Survey Measurement

Sociological Methodology
2016, Vol. 46(1) 121–152
© American Sociological Association 2016
DOI: 10.1177/0081175016641714
http://sm.sagepub.com

4

THE KISS PRINCIPLE IN SURVEY DESIGN: QUESTION LENGTH AND DATA QUALITY

Duane F. Alwin*
Brett A. Beattie*

Abstract

Writings on the optimal length for survey questions are characterized by a variety of perspectives and very little empirical evidence. Where evidence exists, support seems to favor lengthy questions in some cases and shorter ones in others. However, on the basis of theories of the survey response process, the use of an excessive number of words may get in the way of the respondent's comprehension of the information requested, and because of the cognitive burden of longer questions, there may be increased measurement errors. Results are reported from a study of reliability estimates for 426 (exactly replicated) survey questions in face-to-face interviews in six large-scale panel surveys conducted by the University of Michigan's Survey Research Center. The findings suggest that, at least with respect to some types of survey questions, there are declining levels of reliability for questions with greater numbers of words and provide further support for the advice given to survey researchers that questions should be as short as possible, within constraints defined by survey objectives. Findings reinforce conclusions of previous studies that verbiage in survey questions—either in the question text or in the introduction to the question—has

*Pennsylvania State University, University Park, PA, USA

Corresponding Author:
Duane F. Alwin, Pennsylvania State University, 309 Pond Laboratory, University Park, PA 16802, USA
Email: dfa2@psu.edu

negative consequences for the quality of measurement, thus supporting the KISS principle ("keep it simple, stupid") concerning simplicity and parsimony.

Keywords

survey design, measurement, reliability, question length

> One of the things I've learned as a reporter is that you get the best answers, not when you ask long questions, but when you ask short ones.
>
> —Bob Schieffer, broadcast journalist, CBS News, August 31, 2004

1. INTRODUCTION

There is no lack of expert opinion among survey researchers on how to write good questions and develop good questionnaires. Over the past century, vast amounts have been written on the topic of what constitutes a good question, from the earliest uses of surveys down to the present (e.g., see Belson 1981; Converse and Presser 1986; Galton 1893; Krosnick and Fabrigar 1997; Krosnick and Presser 2010; Ruckmick 1930; Saris and Gallhofer 2007; Schaeffer and Presser 2003). On some issues addressed in this literature, there is little consensus on the specifics of question design, which leads some to view the development of survey questions as more of an art than a science. Still, efforts have been made to codify what attributes are possessed by "good questions" and/or "good questionnaires," espousing principles based on a more scientific approach (e.g., Schaeffer and Dykema 2011; Schuman and Presser 1981; Sudman and Bradburn 1974, 1982; Tanur 1992).

With regard to question length, Payne's (1951:136) early writings on surveys, for example, suggested that questions should rarely number more than 20 words. In this tradition, a general rule for formulating questions and designing questionnaires is that questions should be short and simple (e.g., Brislin 1986; Fowler 1992; Sudman and Bradburn 1982; van der Zouwen 1999). Other experts suggest that lengthy questions may work well in some circumstances and have concluded, for example, that longer questions may lead to more accurate reports in some behavioral assessments (see Cannell, Marquis, and Laurent 1977; Marquis, Cannell, and Laurent 1972; Bradburn, Sudman, and Associates 1979). Advice to researchers on question length has therefore been somewhat mixed, and although the testimony of broadcast journalist Bob Schieffer, quoted above, may not be directly relevant to the goals of large-scale survey data collection, there is a common thread there that resonates with many survey researchers, namely, "keep it short and simple."

2. THE KISS PRINCIPLE

The KISS principle (in which KISS is an acronym for "keep it simple, stupid") emphasizes simplicity of design. Other variants of this principle can be found in the vernacular of the day: "keep it short and simple" and "keep it short and sweet," and we are sure there are others. The key goal of the KISS principle is that unnecessary redundancy and complexity should be avoided, and the achievement of perfection depends on parsimony. The principle is applied in scientific reasoning, in which parsimony is revered (as in Occam's razor), as well as in art, in which perfection, it is often claimed, is "reached not when there is nothing left to add, but when there is nothing left to take away."[1] It is relevant to many domains of life, and the principle has been applied in a variety of fields, from software development to film animation.

In survey research, the KISS principle may be applied to the issue of the length of survey questions, both with respect to the length of the actual question and the length of the introduction to the question (e.g., Alwin 2007; Andrews 1984; Saris and Gallhofer 2007; Scherpenzeel and Saris 1997). By focusing on question length, we may be sidestepping another important issue, question comprehension, but we do not normally have independent assessments of comprehension (which ultimately resides in the mind of the individual), whereas question length can be objectively measured (by counting the number of words used in the question). Furthermore, it is possible to distinguish between the length of the introduction to a question, or the introduction to the series of questions in which a given question is embedded, and the length of the question itself. It is argued that either type of question length is fundamentally related to the overall burden felt by the respondent, but lengthy introductions may be viewed by the researcher as a helpful aid to answering the questions. For example, Converse and Presser (1986) suggested that "the best strategy is doubtless to use short questions when possible and slow interviewer delivery—always—so that respondents have time to think" (pp. 12–13). At the same time, they conceded that "in other cases, long questions or introductions may be necessary to communicate the nature of the task" (p. 12). Both respondents and interviewers, on the other hand, may find lengthy introductions time-consuming and distracting.

3. QUESTION LENGTH AND THE SURVEY PROCESS

One of the most basic elements of survey quality is the respondent's comprehension of the question, and the issue of question length is germane to this objective (see, e.g., Tourangeau, Rips, and Rasinski 2000:23–61). Comprehension may be dependent on question length in multiple, and possibly countervailing, ways. Longer questions may be less, rather than more, clear in their meaning, and more complex (longer) questions may reduce comprehension (see, e.g., Holbrook, Cho, and Johnson 2006; Knauper et al. 1997; Yan and Tourangeau 2008). If a question is ambiguous in its meaning, or if parts of the question can have more than one meaning, then the likelihood of measurement error will be increased. Thus, from the point of view of communication, too many words may get in the way of the respondent's comprehension. One study (Holbrook et al. 2006) found that question length was related to both comprehension problems and mapping problems, as measured by behavior coding. The investigators also found that measures of complexity, such as the reading level of the question, were also predictive of respondents understanding questions. In their research, however, the relationship between question length, complexity, and question problems was not always straightforward.[2]

On the other hand, some experts encourage redundancy in survey questions precisely to enhance comprehension (Brislin 1986). Noting the trade-offs between adding redundancy and question length, Converse and Presser (1986) wrote,

> One should consider the use of redundancy now and then to introduce new topics and also to flesh out single questions, but if one larded all questions with "filler" phrases, a questionnaire would soon be bloated with too few, too fat questions. (P. 12)

Although the key element here is probably not question length *per se* but question clarity (Converse and Presser 1986:12), the addition of question text needs to be evaluated in terms of its effects on measurement precision. The phrase "now and then" is vague, and this does not provide clear guidelines regarding the use of redundancy in the phrasing of questions. Additional arguments by experts for using longer questions with redundant information are mentioned by Cannell, Miller, and Oksenberg (1981:406). They mentioned three possible explanations for the finding that a greater number of relevant health events were reported when longer questions were used: longer questions (with redundant information) state the question twice, and this (1) improves the understanding of the question, (2) provides the respondent longer time to think, and (3) encourages (motivates) the respondent to answer by showing higher interest in the interview. Finally, Bradburn et al. (1979:73–74) briefly discussed the advantages of greater question length when requesting information about socially undesirable activities.

