique than any one approach alone.

In split-ballot experiments, respondents are randomly assigned to be asked different versions of the same question. For example, in order to estimate the extent of acquiescence in a question, respondents can be asked to indicate their opinions on an issue by agreeing or disagreeing with an assertion, and two equivalent groups of respondents can be given opposite assertions (e.g., “Individuals are more responsible than social conditions for crime and

lawlessness in this country" vs. "Social conditions are more responsible than individuals for crime and lawlessness in this country;" Schuman and Presser 1981). If no acquiescence occurs, the proportion of people agreeing with the first assertion should not exceed the proportion of people disagreeing with the second assertion. But if acquiescence does occur, the proportion of people agreeing with the first assertion will exceed the proportion of people disagreeing with the second assertion.

The split-ballot approach can also be used to compare the reliability and validity of measurements made with agree/disagree questions and questions with construct-specific response options if panel data are available. For example, to assess reliability, respondents can be randomly assigned to be asked an agree/disagree question on two occasions or to be asked a version of the same question with construct-specific response options on the two occasions instead. Test-retest correlations could be used to assess the reliabilities of the two question formats; a stronger positive correlation for the construct-specific response option version would suggest its measurements are more reliable.

In addition, the split-ballot approach can be used to compare the reliability and validity of the two question formats across multiple interviews. Respondents can be randomly assigned to be asked either an agree/disagree question measuring a target attitude or a construct-specific response option version of the same question, and elsewhere in the questionnaire or in another questionnaire administered during a later interview can be a measure of another variable that should in theory be correlated with the target attitude. If this latter, "criterion" variable is measured in the same way for all respondents, the association between the target attitude measures and the criterion variable can be viewed as an indicator of the reliability and validity of the target attitude measures. If the association of the criterion variable with the question offering

construct-specific response options is stronger than the association of the criterion variable with the agree/disagree measure of the target attitude, it would suggest that the agree/disagree question is less reliable and/or less valid than the measure offering construct-specific response options.

The multitrait-multimethod panel study approach we employed involved an expanded version of this approach. In a multitrait-multimethod study, each respondent is asked many different questions measuring each of various opinions (also called "traits") using each of various methods. In questionnaire studies, a method is an item format (e.g., a 5-point agree/disagree rating scale, a 101-point feeling thermometer). Ideally, methods are completely crossed with traits, meaning that every opinion is measured by every method. With data collected via this sort of design, it is possible to employ structural equation modeling techniques to estimate the reliability and validity of items in each format, as well as the amount of correlated method-induced error variance for items in each format (see Alwin 1974; Andrews 1984; Browne 1984; Coenders and Saris 2000; Marsh and Bailey 1991; Saris and Andrews 1991). Consequently, a researcher can compare the quality of data collected by various methods.

A potential drawback of this approach is the fact that each respondent must be asked multiple questions assessing the same opinion, and early questions might influence answers to later questions, thus distorting their apparent quality. For example, having just reported my opinion on abortion on a 7-point rating scale, I may use my answer to that question as a basis for deriving a report of that same attitude on a 101-point scale later. The precise meanings of the 101 scale points may not be especially clear, but having reported my opinion on a much clearer 7-point scale first, I may be able to simply translate that report onto the 101-point-scale, thereby bypassing the need to interpret all the scale points. A response of 6 on the 7-point scale, for

instance, corresponds proportionally to a response of about 80 on the 101-point scale. Therefore, if respondents are first asked a question involving a format that is easy to use and yields relatively error-free reports, this may help respondents to provide apparently reliable and valid reports of the same opinions on more difficult-to-use rating scales later. As a result, this approach may under-estimate differences between question formats in terms of reliability and validity.

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

In order to avoid problems in estimation of the parameters of an ordinary MTMM structural equation model, it is necessary to have at least three measures of each construct (Saris 1990). The first study we conducted maximized efficiency by combining the analytic leverage afforded by the MTMM approach with the lack of contamination afforded by split-ballot design and multiple interviews in order to permit gauging the reliability and validity of agree/disagree questions and questions with construct-specific response options measuring the same attitudes.

In particular, respondents from a representative sample of adults were asked questions tapping three different and potentially related beliefs: (1) the importance that many people vote

in elections, (2) the degree to which the respondent would feel ashamed to admit not having voted without having a good reason to abstain, and (3) the likelihood that the respondent would admit to not having voted without having a good reason to abstain. All respondents were asked about these beliefs via agree/disagree questions. Half of the respondents were randomly assigned to receive one set of assertions, and the other half of the respondents received opposite assertions. In a later interview, all respondents were asked a common set of questions measuring the same three beliefs and offering construct-specific response options. Using data from all three sets of measures and both groups of respondents, we were able to compare the results obtained by the different question forms and to build a latent variable multitrait-multimethod model to estimate the reliability and validity of the various question formats.

Our second study involved a different approach to exploring the same issues. Rather than employing the between-respondents and within-respondents experimental manipulations of question formats and wordings and the multiple interview approach employed in Study 1, Study 2 involved only within-subjects manipulations of format administered during a single interview. Therefore, we could employ a more conventional MTMM analytic approach to assess item reliability and validity.

Study 1

Data

Sample

The survey data analyzed for our first study were collected in the context of the elections for the second chamber of the parliament in the Netherlands in 1998, which were held on May 6. Initially, a two-stage random sample was drawn from the voting register of the town of Zaanstad, which is close to Amsterdam and consists of three previously independent cities and the rural

areas between and around them. In the first stage, ten voting districts were selected; then, 1,000 people were selected from the official list of voters for these voting districts, proportional to the size of the district. Computer-assisted telephone interviews (CATI) were conducted by the commercial market and opinion research organization O & S of Amsterdam.

Five hundred persons were randomly drawn from the sample frame for the pre-election interviews, which were completed just before May 6, 1998.[1] The response rate for this survey was 57.5% (AAPOR RR1), which is typical for telephone surveys in the Netherlands. Immediately after election day, attempts were made to reinterview all respondents - the response rate for the second wave was 97.4%.

Measures

After answering 26 questions during the pre-election interviews (about media use, party identification, predictions of voting behavior, and satisfaction with government), half of the respondents, selected randomly and dubbed "Subsample 1," were asked Set 1 of the agree/disagree questions, which presented the following assertions and asked respondents to indicate whether they "strongly agree," "agree," "neither agree nor disagree," "disagree," or "strongly disagree:"[2]

1. It is important that many citizens vote in elections.
2. I would not be ashamed if I had to admit that I did not vote in an election without a good reason.
3. I would certainly admit that I did not vote in such a situation.

[1] Only half the sample was interviewed pre-election in order to implement an experiment to estimate the effect of a pre-election interview on post-election survey responses and on voting.
[2] The order in which these response choices were presented to respondents is the conversationally expected order, running from positive/affirmative to negative (see Holbrook,

The other respondents, dubbed "Subsample 2," were instead asked Set 2 of the agree/disagree questions, which asked respondents to use to the same response scale to evaluate three opposite assertions, reversed by adding or leaving out the word "not":

1. It is not important that many citizens vote in elections.
2. I would be ashamed if I had to admit that I did not vote in an election without a good reason.
3. I would certainly not admit that I did not vote in such a situation.