Despite all of this advice, there is little evidence on the issue of the relationship between question length and measurement errors, although it is clear that there are a variety of points of view. Where there is evidence, the support seems to favor longer questions. Using the multitrait, multimethod (MTMM) approach (see Saris and Gallhofer 2007), Andrews (1984) combined the consideration of question length and the length of introductions to questions. He found that "short introductions followed by short questions are not good . . . and neither are long introductions followed by long questions" (p. 431). Errors of measurement were lowest when "questions were preceded by a medium length introduction (defined as an introduction of 16 to 64 words)" followed by medium or long questions (defined by Andrews [1984:431] as "16–24 words and 25+ words, respectively"). By contrast, also using the MTMM approach and similar definitions of question and introduction length, Scherpenzeel and Saris (1997) found long questions with long introductions to be superior, but like Andrews (1984), they did not separate conceptually the two issues of battery/series introduction length and question length.[3] Similarly, using an MTMM approach, Saris and Gallhofer (2007) showed that the length of the question has no significant impact on the reliability of measures, but it does have a significant effect on the trait validity coefficients. On the other hand, in their research, the mean number of words per sentence has a significant effect on reliability but not on validity. Alwin's (2007:202–19) research provides an interesting counterpoint to these results, which shows that question length interacts with the questionnaire context of the question—his results indicating that for stand-alone questions and questions contained within a series of questions on a common topic (but not for questions in batteries), there are diminishing returns to longer questions in terms of the estimated reliability of questions. This research poses an interesting puzzle with respect to the possible interaction between question context, question length, and reliability of measurement.

To be clear, one of the key issues in this debate has to do with the role of comprehension in understanding the relationship between question length and reliability of measurement, but we cannot address this here. Comprehension is a process that occurs within the respondent, and although it cannot be examined in the present study, there are ways in which it can be possibly investigated using cognitive interviews in future research. The assumption is that improved comprehension translates into greater reliability, but what remains unclear is the role that question length contributes to greater comprehension. Greater question length can improve comprehension, but it may also contribute to cognitive burden and confusion, which may reduce accuracy of measurement.

In this article, we focus specifically on the role of survey content and its interaction with question length in the evaluation of the role of question length on reliability of measurement using a longitudinal approach (see Alwin 2007:122–27). There is a long-standing distinction in the survey methods literature between objective and subjective questions (e.g., Kalton and Schuman 1982; Schuman and Kalton 1985; Turner and Martin 1984). The distinction used here between "fact" and "nonfact" is derived from this early work, in which the former refers to information "directly accessible to the external observer" and the latter to phenomena that "can be directly known, if at all, only by persons themselves" (Schuman and Kalton 1985:643).[4] In the words of Schuman and Kalton (1985),

> the distinction is a useful one, since questions about age, sex, or education could conceivably be replaced or verified by use of records of observations, while food preferences, political attitudes, and personal values seem to depend ultimately on respondent self-reports. (P. 643)

From a practical point of view, there is a potential confounding of (1) question length driven by the choice of the variables for study and (2) question length driven by the design considerations mentioned in the above review of the literature. This confounding is problematic because (1) and (2) can lead to very different recommendations to practitioners. If question length driven by design considerations leads only to serious degradation of data quality, then we may need to consider fundamentally different approaches to capturing the information of interest. If length driven by (2) leads to serious degradation of data quality, then we may wish to continue an effort to capture the variable *X* through a survey, but with more restraint on the use of the aforementioned verbiage. Hence, there is the need to attempt to separate the effects of question content from question design (in this case question length) in analyzing their joint effects on data quality.

4. EVALUATING SURVEY QUESTION LENGTH

There are a number of different approaches to the evaluation of the attributes of questions that affect data quality (e.g., see Madans et al. 2011). Indeed, as already noted, there is a large literature that makes an effort to provide practical guidelines for the "best practices" of question and questionnaire design. Many of these approaches use subjective criteria, and rarely do they use rigorous methods for defining the desirable attributes of questions. Sudman and Bradburn (1974) were pioneers in their effort to quantify the "response effects" of various question forms. More recently, several efforts have been made to specify an empirical criteria of data quality—for example, using the MTMM approach to reliability and validity assessment, or the use of longitudinal methods of reliability assessment (see Alwin 1992, 2007; Alwin and Krosnick 1991; Andrews 1984; Saris and Andrews 1991; Saris and Gallhofer 2007; Saris and van Meurs 1990; Scherpenzeel 1995; Scherpenzeel and Saris 1997).

In this article we use the concept of *measurement reliability* as a criterion for evaluating the quality of survey data, and we reevaluate the question of the relationship between question length and reliability of measurement. Reliability refers to consistency of measurement—it is a *sine qua non* of scientific research (see Alwin 2005, 2010). It is typically conceived of as the absence of "errors of measurement" or the obverse of unreliability. Operationally, the psychometric concept of reliability refers to the correlational consistency "between two efforts to measure the same variable, using maximally similar measurements, and independent of any true change in the quantity being measured" (see Campbell and Fiske 1959; Lord and Novick 1968). This analysis builds on the previous analysis of this issue by Alwin (2007) and is based on a reexamination and reanalysis of reliability data assembled by that project.