After answering four questions during the post-election interview (about whether respondent voted and for whom they voted), all respondents were asked the same three questions with construct-specific response options:

1. How important is it that many citizens vote in elections? Very important, important, neutral, unimportant, or very unimportant?
2. How ashamed would you be if you had to admit that you did not vote in an election without a good reason? Very much ashamed, quite ashamed, somewhat ashamed, hardly ashamed, or not at all ashamed?
3. How likely is it that you would admit that you did not vote in such a situation? Very likely, likely, neither likely nor unlikely, unlikely, or very unlikely?

The numeric codings of responses to these questions are shown in Table 1.

Results

Effectiveness of Random Assignment

As would be expected, Subsamples 1 and 2 did not differ significantly in terms of the distributions of various demographic variables, including gender ($\chi^2(1)=0.22$, p=.66), age

---

Krosnick, Carson, and Mitchell 2000). Offering the response options in the reverse order may

($\chi^2(4)=5.48$, p=.24), education ($\chi^2(8)=3.40$, p=.91), occupation ($\chi^2(8)=6.50$, p=.59), and church attendance ($\chi^2(5)=5.30$, p=.37). In addition, Subsamples 1 and 2 did not differ significantly in terms of the distributions of answers to the questions with construct-specific response options, which were asked identically of the two subsamples (see the last two columns of Table 1; Question 1: $\chi^2(4)=7.94$, p=.09; Question 2: $\chi^2(4)=1.93$, p=.75; Question 3: $\chi^2(4)=3.82$, p=.43). Thus, it appears that the random assignment procedure worked effectively to yield two equivalent groups of respondents.

Distributions of Responses

Responses to the agree/disagree questions were coded to range from 1 to 5, such that "1" indicated either strong agreement with positively worded items or strong disagreement with negatively worded items (as shown by the numbers preceding the response options in Table 1). The distribution of responses to Question 1 from Subsample 1 and Subsample 2 were marginally significantly different from one another ($\chi^2(4)=9.23$, p=.056, N=534). Whereas 82.4% of Subsample 1 respondents strongly agreed that voting is important, 89.5% strongly disagreed with the opposite assertion. When we combined the two “agree” responses and the two “disagree” responses, 90.7% of Subsample 1 respondents said either “agree” or “strongly agree,” and 93.0% of Subsample 2 respondents said either “disagree” or “strongly disagree.” The mean score from Subsample 1 (1.35) was not significantly different from the mean score from Subsample 2 (1.27, $t(532)=1.13$, p=.26).

Whereas 45.0% of Subsample 1 respondents strongly agreed with Question 2, only 33.2% of Subsample 2 respondents disagreed with the opposite assertion. These distributions

---

have compromised the quality of responses to these questions (see Holbrook et al. 2000).

differ significantly from one another ($\chi^2(4)=11.67$, p=.02), as do the means for the two subsamples (2.90 vs. 3.34 for Subsamples 1 and 2, respectively, t(491)=2.65, p=.008).

The same pattern appeared for Question 3. The proportion of Subsample 1 respondents who strongly agreed with Question 3 (86.8%) exceeded the proportion of Subsample 2 respondents who strongly disagreed (77.6%). These distributions differed significantly ($\chi^2(4)=17.87$, p=.001), as did the means for the two subsamples (1.38 vs. 1.62 for Subsamples 1 and 2, respectively, t(524)=2.33, p=.020).

The distributions of responses to the agree/disagree questions are also different from the distributions of responses to the comparable questions with construct-specific response options. In general, more extreme responses were given to the agree/disagree questions than were given to the questions with construct-specific response options. For example, for Question 1, 85.6% and 93.0% of respondents chose an extreme response option when answering the agree/disagree items, respectively, whereas only 67.0% and 71.7% of respondents chose an extreme response option when answering the comparable questions with construct-specific response options. Likewise, 84.6% and 80.6% of respondents chose an extreme response option when answering the two agree/disagree versions of Question 2, respectively, whereas 58.8% and 59.0% of respondents chose an extreme response option when answering the comparable construct-specific response option question. And 92.8% and 89.4% of respondents chose an extreme response option when answering the two agree/disagree versions of Question 3, respectively, whereas 81.3% and 80.1% of respondents chose an extreme response option when answering the comparable construct-specific response option question.

Not surprisingly, then, the variances of answers to the agree/disagree questions were consistently larger than the variances of answers to the comparable questions with construct-

specific response options, significantly so in four of the six comparisons (Subsample 1: Question 1 agree/disagree s^2=0.81 (N=278) vs. Question 1 construct-specific response option version s^2=.64 (N=269), $\chi^2(1)=5.92$, $p<.05$; Question 2 agree/disagree s^2=3.53 (N=240) vs. Question 2 construct-specific response option version s^2=2.56 (N=269), $\chi^2(1)=10.08$, $p<.01$; Question 3 agree/disagree s^2=1.17 (N=281) vs. Question 3 construct-specific response option version s^2=1.23 (N=268), $\chi^2(1)=0.22$, n.s.; Subsample 2: Question 1 agree/disagree s^2=0.77 (N=256) vs. Question 1 construct-specific response option version s^2=0.41 (N=258), $\chi^2(1)=29.14$, $p<.01$; Question 2 agree/disagree s^2=3.28 (N=253) vs. Question 2 construct-specific response option version s^2=2.59 (N=249), $\chi^2(1)=6.63$, $p<.05$; Question 3 agree/disagree s^2=1.77 (N=245) vs. Question 3 construct-specific response option version s^2=1.17 (N=239), $\chi^2(1)=12.5$, $p<.01$).

Thus, on the surface, one might be tempted to infer that the two question formats supposedly measuring the same attitude or belief did not in fact do so. However, these distributions are completely contingent on the particular response scale point labels offered for each of the questions. If instead of labeling the end points of the agree/disagree scale “strongly agree” and “strongly disagree,” we had instead labeled them “very strongly agree” and “very strongly disagree,” we would most likely have found fewer respondents choosing those most extreme options (e.g., Klockars and Yamagishi 1988). Likewise, if we had labeled the end-points of the construct-specific response options less extremely, more respondents would probably have selected them. As a result, we could have made the observed distributions of responses more comparable across the agree/disagree and construct-specific response option versions of the questions. Thus, despite the differences between the agree/disagree and the construct-specific response option forms in terms of the observed distributions of responses, the

different forms of the same question may in fact tap the same underlying latent attitude or belief and may do so equally reliably and validly.

Random Measurement Error, Reliability, Validity, and Quality

In MTMM experiments commonly conducted to estimate reliability and validity, a minimum of 3 traits and 3 methods are used, yielding a correlation matrix among 9 variables. In the current split-ballot MTMM design, we measured 3 traits (opinions, in this case), but in each sample, only two methods were implemented (see the correlation matrices in the LISREL input displayed in the Appendix). Each zero correlation occurred because the two questions were not asked of the same respondents.

To analyze conventional MTMM correlation matrices, various analytic models have been suggested (Alwin 1974; Andrews 1984; Browne 1984; Coenders and Saris 2000; Marsh and Bailey 1991; Saris and Andrews 1991). We employed the model suggested by Saris and Andrews (1991), which has the advantage of making an explicit distinction between reliability and validity:

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

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

where Y_{ij} is the observed variable for the j^{th} trait (attitude or belief in this case) and the i^{th} method, T_{ij} is the systematic component of the response Yij, F_j is the j^{th} trait, and M_i represents the variation in scores due to the i^{th} method. This model posits that the observed variable is the sum of the systematic component plus random error. And the systematic component of a measure is the sum of the trait and the method used to assess it, as shown graphically in Figure 1.

In order for the error terms to represent the amount of random measurement error in the responses (which is what we want to estimate), these errors are specified to be uncorrelated with

each other and with the independent variables in the different equations. The trait factors were permitted to correlate with one another, but the method factors were assumed to be uncorrelated with one another and with the trait factors, which is a standard approach in specifying such models (e.g., Jarvis and Petty 1996; Krosnick and Alwin 1988; for more details, see Saris and Andrews 1991).

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

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

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

Although MTMM analyses usually require at least three measures of at least three traits in a single sample of respondents, our design provides only two measures of three traits in a single sample. This might seem to leave the model in Equations (1) and (2) under-identified. However, because we have data from two independent samples, both of which provide estimates

of some of the same correlations and underlying relations among latent variables, it is in fact possible to estimate the model's parameters via the multisample approach in structural equation modeling (e.g., LISREL, Jöreskog and Sörbom 1991). The statistical justification for our approach has been discussed extensively by Allison (1987) and Satorra (1990, 1992).

The agree/disagree items may share method-induced variance to the extent that they manifest acquiescence if the same respondents who acquiesce in answering one agree/disagree question do so when answering another. We, therefore, specified one method factor to represent acquiescence. The second and third opinion questions manifested response patterns consistent with acquiescence (see Table 1), so these items were allowed to load on the acquiescence method factor. We imposed the constraint that the variance of the acquiescence method factor in Subsample 1 should be the same as the variance of the acquiescence method factor in Subsample 2. We also allowed for the possibility that the construct-specific response option questions might manifest a method effect, but this effect was not significant, so we omitted it from the model.

In estimating the effects, we introduced the restriction that the validity coefficients and error variances are the same in the two subsamples for each of the questions with construct-specific response options. This is reasonable, because these questions were asked identically in both subsamples, and respondents were randomly assigned to the subsamples. Furthermore, in the input, the effect of the true score on the observed variables involving questions not measured were all fixed at 0, and the error variances of the questions not asked of a subsample were fixed at 1.0. Automatically, then, the model yields correlations of zero and variances of 1.0 for the not observed "measured" variables. LISREL considers the input correlations of zero to be observed data points even though they were in fact not, so we subtracted a total of 48 degrees of freedom

from the number of degrees of freedom given by LISREL to compensate for these 48 illusory correlations. The LISREL input instructions can be found in the Appendix.

When the parameters of this model were estimated using our data, they fit the data adequately ($\chi^2(20)=28.9$; n.s., χ^2/d.f.=1.44, Δ=.96; ρ=.94). In this model, the variance of the acquiescence method factor (s^2=.12) was significantly different from zero (z=1.75, p=.04), consistent with the assumption that acquiescence was present in answers to the agree/disagree questions tapping the second and third beliefs.

To see whether the agree/disagree question format yielded more random error variance in responses, we examined the amounts of variance in each item attributable to random measurement error. These variances are shown in Table 2 for all of our measures, and they confirm our expectations in two regards. First, the amounts of random error variance in answers to the construct-specific response option items are always smaller than the amounts of random error variance in answers to the comparable agree/disagree items worded in an affirmative direction. This difference is statistically significant for Question 1 (in Subsample 1, Δ =.32, $\chi^2(1)=14.9$, $p<.05$) and Question 2 (in Subsample 2, Δ =.89, $\chi^2(1)=6.2$, $p<.05$), though not for Question 3 (in Subsample 1, Δ =.03, $\chi^2(1)=0.0$, n.s.). This is consistent with the conclusion that the agree/disagree format introduced additional random variance over and above that which was present in questions with construct-specific response options.

In addition, we see that the amounts of random error variance in affirmatively-worded agree/disagree questions were always smaller than the amounts of random error variance in answers to the negatively-worded agree/disagree questions (shown in bold in Table 2), which included the word “not.” This difference was again significant for Question 1 (Δ=.13, $\chi^2(1)=4.3$,

$p<.05$) and Question 2 ($\Delta=.63$, $\chi^2(1)=6.3$, $p<.05$), though not for Question 3 ($\Delta=.15$, $\chi^2(1)=0.3$, n.s.).

Shown in the upper portion of Table 3 are estimates of the item reliabilities. For Question 1, the construct-specific response option item's reliability (.88) is considerably greater than those for the two agree/disagree formats (.46 and .36). For Question 2, the same pattern appeared: the construct-specific response option question was considerably more reliable (.87) than were the two agree/disagree formats (.44 and .61). For Question 3, all the reliabilities were very low. The reliabilities of the questions using construct-specific response options were significantly larger than the reliabilities of the agree/disagree questions in Subsample 1, $\chi^2(3)=27.37$, $p<.001$. Likewise, the reliabilities of the questions using construct-specific response options were significantly larger than the reliabilities of the agree/disagree questions in Subsample 2, $\chi^2(3)=29.90$, $p<.001$. Thus, the use of construct-specific response options led to significantly higher reliabilities.

Shown in the middle portion of Table 3 are the validities of the measures. Because we expected no method-induced systematic variance in answers to the questions with construct-specific response options (because the rating scales were all different; the plausibility of this assumption was confirmed by the non-significance of the construct-specific response option items' method factor's variance in our initial model), the construct-specific response option items all had validities of 1.0. Furthermore, because we saw no evidence of acquiescence in answers to the agree/disagree measures of the first attitude, the validities for those two items were also 1.0. In other words, all systematic variance in those items is attributable to the latent attitudes or beliefs underlying them. Reflecting the presence of acquiescence-induced method variance in the second and third agree/disagree questions, these items' validities are less than 1.0

(.92 and .76 in Subsample 1; .94 and .88 in Subsample 2). However, the validities of the items with construct-specific response options were not significantly different from the reliabilities of the agree/disagree items ($\chi^2(2)=0.47$, n.s.).

The bottom portion of Table 3 displays estimates of what we call the "quality" of the measures, which are the products of the reliabilities and the validities. These measures of quality squared indicate the portion of the variance in responses to each item that is attributable to the true underlying attitude or belief being measured. And here, the differences between the construct-specific response option items and the agree/disagree items, in terms of quality, are large. In five of the six comparisons, the construct-specific response option measure of an attitude or belief had substantially higher quality than the agree/disagree measure of the same attitude or belief.