The concept of reliability has been applied to survey measurement previously (e.g., Alwin 1989, 1992, 2007, 2010; Alwin and Krosnick

1991; Marquis and Marquis 1977), and it has proved useful as a measure of data quality (Biemer et al. 1991; Groves 1989); however, there is a reluctance on the part of many survey methods experts to evaluate questions in terms of their reliability (e.g., see Krosnick and Presser 2010; Schaeffer and Dykema 2011). In general, the psychometric approach defines the observed score as a function of a true score and an error score—that is, as $y = \tau + \varepsilon$, where $E(\varepsilon) = E(\tau\varepsilon) = 0$ and $E(\tau) = E(y)$. The idea of a true value may be difficult for some analysts to accept (see Lord and Novick 1968:27), but it follows from this classical true score theory (CTST) model that the sample estimate of *reliability* is the squared correlation between observed and true scores (i.e., $\rho^2_{y\tau}$), which equals the ratio of true variance to the observed variance (i.e., σ^2_τ/σ^2_y), or the proportion of the observed variance that is true variance. The challenge is to design surveys that will produce a valid estimate of this ratio (see Section 5.2).

5. RESEARCH METHODS

The purpose of this article is to present a more detailed analysis of the issue of the linkage between question length and reliability of measurement on the basis of a reanalysis of the data assembled in Alwin (2007:202–10), a project dealing with the relationships of various attributes of questions and the reliability of measurement. Here we provide a more thorough investigation of this topic, examining the relationship between question length and the reliability of measurement, controlling for question content, question context, and length of unit (series and battery) introductions. In this section we discuss the source of the data on which the present analysis is based, the methods we use to estimate the reliability of measurement, and our strategy for analyzing these data.

5.1. *Samples and Data*

Our study design requires the use of large-scale panel studies that are representative of known populations, with a minimum of three waves of measurement separated by two-year reinterview intervals. Questions were selected for use only if they were exactly replicated (exact wording, response categories, mode of interviewing, etc.) across the three waves and if the underlying variable measured was continuous (rather than categorical) in nature. Specifically, this research is based on six nationally (or regionally) representative panel surveys of the U.S. population, all involving probability samples and all using face-to-face interviews, as shown in Table 1. These data sets are as follows: (1) the 1956, 1958, and 1960 National Election Study (NES) panel; (2) the 1972, 1974, and 1976 NES panel; (3) the 1992, 1994, and 1996 NES panel; (4) the 1986, 1989, and 1994 Americans' Changing Lives panel study; (5) the Study of American Families (Detroit Area) panel study of mothers; and (6) the Study of American Families (Detroit Area) panel study of children (see Alwin 2007:119–22). These selection criteria yielded 426 self-report and proxy-report questions. Table 1 presents descriptive information on these six panel studies (see Alwin 2007:118–22). This table presents the total sample sizes of these studies, along with the number of cases with data present at all three waves of the panel (i.e., listwise cases).

One of the main advantages of the reinterview or panel design using long reinterview intervals is that under appropriate circumstances, it is possible to eliminate the confounding of the systematic and random error components. To address the question of stable components of error, the panel survey must deal with the problem of memory, because in the panel design, by definition, measurement is repeated. So, although this overcomes one limitation of cross-sectional surveys—namely, the failure to meet the assumption of the independence of errors—it presents problems if respondents can remember what they said in a previous interview and are motivated to provide consistent responses (Moser and Kalton 1972). Estimation of reliability from reinterview designs makes sense only if we can rule out memory as a factor in the covariance of measures over time, and thus the occasions of measurement must be separated by sufficient periods of time to rule out the operation of memory. In cases where the remeasurement interval is insufficiently large to permit appropriate estimation of the reliability of the data, the estimate of the amount of reliability will most likely be inflated (see Alwin 1989, 1992; Alwin and Krosnick 1991), and the results of these studies suggest that longer remeasurement intervals, such as those used here, are highly desirable.

As noted, we include survey measures of continuous variables only, and within this class of variables, we implement estimates of reliability that are independent of scale properties of the observed measures, which may be dichotomous, polytomous-ordinal, or interval. In each of

Table 1. Sources of Data for Question-specific Estimates of Reliability and Attributes of Questions ($n = 426$)

Panel Studies	Population Sampled	Acronym	Total Sample	Listwise Sample	Number of Measures
1956, 1958, and 1960 NES panel	National household population	NES60s	2,529	1,132	42
1972, 1974, and 1976 NES panel	National household population	NES70s	4,455	1,296	100
1992, 1994, and 1996 NES panel	National household population	NES90s	2,439	597	98
ACL panel	National household population	ACL	3,617	2,223	86
Study of American Families, mother panel	Detroit area, mothers of 1961 births	SAF-Mo	1,113	879	46
Study of American Families, children panel	Detroit area, 1961 births	SAF-Ch	1,113	875	54
Total					426

Source: Alwin (2007).

Note: ACL = Americans' Changing Lives; NES = National Election Study.

these cases, the analysis uses a different estimate of the covariance structure of the observed data, but the model for reliability is the same. That is, when the variables are dichotomies, the appropriate covariance structure used in reliability estimation is based on tetrachoric correlations (Jöreskog 1990, 1994; Muthén 1984); when the variables are polytomous-ordinal, the appropriate covariance structure is either the polychoric correlation matrix or the asymptotic covariance matrix based on polychoric correlations; and when the variables can be assumed to be interval, ordinary Pearson-based correlations and covariance structures for the observed data are used (Brown 1989; Jöreskog 1990, 1994; Lee, Poon, and Bentler 1990; Muthén 1984). As noted, all of these models assume that the latent variable is continuous.

5.2. *Methods of Reliability Estimation*

Following Campbell and Fiske's (1959) famous definition of reliability as the "agreement between two efforts to measure the same thing, using maximally similar methods" (p. 83), the concept of reliability is often conceptually defined in terms of the consistency of measurement. This is an appropriate characterization, indicating the extent to which "measurement remains constant as it is repeated under conditions taken to be constant" (see Kaplan 1964:200). The key idea here is expressed by Lord and Novick (1968) in their classical statement of true score theory, wherein they state that "the correlation between truly parallel measurements taken in such a way that the person's true score does not change between them is often called *the coefficient of precision*" (p. 134). In this case, the only source contributing to measurement error is the unreliability or imprecision of measurement. The assumption here, as is true in the case of cross-sectional designs, is that "if a measurement were taken twice and *if no practice, fatigue, memory, or other factor* [emphasis added] affected repeated measurements," the correlation between the measures reflects the precision, or reliability, of measurement (Lord and Novick 1968:134). In practical situations in which there are in fact practice effects, fatigue, memory, or other spurious factors contributing to the correlation between repeated measures, the simple idea of the correlation between Y_1 and Y_2 is not the appropriate estimate of reliability. Indeed, in survey interviews of the type commonly used it would be almost impossible to ask the same question twice without memory or other factors contributing to the correlation of repeated measures. Thus, in general, asking the same question twice within the same interview would be the incorrect design for estimating the reliability of measuring the trait *T*, because the two observations are likely not independent.