Discussion

The evidence from our first study is consistent with the notion that agree/disagree rating scales yield responses containing more random measurement error than do questions with construct-specific response options. However, there is an alternative possible explanation for the results observed: the questions with construct-specific response options were asked after only four questions during the post-election interviews, whereas the agree/disagree questions were asked after 26 other questions during the pre-election interviews. If respondents became fatigued as they answered an increasing number of questions, the effort they devoted to generating optimal answers may have declined, so any question asked later may have lower reliability and validity than the same question would have if it had been asked early in an interview. In fact, some past studies suggest that people are more attracted to "don't know" response options later in a long questionnaire (e.g., Culpepper, Smith, and Krosnick 1992;

Dickinson and Kirzner 1985; Ferber 1966; Krosnick et al. 2002; Ying 1989), are more likely to manifest acquiescence response bias (e.g., Clancy and Wachsler 1971), and will non-differentiate when responding to batteries of ratings using the same scale (e.g., Herzog and Bachman 1981; Kraut, Wolfson, and Rothenberg 1975). This evidence suggest that when people do give substantive answers to later questions, those answers are generated less thoughtfully and precisely.

Study 2

To explore whether this explanation fully accounts for Study 1's results, we conducted a second study using a different analytic approach applied to data that were collected during a single interview, with the agree/disagree questions asked before the construct-specific response option questions. Study 2 also explored the generality of Study 1's findings to a new set of attitude questions. Three agree/disagree questions were asked of respondents early in an interview, followed later by the same three questions asked with construct-specific response options, followed still later by re-administration of the agree/disagree questions.

This design allowed us to explore whether later placement of the same questions in a questionnaire improves or compromises measurement precision. Furthermore, we could compare the amounts of random measurement error, reliabilities, and validities of the same questions asked in different formats.

Data

Sample

The survey data analyzed in our second investigation were collected as a pilot study for the European Social Survey. A simple random sample of 2,000 addresses and telephone numbers were drawn from a list of the Dutch population. These addresses were spread over 30

geographic areas. In each area, one interviewer conducted computer-assisted personal interviews under supervision of the scientific research organization SRF. The resident of each household with the most recent birthday was asked to participate (Salmon and Nichols 1983).

Because one month was allocated for data collection, the interviewers were instructed to complete as many interviews as quickly as possible, with a target total sample size of 500. After one month, 427 interviews had been completed, yielding a response rate a bit below 25% (AAPOR RR2). A total of 409 respondents answered all of the questions that we analyzed. 44.7% of them were male and 55.3% were female. 93.4% were born in the Netherlands and 6.6% were born elsewhere. 21.8% were age 24 or younger, 10.3% were ages 25-30, 15.2% were ages 31-40, 18.1% were ages 41-50, 20.2% were ages 51-64, and 14.4% were age 65 or older. 4.6% of respondents had only primary school education, 8.1% has completed lower vocational school education, 13.9% had completed lower general education, 18.3% had completed higher vocational school, 7.6% had completed higher general education, 4.6% had completed school preparing for University, 29.8% had some higher education, and 13.0% had completed University education. As is routinely true of surveys, this sample under-represented older people and people with relatively little education.

Measures

After nine questions at the beginning of the interview about news media use and interest in politics, respondents were handed a showcard displaying the response choices “strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree.” Respondents were then asked: “Using this card, how much do you agree or disagree with each of the following statements?”

Q1. Sometimes politics and government seem so complicated that I can't really understand what is going on.

Q2. I think I can take an active role in a group that is focused on political issues.

Q3. I understand and judge important political questions very well.

The next set of questions took approximately one hour to ask and addressed a wide range of topics, including political attitudes, social networks, crime victimization, life satisfaction, religion, ethnicity and ethnocentrism, internet use, demographics, and much more.

Next, respondents were asked questions comparable to Q1, Q2, and Q3 with construct-specific response options:

Q1. How often do politics and government seem so complicated that you can't understand what is going on? Never, seldom, occasionally, regularly, or frequently?

Q2. Do you think that you could take an active role in a group that is focused on political issues? Definitely not, probably not, not sure, probably, or definitely?

Q3. How good are you at understanding and judging important political questions? Very bad, bad, neither good nor bad, good, or very good?

After answering 27 additional questions (about interpersonal trust, approval of government performance, and other topics), respondents were asked the agree/disagree forms of the three target questions again.

The numeric coding of responses to these questions are shown in Table 4.

Results

Distributions of Responses

Table 4 displays the distributions of responses to the agree/disagree questions (see columns 1 and 2). These figures reveal little change in the distribution between the first and second administrations of each question. The shapes of the distributions varied from item to item, with a plurality of respondents disagreeing with Q1's statement, a majority disagreeing with Q2's statement, and a plurality agreeing with Q3's statement. Questions with construct-specific response options (see column 3 of Table 4) revealed that most respondents rarely failed to understand politics and government; most said they could not take an active role in a politically focused group; and a plurality said they were good at understanding and judging political questions. Thus, the two question forms yielded similar portraits of the distributions of opinions on these issues.

Random Measurement Error, Reliability, Validity, and Quality

To gauge the reliability and validity of these items, we estimated the parameters of a structural equation model that posited each observed measure was a function of the latent attitude it tapped and a method factor. We began estimating and testing the parameters with the model shown in Figure 1, with the same restrictions as were imposed in Study 1 (that the method effects for the items with construct-specific response options were zero). Because all respondents answered all questions, estimation of the parameters was readily accomplished using conventional methods for a single sample of respondents.

Goodness-of-fit indices suggested that the model did not fit the data adequately ($\chi^2(22)=86.21$; $p<.001$; $\chi^2/df=3.92$, RMSEA=.10; $\Delta=.95$; $\rho=.94$). The modification indices suggested that we include a method factor for the items with construct-specific response options,

and adding this factor significantly improved the fit of the model ($\Delta\chi^2(1)=22.00$; $p<.01$), but the model's fit was still not adequate. The modification indices suggested the need for a correlation between the error terms of Q1 and Q3, which is quite sensible, because both of these questions asked about the extent to which respondents understood politics and explicitly mentioned the concept of "understanding." The model including these correlations fit the data adequately ($\chi^2(18)=30.58$; $p=.03$; $\chi^2/df=1.70$, RMSEA=.04; $\Delta=.98$; $\rho=.97$).

In this model, the method factor variances were all statistically significant (the first administration of the agree/disagree items: $s^2=0.07$, $t=3.47$, $p<.05$; the second administration of the agree/disagree items: $s^2=0.11$, $t=4.01$, $p<.05$; the questions with construct-specific response options: $s^2=0.06$, $t=3.23$, $p<.05$).

To test our principal hypothesis, we again examined whether the agree/disagree question format yielded more random error variance in responses. As shown in Table 5, these variances confirmed our expectations: the amounts of random error variance in answers to the construct-specific response option items were always smaller than the amounts of random error variance in answers to the first administration of the agree/disagree questions (Q1: .28 vs. .78, $\chi^2(1)=35.91$, $p<.01$; Q2: .18 vs. .73, $\chi^2(1)=52.39$, $p<.01$; Q3: .22 vs. .54, $\chi^2(1)=24.63$, $p<.01$) and in answers to the second administration of those questions (Q1: .28 vs. .50, $\chi^2(1)=7.72$, $p<.01$; Q2: .18 vs. .32, $\chi^2(1)=3.50$, $p=.06$; Q3: .22 vs. .36, $\chi^2(1)=6.49$, $p=.01$). This is consistent with the conclusion that the agree/disagree format introduced additional random variance over and above that which was present in questions with construct-specific response options.