It can be argued that for purposes of assessing the reliability of survey data, longitudinal data provide an optimal design (see Alwin 2007). Indeed, the idea of replication of questions in panel studies as a way of getting at measurement consistency has been present in the literature for decades, the idea of "test-retest correlations" as an estimate of reliability being the principle example of a longitudinal approach. The limitations of the test-retest design are well known, but they can be overcome by incorporating three or more waves of data separated by lengthy periods of time (see Alwin 2007:96–116). The multiple-wave reinterview design discussed in this article goes well beyond the traditional test-retest design (see Moser and Kalton 1972:353–54), and specifically by using models that permit change in the underlying true score (using the quasi-Markov simplex approach) allows us to overcome one of the key limitations of the test-test design (see, e.g., Heise 1969; Wiley and Wiley 1970). The literature discussing the advantages of the quasi-Markov simplex approach for separating unreliability from true change is extensive (see Appendix A in the online journal).[5]

Through the use of design strategies with relatively distant reinterview intervals (e.g., two-year intervals), the problem of consistency due to retest effects or memory can be remedied, or at least minimized. There are two main advantages of the reinterview design for reliability estimation. First, the estimate of reliability obtained includes all reliable sources of variation in the measure, both common and specific variance. Second, under appropriate circumstances it is possible to eliminate the confounding of the systematic error component discussed earlier, if systematic components of error are not stable over time. To address the question of stable components of error, the panel survey must deal with the problem of memory, because in the panel design, by definition, measurement is repeated. So, although this overcomes one limitation of cross-sectional surveys, it presents problems if respondents can remember what they say and are motivated to provide consistent responses. If reinterviews are spread over months or years, this can help rule out sources of bias that occur in cross-sectional studies. Given the difficulty of estimating memory functions, estimation of reliability from reinterview designs makes sense only if we can rule out memory as a factor in the covariance of measures over time, and thus, the occasions of measurement must be separated by sufficient periods of time to rule out the operation of memory.

The model used here falls into a class of autoregressive or quasi-Markov simplex models that specifies two structural equations for a set of p over-time measures of a given variable Y (where $t = 1, 2, \ldots, p$) as follows:

$$Y_t = T_t + E_t \tag{1}$$

and

$$T = \beta_{t,t-1} T_{t-1} + Z_t. \tag{2}$$

Equation (1) represents a set of measurement assumptions indicating (1) that over-time measures are assumed to be τ-equivalent, except for true score change and (2) that measurement error is random (see Alwin 1989, 2007, 2011; Heise 1969; Jöreskog 1970; Wiley and Wiley 1970). Equation (2) specifies the causal processes involved in change of the latent variable over time. A formal statement of the model is provided in Appendix A in the online journal.

It is important to note that this model assumes that the latent variable will change over time and that it follows a Markovian process in which the distribution of the true variables at time t is dependent only on the distribution at time $t-1$ and not directly dependent on distributions of the variable at earlier times. If these assumptions do not hold, then this type of simplex model may not be appropriate. In order to estimate such models, it is necessary to make some assumptions regarding the measurement error structures and the nature of the true change processes underlying the measures. All estimation strategies available for such three-wave data require a lag-1 assumption regarding the nature of the true change. This assumption in general seems a reasonable one, but erroneous results can result if it is violated. The various approaches differ in their assumptions about measurement error. One approach assumes equal reliabilities over occasions of measurement (Heise 1969). This is often a realistic and useful assumption, especially when the process is not in dynamic equilibrium, that is, when the observed variances vary with time. Another approach to estimating the parameters of the above model is to assume constant measurement error variances rather than constant reliabilities (Wiley and Wiley 1970). Where

$P = 3$, either model is just-identified, and where $P > 3$, both models are overidentified with degrees of freedom equal to $.5[P(P+1)] - 2P$. The four-wave model has two degrees of freedom, which can be used to perform likelihood-ratio tests of the fit of the model.

Wiley and Wiley (1970) showed that by invoking the assumption that the measurement error variances are equal over occasions of measurement, the $P = 3$ model is just-identified, and parameter estimates can be defined. They suggested that measurement error variance is "best conceived as a property of the measuring instrument itself and not of the population to which it is administered" (p. 112). Following this reasoning, we might expect that the properties of our measuring instrument would be invariant over occasions of measurement and that such an assumption would be appropriate. Following the CTST model for reliability, the reliability for the observed score Y_t is the ratio of the observed variance—that is, $\mathrm{VAR}(T_t)/\mathrm{VAR}(Y_t)$—and this model permits the calculation of distinct reliabilities at each occasion of measurement using the above estimates of $\mathrm{VAR}(E_t)$ and $\mathrm{VAR}(T_t)$ (see Wiley and Wiley 1970; see also Alwin 2007:109).

To summarize, the three-wave model is just-identified—that is, there are no overidentifying restrictions that allow for the possibility of testing the model against a null model of interest. Where $P > 3$, the simplex model is overidentified, and a test of the model (with degrees of freedom equal to $.5[P(P+1)] - 2P$) is possible. The four-wave model has two degrees of freedom, which can be used to perform likelihood-ratio tests of the fit of the model. We restricted the present analysis to only three waves because the panel studies on which we draw include only three waves. Note that because the models we use are just-identified, standard errors for the parameter estimates cannot be computed. Further testing of these models using more than three waves is essential for generalization of the findings reported here. On the other hand, although it is important to use multiwave panel data in this case, it is also true that more waves add complexities that must be dealt with. It is important to be able to deal with attrition, for example, and more waves add to the problems of missing data.

5.3. *Analysis Strategy*

In the following analysis we use several measures of the attributes of the questions and use these to predict the estimated reliability of the questions. First, our primary explanatory variable is question length (i.e., the number of words in a question), and we examine several aspects of question length, including the number of words in both the "prequestion text" and the actual question. In our previous discussion of the reliability of questions in series and batteries, we considered the role of the presence and length of subunit introductions on estimates of measurement reliability, finding that it had modest effects (see Alwin 2007:208). Here we consider the number of words appearing in the text for a given question following any prequestion text that might be included as an introduction. Our analysis focuses on both the effects of question length and introduction length as independent factors. In both cases, we express question and introduction length in units of 10 (or fractions thereof) and center this measure by expressing it as a deviation from its mean.

Second, we use a measure of question content, an important predictor of question reliability, to control for this source of variation. In this case, question content refers to whether the content measured is a fact (content that can potentially be verified by consulting other sources) versus a nonfact such as beliefs, values, attitudes, or self-descriptions—content that is primarily subjective and cannot be verified by consulting other sources.

Third, we use a three-category variable that we refer to as question context, which classifies questions according to the architecture of survey questionnaires: (1) stand-alone questions, which do not bear any particular topical relationship with adjacent questions; (2) questions that are a part of a series of questions that all focus on the same specific topical content; and (3) questions that appear in batteries focusing on the same or similar subject matter, and more specifically use the identical response format (for some examples, see Alwin 2007:205–207). (Appendix C in the online journal presents illustrative examples of these types of contexts from an actual survey questionnaire used here.) In our analyses of the total sample of questions, we use a set of dummy variables to represent this variable and omit the category of "questions in batteries" to deal with the redundancy.

Finally, in subclass regressions for survey context, we additionally consider as a separate factor the number of words in the introduction to a series or battery.

Table 2. Regression of Reliability Estimates on Length of Question and Attributes of Question Content and Question Context

	Model				
Predictor	1	2	3	4	5
Intercept	.668	.636	.625	.620	.600
QL	.000	−.001	−.006	−.004	−.006
Fact vs. nonfact		.173***		.117***	.092***
SA			.228***	.146***	.263***
In series			.098***	.060***	.156***
QL × SA					−.040***
QL × In Series					−.034***
R^2	.000	.158	.149	.200	.239
Number of cases	426	426	426	426	426

Note: QL = question length; SA = stand-alone.
***$p \leq .001$.

6. RESULTS

In Table 2 we present a set of regression models in which we predict the reliability of survey measures from several question characteristics using the entire database of 426 questions. These models use predictor variables for question length, a dummy variable representing question content (nonfact is the omitted category), a set of two dummy variables representing question context (questions in batteries form the omitted category), and two interaction terms expressing the interaction between question length and questions in the stand-alone category and questions in series.[6] These interaction terms are important, in that as specified they express the role of question length within the two categories of stand-alone questions and questions in series. There are two key differences between the following results and Alwin's (2007) presentation of these data. First, we control here for question content when assessing the combined effects of question length and reliability of measurement; second, we separate the effects of question length and the length of introductions to series and batteries.

In this table, model 1 contains question length as the sole predictor; this variable is statistically independent of question reliability in the total set of 426 measures (see also Figure 1). The second model adds question content (fact vs. nonfact), as described above, which is an important predictor of reliability; the coefficient of .173 indicates that

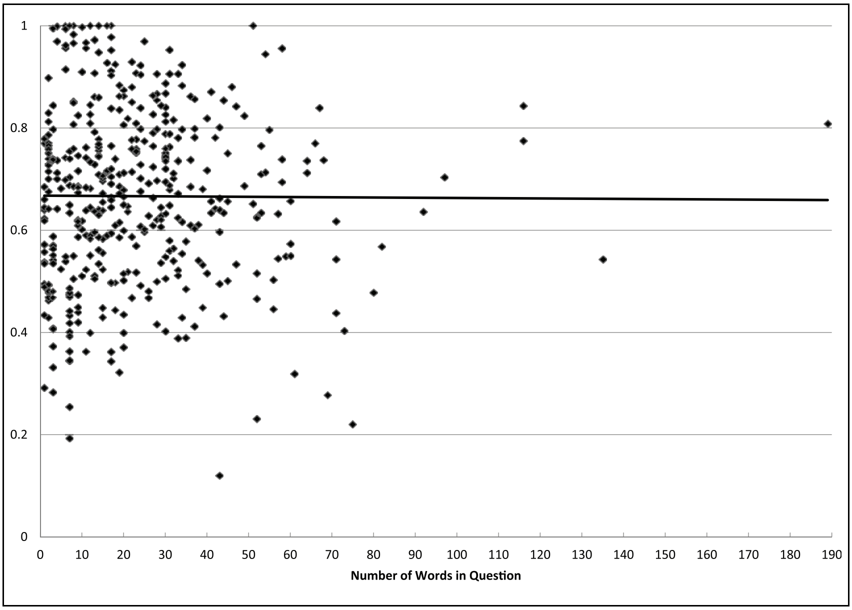

Figure 1. Regression of reliability estimates on question length for all factual and nonfactual questions ($n = 426$).

there is a predicted difference of this magnitude on average between questions that measure facts and those that measure nonfacts. This variable alone accounts for about 15 percent of the variation in reliability among survey questions. Model 3 adds our classification of survey context (removing the measure of question content), and model 4 includes both survey content and survey context. In model 3, the reference category (or omitted category) is the category of questions in batteries, and thus the regression coefficients for the dummy variables for "stand-alone" and "in series" refer to the differences in mean reliability for these categories relative to questions in batteries. The R^2 value for model 3 indicates that the survey context classification also by itself accounts for about 15 percent of the variation in reliability, when entered alone. Stand-alone questions are the most reliable, followed by questions in series; questions in batteries (which have an average reliability of .624) are the least reliable. We present the results of several alternative specifications of the model in the supplementary tables provided in Appendix B in the online journal.[7]

Table 3. Regression of Reliability Estimates on Length of Question and Question Content: Stand-alone Questions

	Model	
Predictors	1	2
Intercept	.956	.763
Question length	−.045***	−.023†
Fact vs. nonfact		.190**
R^2	.381	.556
Number of cases	30	30

†$p \leq .10$. **$p \leq .01$. ***$p \leq .001$.

Returning to the results in Table 2, survey content and survey context are clearly not independent, as indicated by the results of model 4: factual questions are more likely to be measured in a stand-alone format or in a series, whereas nonfacts are more likely to be placed in series and batteries. Together these two sets of factors account for about 20 percent of the variance, a rather remarkable result. In other words, by knowing only two things about survey questions—what they are measuring and the placement of the question in the organizational context of the questionnaire—we can account for roughly one fifth of the variation in reliability of measurement. The addition of the interaction terms between question length and question context, which allow the effects of question length to vary by question context, significantly improves the prediction of question reliability—to about 25 percent of the variation. These results indicate that although there is no relationship between question length and reliability among questions included in batteries (note that this effect is represented by the effects of question length in model 5), there is a significant decline in reliability linked to question length for both stand-alone questions and questions in series.[8]

To explore these relationships further, we examine the patterns of association between reliability and question length separately for stand-alone questions and questions in series. Table 3 presents results for the 30 stand-alone questions, and although this represents a relatively small pool of survey questions, these results are very revealing. As noted above, among stand-alone questions, there is a modest decline in reliability as the length of the question increases (see Figure 2). Very short questions are highly reliable, but as the length of the question increases,