The second administration of the agree/disagree questions contained less random error variance than the first administration did (Q1: .50 vs. .78, $\chi^2(1)=11.26$, $p<.01$; Q2: .32 vs. .72, $\chi^2(1)=34.00$, $p<.001$; Q3: .36 vs. .54, $\chi^2(1)=6.87$, $p<.01$).

Shown in upper portion of Table 6 are estimates of the item reliabilities, which manifested the expected pattern. The construct-specific response option items' reliabilities were significantly greater than those for the first administration of the agree/disagree format questions (Q1: .88 vs. .68; Q2: .93 vs. .71; Q3: .88 vs. .73; $\chi^2(3)=48.5$, $p<.001$. The reliabilities of the agree/disagree items were consistently greater the second time they were answered than the first time (Q1: .78 vs. .68; Q2: .97 vs. .71; Q3: .81 vs. .73). And the construct-specific response option items' reliabilities were greater than those for the second administration of the agree/disagree questions for two of the three questions (Q1: .88 vs. .78; Q2: .93 vs. .97; Q3: .88 vs. .81). When tested using all three items, however, the questions using construct-specific response options had significantly higher reliabilities than the second administration of the agree/disagree questions, $\chi^2(3)= 8.5$, $p<.05$).

Shown in the middle portion of Table 7 are the validities of the measures, which were all quite high. The construct-specific response option questions' validities (.97, .98, and .96) were larger than those for the first administration of the agree/disagree questions (.92, .97, and .95) and than those for the second administration of the agree/disagree questions (.95, .96, and .95), but not significantly so. As shown in the bottom portion of Table 7, the qualities of the construct-specific response option question responses (.85, .91, and .84) were substantially larger than the qualities of the agree/disagree question responses from the first administration (.63, .69, and .67) and larger than two of the three qualities for the second administration (.74, .93, and .77).

Discussion

Many prior studies have documented that some people answer agree/disagree questions by agreeing with any assertion, regardless of its content (Krosnick and Fabrigar forthcoming).

Furthermore, we outlined earlier how the cognitive processes entailed in answering an agree/disagree question are likely to be more burdensome and complex than the cognitive processes entailed in answering a comparable construct-specific response option question. And we outlined why responses to agree/disagree items do not necessarily have monotonic relations with the underlying constructs. Presumably, because questions with construct-specific response options avoid acquiescence, minimize cognitive burden, and do indeed produce answers with monotonic relations with the underlying constructs of interest, this format may yield more reliable self-reports.

Few previous studies have compared the amount of measurement error in responses to agree/disagree questions with the amount of such error in responses to comparable questions asked with construct-specific response options. We have reported two studies using two different survey datasets in two different modes involving two different sets of measures addressing two different topics with two different designs (panel vs. cross-section, split-ballot vs. not) and applied two different analytic methodologies to explore this issue. The evidence from both studies is consistent with the conclusion that data quality is indeed higher for questions offering construct-specific response options.

Our first study also indicated that agree/disagree items with stems involving negations (including the word “not”) had lower reliabilities than items without negations. This finding is consistent with past studies that used very different methods to demonstrate that people make more cognitive errors when processing negation statements than they make when processing affirmative statements. We documented this same general phenomenon using a new methodological approach and a new indicator of data quality: amount of random error variance.

This finding suggests caution before presuming that battery balancing is an effective and

wise solution to the acquiescence problem, for a few reasons. First, negation items bring with them an inherent cost: lower data quality due to reduced reliability. And the greater cognitive difficulty entailed in generating answers to these items is likely to enhance respondent fatigue, which may compromise the quality of people's responses to items later in a questionnaire. Furthermore, the "balancing" approach simply places all acquiescing respondents at or near the middle of the response dimension, regardless of the fact that there is no reason to believe that these individuals belong there. This relatively arbitrary placement of those individuals may hurt data quality as well. Therefore, solving the acquiescence problem seems to be accomplished more effectively by using questions with construct-specific response options instead of by balancing large batteries of agree/disagree questions.

Study 2 showed that respondents provided more precise responses to agree/disagree questions the second time they were asked as compared to the first time. There are at least three possible explanations for this difference. First, perhaps people become more precise when answering a question if they have had practice answering the identical question earlier in the same questionnaire. That is, answering the agree/disagree questions early may have improved the quality of responses to the same questions later. This is consistent with the general notion that practice at answering a question improves the precision of later answers to the same question (see, e.g., Donovan and Radosevich 1999; Smith, Branscombe, and Bormann 1988).

However, there are at least three other possible explanations for our finding in this regard. It is possible that answering any questions on a topic, regardless of their format, may enhance the reliability and validity of responses to later questions on the same topic by providing practice at thinking about the topic, not necessarily practice with the particular judgment being requested or response format employed. This interpretation is consistent with other evidence of the same sort

reported by Knowles and colleagues (Knowles 1988; Knowles and Byers 1996). Those investigators experimentally rotated the order in which large sets of questions on a topic were asked of respondents and found that an item's reliability was greater after people had answered many questions measuring the same construct.

Another possibility is that answering the questions offering construct-specific response options produced the improvement in the quality of responses to the later agree/disagree questions. That is, asking questions on the same topic in a more direct way may have facilitated generating answers to more cognitively difficult questions measuring the same attitudes.

Finally, it is technically possible that simply asking the agree/disagree questions later in the interview improved the quality of responses to them, regardless of what items preceded them. However, various studies have indicated that later placement of a question in a questionnaire increases respondent fatigue at the time of answering and thereby reduces response quality. For example, later placement enhances the likelihood that respondents will manifest acquiescence response bias (e.g., Clancy and Wachsler 1971), non-differentiation in responding to batteries of ratings using the same scale (e.g., Herzog and Bachman 1981; Kraut et al. 1975), and selection of an offered "don't know" response option (e.g., Culpepper et al. 1992; Dickinson and Kirzner 1985; Ferber 1966; Krosnick et al. 2002; Ying 1989). Therefore, later question placement *per se* seems unlikely to have been responsible for the improvement in response quality that we observed. Our results suggest a potentially fruitful avenue for future research: exploring what processes were at work to yield the question order effects we observed on responses to the agree/disagree items.

Also of value here is the demonstration of the effectiveness of the split-ballot MTMM approach to assessing the quality of data from various item formats. This approach brings

together the strengths of the traditional split-ballot experiment, the strengths of the MTMM approach, and the strength of panel re-interviews to efficiently assess item reliability and validity while minimizing the burden on respondents and the cost of data collection for researchers. Stated most generally, this technique is appealing because it takes information about item covariances obtained from multiple groups of respondents and combines that information into a single structural equation model to yield general parameter estimates that could not be obtained from any one of the groups of respondents alone. In this way, each respondent has to answer similar questions only twice instead of three times, thereby reducing any effects on the reliability and validity estimates due to repeated observations. More complex split-ballot MTMM designs allow estimation of the effect of repeated observations as well, so the model we employed can be expanded and enhanced in various ways (Saris, Satorra and Coenders forthcoming). The effectiveness of this method demonstrated here encourages further use of this technique in future studies.