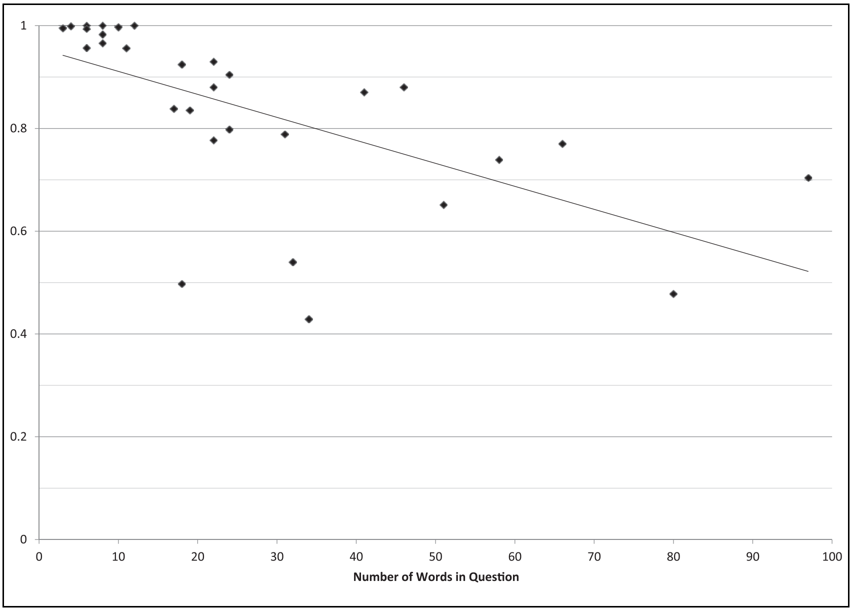

Figure 2. Regression of reliability estimates on question length for stand-alone questions ($n = 30$).

reliability of responses suffers. This relationship is, however, due almost entirely to the fact that stand-alone questions are much more likely to be measuring factual content: when this variable is added to the equation (see model 2 in Table 3), the relationship with question length is almost entirely removed. It is noteworthy, however, that there is a very slight negative decline in reliability due to question length. This pattern was stronger (statistically speaking) in Table 2, in which the overall test of this effect had the benefit of greater power due to the overall larger sample size.

When we examine these relationships separately for questions in series (see Table 4 and Figure 3), the same patterns emerge as those presented in Table 2—greater question length suppresses reliability of measurement. The effects of question length are not removed by controlling for survey content, and it is not possible to argue on the basis of these results that the effect of question length is due to its confounding with question content. As above, the major factor affecting measurement reliability is survey content, indicating that among questions in series,

Table 4. Regression of Reliability Estimates on Length of Question and Question Content: Questions in Series

	Model					
Predictor	1	2	3	4	5	6
Intercept	.799	.762	.759	.832	.738	.755
Question length	−.033***	−.029***	−.032***	−.037**	−.025*	−.024*
Fact vs. nonfact		.079***	.078**	.063***	.085***	.070***
First in series			.048†			
Series introduction length						−.009*
R^2	.098	.150	.166	.201	.144	.175
Number of cases	177	177	177	42	135	135

Note: Model 4 is for those cases first in series. Models 5 and 6 are for those cases second or later in series.

$^{\dagger}p \leq .10$. $^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

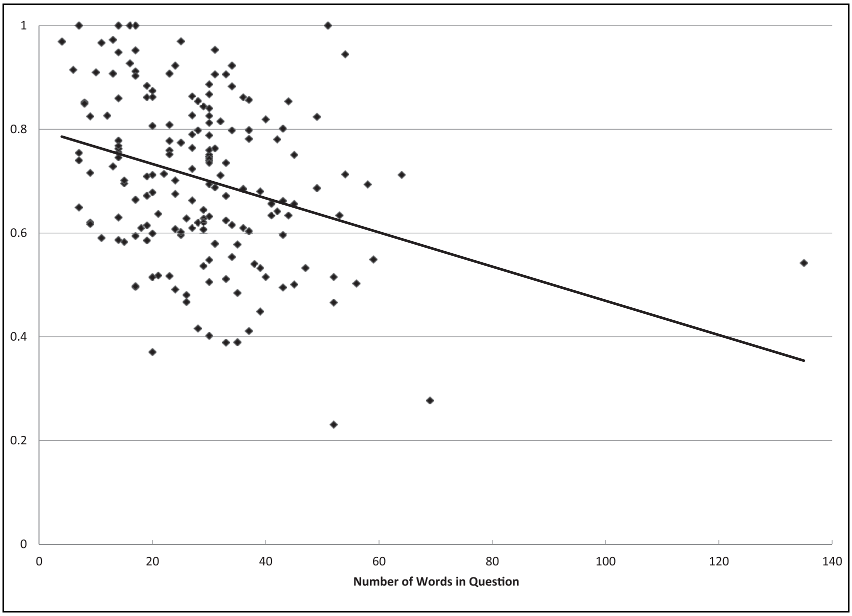

Figure 3. Regression of reliability estimates on question length for question in series (n = 177).

factual questions exceed nonfactual ones by about .07 to .08 in reliability, depending on which subset of questions we consider (compare model 2 in Table 3 and model 2 in Table 4). There is a slight advantage in reliability for the first questions in a series (see model 3 in Table 4), but this effect is marginally significant.

Note that in Table 4, the first three models apply to the full set of questions in the pool that were placed in series, whereas the later models apply to specific subsets. As already noted, among questions in series, there is a significant reduction in reliability associated with greater question length (see model 1), which is not removed when controlling for question content (see model 2). Model 3 tests for whether being the first question in a series improves reliability, and there is a marginally ($p < .10$) significant improvement. We tested for statistical interaction between being first in the series and question length, which was not significant ($p = .4326$). Note that model 4 is based only on those 42 questions that were the first items in the series. For this subsample of questions, the negative effect of question length is slightly enhanced in this subsample. Models 5 and 6 pertain to the subset of 135 questions that were second or later in the series, and there continues to be a significant effect of question length. The advantages that accrue to shorter questions and measures of facts do not depend on whether the question is the first one in the series or the second or later; compare models 4 and 5 in Table 4, for example. When we compare the effect of the length of the introductions with the questions in series (see model 6), we find a slight depressing effect on reliability of longer introductions; this effect, however, is quite small.