References

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

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

Alwin, Duane F. 1974. "Approaches to the Interpretation of Relationships in the Multitrait-Multimethod Matrix." In *Sociological Methodology 1973-1974*, ed. Herbert L. Costner, pp. 79-105. San Francisco, CA: Jossey Bass.

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

Andrews, Frank M. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 48:409-22.

Bem, Daryl J., and H. Keith McConnell. 1970. "Testing the Self-Perception Explanation of Dissonance Phenomena: On the Salience of Premanipulation Attitudes." *Journal of Personality and Social Psychology* 14:23-31.

Berkowitz, Norman H., and George H. Wolkon. 1964. "A Forced Choice Form of the F Scale-Free of Acquiescent Response Set." *Sociometry* 27:54-65.

Billiet, J., G. Loosveldt, and L. Waterplas. 1985. *Het Survey-Interview Onderzocht: Effecten Van Het Ontwerp En Gebruik Van Vragenlijsten Op De Kwaliteit Van De Antwoorden.*

[Research on Surveys: Effects of the Design and Use of Questionnaires on the Quality of the Response]. Leuven, Belgium: Sociologisch Onderzoeksinstituut KU Leuven.

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

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

Carr, Leslie. G. 1971. "The Srole Items and Acquiescence." *American Sociological Review*, 36:287-293.

Clancy, Kevin J., and Robert A. Wachsler. 1971. "Positional Effects in Shared-Cost Surveys." *Public Opinion Quarterly* 35:258-65.

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

Coenders, Germà, and Willem E. Saris. 2000. "Testing Additive and Multiplicative MTMM Models." *Structural Equation Modeling* 7:219-51.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, ed. David E. Apter, pp. 206-61. New York: Free Press.

Costa, Paul T., and Robert R. McCrae. 1988. "From Catalog to Classification: Murray's Needs and the Five-Factor Model." *Journal of Personality and Social Psychology* 55:258-65.

Counte, Michael A. 1979. "An Examination of the Convergent Validity of Three Measures of Patient Satisfaction in an Outpatient Treatment Center." *Journal of Chronic Diseases* 32:583-588.

Culpepper, Irving J., Wendy R. Smith, and Jon A. Krosnick. 1992. "The Impact of Question Order on Satisficing in Surveys." Paper presented at the *Midwestern Psychological Association Annual Meeting*, Chicago, Illinois.

Davis, James A., and Tom M. Smith. 1996. *General Social Surveys, 1972-1996: Cumulative Codebook*. Chicago: National Opinion Research Center.

Dickinson, John R., and Eric Kirzner. 1985. "Questionnaire Item Omission as a Function of Within-Group Question Position." *Journal of Business Research* 13:71-75

Donovan, John J., and David J. Radosevich. 1999. "A Meta-Analytic Review of the Distribution of Practice Effect: Now You See It, Now You Don't." *Journal of Applied Psychology* 84:795-805.

Dudycha, Arthur L., & Carpenter, James B. 1973. "Effects of Item Format on Item Discrimination and Difficulty." *Journal of Applied Psychology* 58:116-21.

Eifermann, Rivka R. 1961. "Negation: A Linguistic Variable." *Acta Psychologica* 18:258-73.

Ferber, Robert. 1966. "Item Nonresponse in a Consumer Survey." *Public Opinion Quarterly* 30:399-415.

Forsman, Gösta, and Irwin Schreiner. 1991. "The Design and Analysis of Reinterview: An Overview." In *Measurement Errors in Surveys*, ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, pp. 279-302. New York: Wiley.

Goethals, George R., and Richard F. Reckman. 1973. "The Perception of Consistency in Attitudes." *Journal of Experimental Social Psychology* 9:491-501.

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

Gough, Phillip B. 1965. "Grammatical Transformations and Speed of Understanding." *Journal of Verbal Learning and Verbal Behavior* 4:107-11.

Herzog, A. Regula, and Jerald G. Bachman. 1981. "Effects of Questionnaire Length on Response Quality." *Public Opinion Quarterly* 45:549-59

Holbrook, Allyson L., Jon A. Krosnick, Richard T. Carson, and Robert C. Mitchell. 2000. "Violating Conversational Conventions Disrupts Cognitive Processing of Attitude Questions." *Journal of Experimental Social Psychology* 36:465-94.

Jarvis, W. Blair G., and Richard E. Petty. 1996. "The Need to Evaluate." *Journal of Personality and Social Psychology* 70:172-94.

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

Just, Marcel A., and Patricia A. Carpenter. 1976. "The Relation Between Comprehending and Remembering Some Complex Sentences." *Memory and Cognition* 4:318-22.

Klockars, Alan J., and Midori Yamagishi. 1988. "The Influence of Labels and Positions in Rating Scales. *Journal of Educational Measurement* 25:85-96.

Knowles, Eric S. 1988. "Item Context Effects on Personality Scales: Measuring Changes the Measure." *Journal of Personality and Social Psychology* 55:312-20.

Knowles, Eric S., and Brenda Byers. 1996. "Reliability Shifts in Measurement Reactivity: Driven by Content Engagement or Self-Engagement?" *Journal of Personality and Social Psychology* 70:1080-90.

Kraut, Allen I., Alan D. Wolfson, and Alan Rothenberg. 1975. "Some Effects of Position on Opinion Survey Items." *Journal of Applied Psychology* 60:774-776.

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 Duane F. Alwin. 1988. "A Test of the Form-Resistant Correlation Hypothesis: Ratings, Rankings, and the Measurement of Values." *Public Opinion Quarterly* 52:526-538.

Krosnick, Jon A., and Leandre R. Fabrigar. forthcoming. *Designing Great Questionnaires: Insights From Psychology*. New York: Oxford University Press.

Krosnick, Jon A., Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Robert C. Mitchell, Stanley Presser, Paul A. Ruud, V. Kerry Smith, Wendy R. Moody, Melanie C. Green, and Michael Conaway. 2002. "The Impact of 'No Opinion' Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice?" *Public Opinion Quarterly* 66:371-403.

Leech, Geoffrey N. 1983. *Principles of Pragmatics*. London; New York: Longman.

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., and Michael Bailey. 1991. "Confirmatory Factor Analyses of Multitrait-Multimethod Data: A Comparison of Alternative Models." *Applied Psychological Measurement* 15:47-70.

Mehler, Jacques. 1963. "Some Effects of Grammatical Transformations on the Recall of English Sentences." *Journal of Verbal Learning and Verbal Behavior* 2:346-51.

Miller, Warren E., and Santa Traugott. 1989. *American National Election Studies Data Sourcebook, 1952-1986.* Cambridge, MA: Harvard University Press.

Ray, John J. 1979. "A Quick Measure of Achievement Motivation – Validated in Australia and Reliable in Britain And South Africa." *Australian Psychologist* 14:337-44.

Ray, John J. 1980. "The Comparative Validity of Likert, Projective, and Forced-Choice Indices of Achievement Motivation." *Journal of Social Psychology* 111:63-72.

Richardson, Stephen A., Barbara Snell Dohrenwend, and David Klein. 1965. "Expectations and Premises: The So-Called 'Leading Question.'" In *Interviewing: Its Forms and Functions*, ed. Richardson, Stephen A., Barbara Snell Dohrenwend, and David Klein. New York: Basic Books.

Robinson, John P., and Phillip R. Shaver. 1973. *Measures of Social Psychological Attitudes*. Ann Arbor, Michigan: Institute for Social Research.

Robinson, John P., and Lawrence S. Wrightsman. 1999. *Measures of Political Attitudes*. San Diego, CA: Academic Press.

Robinson, John P., Phillip R. Shaver, and Lawrence S. Wrightsman. 1991. *Measures of Personality and Social Psychological Attitudes.* San Diego, CA: Academic Press, Inc.

Ross, Caroline K., Colette A. Steward, and James M. Sinacore. 1995. "A Comparative Study of Seven Measures of Patient Satisfaction." *Medical Care* 33:392-406.

Ross, Michael, and Ronald F. Shulman. 1973. "Increasing the Salience of Initial Attitudes: Dissonance Versus Self-Perception Theory." *Journal of Personality and Social Psychology* 28:138-44.

Salmon, Charles T., and John S. Nichols. 1983. "The Next-Birthday Method of Respondent Selection." *Public Opinion Quarterly* 47: 270-6.

Saris, Willem E. 1990. "The Choice of a Research Design For MTMM Studies." In *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, ed. Willem E. Saris and A. van Meurs. Amsterdam: North Holland.

Saris, Willem E., and Ákos Münnich. 1995. *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments.* Budapest, Hungary: Eötvös University Press.

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

Saris, Willem E, Albert Satorra, and Germà Coenders. forthcoming. *A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design.*

Satorra, Albert. 1992. "Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures." *Sociological Methodology* 22:249-278.

Satorra, Albert. 1990. "Robustness Issues in the Analysis of MTMM and RMM Models." In *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies,* ed. Willem E. Saris and A. van Meurs. Amsterdam: North Holland.

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

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York, NY: Academic Press.

Shaffer, David R. 1975a. “Another Look at the Phenomenological Equivalence of Pre- and Postmanipulation Attitudes in the Forced-Compliance Experiment.” *Personality and Social Psychology Bulletin* 1:497-500.

Shaffer, David R. 1975b. “Some Effects of Consonant and Dissonant Attitudinal Advocacy on Initial Attitude Salience and Attitude Change.” *Journal of Personality and Social Psychology* 32:160-68.

Shaw, Marvin E., and Jack M. Wright. 1967. *Scales for the Measurement of Attitudes*. New York: McGraw Hill.

Slobin, Dan I. 1966. “Grammatical Transformation and Sentence Comprehension in Childhood and Adulthood. *Journal of Verbal Learning and Verbal Behavior* 5:219-27.

Smith, Eliot R., Nyla Branscombe, and Carol Bormann. 1988. “Generality of the Effects of Practice on Social Judgement Tasks.” *Journal of Personality and Social Psychology* 54:385-95.

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

van Meurs, A., and Willem E. Saris. 1990. “Memory Effects in MTMM Studies.” In *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, ed. Willem E. Saris and A. van Meurs. Amsterdam: North Holland.

Wason, P. C. 1959. “The Processing of Positive and Negative Information.” *Quarterly Journal of Experimental Psychology* 11:92-107.

Wason, P. C. 1961. “Response to Affirmative and Negative Binary Statements. *British Journal of Psychology* 52:133-142.

Wason, P. C. 1962. *Psychological Aspects of Negation.* University College London: Communication Research Centre.

Wason, P. C. 1965. "The Contexts of Plausible Denial." *Journal of Verbal Learning and Verbal Behavior* 4:7-11.

Wason, P. C. 1972. "In Real Life Negatives are False." *Logique et Analyse* 57-58:17-38.

Wason, P. C., and P. N. Johnson-Laird. 1972. *Psychology of Reasoning; Structure and Content*. Cambridge, MA: Harvard University Press.

Wason, P. C., and Shelia Jones. 1963. "Negatives: Denotation and Connotation." *British Journal of Psychology* 54:299-307.

Wembridge, Eleanor R., and Edgar R. Means. 1918. "Obscurities in Voting Upon Measures Due to Double-Negative." *Journal of Applied Psychology* 2:156-63.

Wixon, D. R., and James D. Laird. 1976. "Awareness and Attitude Change in the Forced-Compliance Paradigm: The Importance of When." *Journal of Personality and Social Psychology* 34:376-84.

Ying, Yu-wen. 1989. "Nonresponse on the Center for Epidemiological Studies-Depression Scale in Chinese Americans." *International Journal of Social Psychiatry* 35:156-63

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

```
analysis of split-ballot MTMM experiments/ CM model : group 1
da ni=9 ng=2 no=270 ma=cm
km
*
1.0
.467    1.0
.006    -.086   1.0
.671    .368    -.026   1.0
.356    .585    -.076   .313    1.0
-.134   -.078   .399    -.103   -.160   1.0
0       0       0       0       0       0       1.0
0       0       0       0       0       0       0       1.0
0       0       0       0       0       0       0       0       1.0
sd
*
.797 1.603 1.110   .890 1.877 1.081   1.00 1.00 1.00
model ny=9 ne=15 ly=fu,fi be=fu,fi ps=sy,fi te=sy,fi
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
value 0 ly 7 7 ly 8 8 ly 9 9
free be 1 10 be 2 11 be 3 12
free be 4 10  be 5 11   be 6 12
free be 7 10 be 8 11 be 9 12
value 1 ps 10 10 ps 11 11 ps 12 12
value 1 be 1 13 be 2 13 be 3 13
value 0  ps 13 13
value  -1  be 5 14
value 1 be 6 14
free ps 14 14
value 1  be 8 15
value -1  be 9 15
eq ps 14 14  ps 15 15
start 1 ps 10 10 ps 11 11 ps 12 12
free ps 10 11 ps 10 12 ps 11 12
free 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
start .5 all
out sc ns ad=off
analysis of split-ballot MTMM experiments: group 2
da ni=9  no=240 ma=cm
km
*
1.0
```

```
.401    1.0
-.092   -.186   1.0
0       0       0       1.0
0       0       0       0       1.0
0       0       0       0       0       1.0
.523    .207    -.021   0       0       0       1.0
.304    .697    -.143   0       0       0       .174    1.0
-.053   -.165   .477    0       0       0       -.121   -.148   1.0
sd
*
.646 1.608 1.084 1.00 1.00 1.00 .874 1.813 1.327
model ny=9 ne=15 ly=fu,fi be=in ps=in te=sy,fi
value 1 ly 1 1 ly 2 2 ly 3 3
value 0 ly 4 4 ly 5 5 ly 6 6
value 1 ly 7 7 ly 8 8 ly 9 9
eq te 1 1 1  te 1 1
eq te 1 2 2  te 2 2
eq te 1 3 3  te 3 3
free te 8 8
free  te 7 7  te 9 9
value 1 te 4 4 te 5 5 te 6 6
start .5 all
out sc ns ad=off
```

Note: In light of the coding of responses (shown in Table 1), the loadings of negatively-worded agree/disagree items on the acquiescence method factors are negative, so that a high score on the method factor is always associated with a tendency to agree with each statement.