Finally, Table 5 presents the results of an analysis for questions in batteries. The first two models in this table were estimated on the basis of all of the questions in the pool that were placed with batteries; models 3 and 4 pertain to subsets of these questions in batteries. Model 3 was conducted on the questions that were the first questions in the batteries, whereas model 4 applies to the questions appearing second or later in the batteries. In these models, we have not included the survey content variable, in that virtually all questions in batteries measure nonfactual content. Consistent with prior results (Alwin 2007), there appears to be little evidence that question length for questions in batteries has any bearing on their reliability. We tested for statistical interaction on the basis of whether the question was first in a battery and question length in model 2, which was nonsignificant ($p = .3694$). Questions in batteries

Table 5. Regression of Reliability Estimates on Length of Question and Other Content: Questions in Batteries

	Model			
Predictor	1	2	3	4
Intercept	.600	.601	.647	.582
Question length	.007	.004	.000	.015
First in batteries		.026		
Batteries introduction length				.002
R^2	.012	.014	.000	.000
Number of cases	219	219	41	178

Note: Model 3 is for those cases first in batteries; model 4 is for those cases second or later in batteries.

are overall less reliable (e.g., compare the intercepts across the models in Table 5 with Table 4), but there is no association between question length and reliability. We consider the possible explanations for this in Section 7.

Question length is related to reliability for questions in series and to a lesser extent in stand-alone questions once content is controlled for. Given the importance of content, the question was raised about whether question length interacts with content—that is, are there differences in the relationship of question length and reliability by survey content (fact vs. nonfact)?[9] We explored this hypothesis and found that there was an interaction between content and question length, but it was driven entirely by question context. Fact-based questions are asked only in stand-alone questions or within a series of questions, and once question context is controlled, the relationship between content and question length disappears. The overall results suggest that content and context are clearly predictors of reliability, and question length adds to the prediction only within specific contexts, questions in series, and to some extent, stand-alone questions.

7. DISCUSSION

There is a growing literature that addresses the practical question of the desirable length of questions in surveys. There is no consensus on this issue and a mixture of opinions, few of which are grounded in empirical assessments. The present research builds on work that established an empirical regularity of a relationship between question length and data quality in some survey contexts, as assessed by measurement reliability. The focus of this research entertained the modest objective of the possible confounding in those results of question length with question content, while controlling for question context (i.e., the placement of the question within the organization of the questionnaire). There is no question that there may be substantial confounding between question length and the substantive focus of the question, wherein subjective content tends to be measured using longer questions. The present research has focused explicitly on the confounding of question length and survey content, net of survey context.

This research assesses question content using the long-standing focus in the survey methods literature on objective and subjective questions, operationalized here in terms of questions seeking factual versus nonfactual information. Prior research suggests that factual (or objective) material can be more precisely measured than content that is essentially subjective (Alwin 1989, 2007; Kalton and Schuman 1982; Schuman and Kalton 1985), although there is considerable overlap. Few survey questions are perfectly reliable—but the typical factual question can be shown to be substantially more reliably measured on average than the typical nonfactual question. Our more detailed examination of this issue in the present article confirms the strong effects of question content (fact vs. nonfact) and question context (stand-alone questions, questions in series, and questions in batteries) as important predictors of reliability, together accounting for some 20 percent of the variation in measurement reliability.

Within the constraints of the purpose of the survey, one element of survey question writing to which the majority of (but not all) researchers subscribe is that questions should be as short as possible, although there are opposing views. The question raised in the present research is whether the length of questions (and for questions in series and batteries, the length of introductory text) produces any significant decrement to reliability of measurement. Bear in mind our findings are conditional on our parameters for question inclusion (see Section 5.1 for qualifying criteria). Certainly, there may be limitations to any generalization concerning the practical advantages and disadvantages of any particular study, but it is important to address the practical conclusions of the present research. The major practical implications are that, exclusive of questions in batteries, other things being equal, shorter questions are more reliable. In the case of questions in batteries, the concept of question length, apart from the length of the introduction, is somewhat ambiguous. It will therefore be valuable for future research on the effects of question length to introduce further clarification of the types of questions we have considered to be part of questionnaire batteries (see Alwin 2007:205–207).

The overarching practical consideration in the case of batteries, then, is not the length of the questions but whether to use them at all. Consistent with prior research (e.g., Alwin 2007; Andrews 1984), our results provide strong support for the view that questions in what we referred to as a "topical series" are less reliable than "stand-alone questions" (at least among factual questions) and questions in "batteries" are less reliable than questions in series (among nonfactual questions) (Alwin 2007:171–72). Perhaps the best explanation of this phenomenon is that the same factors motivating the researcher to group questions together—contextual similarity—are the same factors that promote measurement errors (see Andrews 1984:431). Similarity of question content and response formats may actually distract the respondent from fully considering what information is being asked for, and this may reduce the respondent's attention to the specificity of questions (i.e., they may increase his or her tendency to "satisfice"; Krosnick and Alwin 1987). Thus, measurement errors may be generated in response to the "efficiency" features of the questionnaire, and unfortunately, as Andrews (1984:431) concluded, it appears that the respondents may also be more likely to "streamline" their answers when the investigator "streamlines" the questionnaire.

8. CONCLUSION

Our research concludes that the consideration of the length of questions adds to the understanding of levels of measurement error associated with question attributes, but the results must be understood in terms of the interaction of question length and question context. That is, as indicated above, the question length of stand-alone questions and questions in a topical series are found to have a negative effect on measurement reliability. The length of questions in batteries, however, reveals no relationship to reliability. This may be due in part to the fact that questions in batteries are quite a lot shorter; the typical nonfactual question in batteries has a question length of 12 words (see Alwin 2007:204).

This is because often the actual question stimulus is just a few words, given the existence of a lengthy introduction to the battery that explains how the respondent should use the rating scale. One of the important contributions of the present research is its emphasis on the separation of the length of questions from the length of introductions to series and batteries. Although it is true that the length of a question is not independent of the nature and length of the unit (series or battery) introduction, neither the length of the battery introduction nor the length of questions in batteries have any measurable effect on reliability of measurement. Indeed, the very existence of lengthy introductions in batteries of questions may promote greater reliability.

There are admitted limitations related to the present research. First, as noted at the outset, we focus on only one aspect of data quality, and given the limitations of the research design, we cannot examine aspects of data quality, such as bias and nonresponse (see Groves 1989). Similarly, there are other aspects of question complexity that are beyond the scope of the present article, which may limit its applicability to assessing question characteristics. Prior research has found that other measures of complexity, such as the reading level of the question, were predictive of the comprehension of the question. Ultimately, question comprehension is an attribute of the respondent and not necessarily an objective characteristic of the question. Clearly, the relationship between question length, complexity, and other aspects of questionnaires is an important set of issues, and the narrow focus on the issues of question length leaves other issues unaddressed. The assumption is that improved comprehension translates into greater reliability, but what is unclear is the role that question length contributes to greater comprehension. There are ways in which the issue of comprehension can be investigated in future research using cognitive interviews in laboratory settings.