Table 1. Distributions of Responses to the Agree/Disagree and Questions with Construct-Specific Response Options (Study 1)

Agree/Disagree Questions				Questions with Construct-Specific Response Options		
Subsample 1		Subsample 2		Subsample 1		Subsample 2
"It is important that many citizens vote in elections."		"It is not important that many citizens vote in elections."		How important is it that many citizens vote in elections?		
1 Strongly agree	82.4%	5 Strongly agree	3.5%	1 Very important	65.9%	71.7%
2 Agree	8.3%	4 Agree	2.0%	2 Important	22.8%	22.3%
3 Neither agree nor disagree	4.3%	3 Neither agree nor disagree	1.6%	3 Neutral	8.6%	4.5%
4 Disagree	1.8%	2 Disagree	3.5%	4 Unimportant	1.5%	1.6%
5 Strongly disagree	3.2%	1 Strongly disagree	89.5%	5 Very unimportant	1.1%	0.0%
N	278	N	256	N	267	247
"I would not be ashamed if I had to admit that I did not vote in an election."		"I would be ashamed if I had to admit that I did not vote in an election."		How ashamed would you be if you had to admit that you did not vote in an election?		
5 Strongly agree	45.0%	1 Strongly agree	47.4%	1 Very ashamed	26.8%	31.8%
4 Agree	6.7%	2 Agree	9.9%	2 Quite ashamed	17.1%	16.3%
3 Neither agree nor disagree	1.7%	3 Neither agree nor disagree	4.7%	3 Somewhat ashamed	14.5%	16.3%
2 Disagree	7.1%	4 Disagree	4.7%	4 Hardly ashamed	9.6%	8.4%
1 Strongly disagree	39.6%	5 Strongly disagree	33.2%	5 Not at all ashamed	32.0%	27.2%
N	240	N	253	N	228	239
"I would certainly admit that I did not vote in such a situation."		"I would certainly not admit that I did not vote in such a situation."		How likely is it that you would admit that you did not vote in such a situation?		
1 Strongly agree	86.8%	5 Strongly agree	11.8%	1 Very likely	75.7%	75.7%
2 Agree	3.2%	4 Agree	0.4%	2 Likely	10.4%	10.6%
3 Neither agree nor disagree	1.1%	3 Neither agree nor disagree	3.7%	3 Neutral	4.9%	5.8%
4 Disagree	2.8%	2 Disagree	6.5%	4 Unlikely	3.4%	3.5%
5 Strongly disagree	6.0%	1 Strongly disagree	77.6%	5 Very unlikely	5.6%	4.4%
N	281	N	245	N	268	226

Table 2. Random Error Variance Estimates for Each Measure (Study 1)

	Subsamples 1 & 2: Construct-Specific Response Options	Subsample 1: Agree/ Disagree	Subsample 2: Agree/ Disagree
Question 1	.07	.39	**.52**
Question 2	.41	**1.93**	1.30
Question 3	.63	.66	**.81**

Note: The estimates in bold and underlined are for the negatively-worded agree/disagree items.

Table 3. Reliability, Validity, and Quality Estimates for the Measures (Study 1)

	Subsamples 1 & 2: Construct-Specific Response Options	Subsample 1: Agree/ Disagree	Subsample 2: Agree/ Disagree
Reliability			
Question 1	.88	.46	**.36**
Question 2	.87	**.44**	.61
Question 3	.48	.44	**.55**
Validity			
Question 1	1.00	1.00	**1.00**
Question 2	1.00	**.92**	.94
Question 3	1.00	.76	**.88**
Quality			
Question 1	.88	.46	**.36**
Question 2	.87	**.41**	.57
Question 3	.48	.33	**.48**

Note: Quality is the product of reliability and validity. The estimates in bold and underlined are for the negatively-worded agree/disagree items.

Table 4. Distributions of Responses to Agree/Disagree Questions and Questions with Construct-Specific Response Options (Study 2)

Agree/Disagree Questions				Questions with Construct-Specific Response Options			
		First Administration	Second Administration				
"Sometimes politics and government seem so complicated that I can't really understand what is going on."	5 Strongly Agree	9.1	5.6	"How often do politics and government seem so complicated that you can't understand what is going on?"	5 Frequently	7.3
	4 Agree	27.3	24.9		4 Regularly	19.1
	3 Neither	17.9	21.0		3 Occasionally	29.4
	2 Disagree	33.7	36.9		2 Seldom	30.8
	1 Strongly Disagree	12.0	11.5		1 Never	13.4
	Total	100%	100%		Total	100%
	N	425	427		N	426
"I think I can take an active role in a group that is focused on political issues."	5 Strongly Agree	5.2	2.9	"Do you think that you could take an active role in a group that is focused on political issues?"	5 Definitely	3.9
	4 Agree	17.9	21.5		4 Probably	16.1
	3 Neither	12.6	16.6		3 Not sure	15.9
	2 Disagree	37.4	35.7		2 Probably not	30.8
	1 Strongly Disagree	26.7	23.2		1 Definitely not	33.2
	Total	100%	100%		Total	100%
	N	422	427		N	426
"I understand and judge important political questions very well."	5 Strongly Agree	7.3	5.1	"How good are you at understanding and judging important political questions?"	5 Very good	7.3
	4 Agree	32.0	40.7		4 Good	38.8
	3 Neither	26.7	26.7		3 Neither	33.7
	2 Disagree	25.4	20.1		2 Bad	15.6
	1 Strongly Disagree	7.6	7.4		1 Very bad	4.6
	Total	100%	100%		Total	100%
	N	423	426		N	426

Table 5. Random Error Variance Estimates for Each Measure (Study 2)

	Construct-Specific Response Options	Agree/ Disagree (First Admin.)	Agree/ Disagree/ (Second Admin.)
Question 1	.28	.78	.50
Question 2	.18	.73	.32
Question 3	.22	.54	.36

Note: All errors variances are statistically significant ($p<.05$).

Table 6. Reliability, Validity, and Quality Estimates for Each Measure (Study 2)

	Construct-Specific Response Options	Agree/ Disagree (First Admin.)	Agree/ Disagree/ (Second Admin.)
Reliability			
Question 1	.88	.68	.78
Question 2	.93	.71	.97
Question 3	.88	.73	.81
Validity			
Question 1	.97	.92	.95
Question 2	.98	.97	.96
Question 3	.96	.95	.95
Quality			
Question 1	.85	.63	.74
Question 2	.91	.69	.93
Question 3	.84	.67	.77

Note: Quality is the product of reliability and validity.

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