Second, as noted, our estimates of reliability are limited to three-wave panels, which are just-identified and assume a lag-1 structure for the latent variables. Our models also indicate that the error variances are assumed to represent random error and that stable nonrandom sources of error are included in the underlying latent trait variable. In our defense, it should be noted that multiple-wave ($P > 3$) studies are generally unavailable, which limits the possibilities of developing overidentified models. Future research will need to address these issues using the expanding opportunities in longitudinal studies that meet the design requirements of this approach. Also as a minor but still important issue, we note that the present article addresses only questions used in the United States (and more specifically only questions from the University of Michigan surveys). One critical reviewer commented that the questions in the United States are in general “considerably longer than questions in Europe,” which, if true, admittedly may influence the generalizability of the results of the present article. We therefore caution the reader to not overgeneralize the findings of the present study.

Finally, we should note we can rule out the broad features of question content as an explanation of the relationship between question length and reliability. As we stressed at the beginning of the article, the issue of question length may be sidestepping an important set of issues, namely, question clarity and comprehension (Converse and Presser 1986). Although it is an important practical concern of many survey methodologists, it is probably the case that some people would think that question length and introduction length *per se* are not the issue. However, clarity and comprehension ultimately reside in the mind of the individual respondent and are not specifically linked to the properties of the question itself. Question length can be measured objectively, and if a variable such as this is related to increased errors of measurement, we should attempt to understand why. One problem with long questions is that, if there are a lot of them, this may lead to satisficing on the part of the respondent (see Krosnick and Alwin 1987), and this may be likely to hurt reliability. But again this is beyond the potential question-level analysis that is possible here. Moreover, satisficing is an interpretative tool for respondent behavior, and as far as we know, no one has figured out a way of measuring it within the survey context. There are some avenues, however, that may be pursued in future studies as a way of explaining the effects of question length. For example, as we mentioned earlier, if a question gets to be too complex syntactically, it may lead to poor comprehension, which again is likely to reduce reliability. The relation between question length and question complexity is a topic that should be pursued in future work. On the other hand, a long question may be long because it includes a definition for a vague or unfamiliar term and that may actually improve comprehension (and reliability). Or, a long question may incorporate lots of memory cues, and that may increase reliability (or at least accuracy). Each of these hypotheses suggests interactions that could be incorporated into future studies that would try to account for the present findings. For example, if question length is associated with syntactic complexity, then it may have fewer (or different) effects on highly educated respondents. If length is associated with more memory cues for factual questions involving memory retrieval, then it may have positive effects for such questions. There are a number of ways in which hypotheses may be developed to afford better explanations for the findings presented here.

Acknowledgments

This article was submitted and reviewed during the editorship of Tim Futing Liao, editor-in-chief of *Sociological Methodology* through 2015. Decision making on this article was made entirely during his editorship. Tim Liao, previous reviewers, and Aaron Maitland provided valuable comments and suggestions on earlier versions. We acknowledge the research assistance of Mona Ostrowski, Audrey Jackson, Alyson Otto, and Judy Bowes.

Notes

1. This quotation is attributed to Antoine de Saint-Exupéry (1900–1944), noted French writer and aviator.
2. Although we believe that Holbrook et al.'s (2006) research helps account for the mechanisms by which question length may affect the quality of survey data, any exploration here of the complexity of questions, such as reading level and comprehensiveness, is beyond the scope of the present research.
3. Results reported by Andrews (1984) and Scherpenzeel and Saris (1997) are somewhat confusing because in the typical survey, questions by themselves do not have introductions. On the other hand, series of questions, or batteries, or entire sections, do typically have introductions (see below). We see the introduction to series and/or batteries of questions to be a separate topic conceptually from that of question length, and we therefore distinguish between question length and the length of introductions to organizational units larger than the question.
4. The distinction used here between "fact" and "nonfact" is derived from early work in survey methods (e.g., Kalton and Schuman 1982; Schuman and Kalton 1985; Turner and Martin 1984). The distinction was further used in Alwin's (1989, 2007) work, and it has been proved to be easily coded. Facts are defined as objective information regarding the respondent or members of the household, which can be verified against objective records—for example, information on the respondent's personal characteristics, such as the date of birth, amount of schooling, amount of family income, and the timing, duration, and frequencies of certain behaviors (Alwin 2007:123; see also Alwin 2007:157, Table 7.4). Nonfacts include beliefs, attitudes, and values that are a matter of personal judgment for which no objective information exists; nonfacts also include self-descriptions—that is, subjective assessments or evaluations of the state of the respondent within certain domains (see Alwin 2007:123–24, Table 6.1). In the present research, agreement was achieved among four investigators, and little ambiguity exists for the vast majority of

questions. It is worth noting that this dichotomous predictor may not always be adequate, and that the nonfact category is normally broken down further into attitudes, beliefs, values, self-appraisals, and self-evaluations (see Alwin 2007:153–62). In the present research, the fact-nonfact distinction was deemed adequate given the goals of the research.

5. Appendix A in the online journal presents an extended discussion of the simplex model as applied to multiwave panel data. In the three-wave case, the parameters of the model are just-identified and can easily be estimated by hand given the correlations among the variables. As noted in the text, for ordinal variables we base our estimates on polychoric correlations, and for continuous variables we use Pearson correlations. The reader is invited to contact us with any additional questions regarding the estimation models used to obtain the correlations involved.
6. The reader is referred to introductory statistics materials that cover the use of dummy variables and interaction terms in regression analysis (e.g., Hardy 1993).
7. See Appendix B in the online journal for a detailed discussion of these alternative models.
8. We concede that statistical tests on differences in attributes of questions measured by the predictor variables are not technically appropriate, because the questions have neither been randomly selected from some known universe of questions, nor are they independent in a sampling sense. Nonetheless, we present information from statistical tests as a qualitative measure of the relative magnitude of a particular relationship, not as a basis for generalizing to some known universe of questions.
9. One reviewer made the case for considering the interaction of content with question length. The net result is that there are no interactions of question length with survey content (fact vs. nonfact) in predicting reliability.

References

Achen, Christopher H. 1975. “Mass Political Attitudes and the Survey Response.” *American Political Science Review* 69:1218–31.

Alwin, Duane F. 1989. “Problems in the Estimation and Interpretation of the Reliability of Survey Data.” *Quality and Quantity* 23(3):277–331.

Alwin, Duane F. 1992. “Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement.” *Sociological Methodology* 22:83–118.

Alwin, Duane F. 2005. “Reliability.” Pp. 351–59 in *Encyclopedia of Social Measurement*, edited by K. Kempf-Leonard. New York: Academic Press.

Alwin, Duane F. 2007. *Margins of Error—A Study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley.

Alwin, Duane F. 2010. “How Good Is Survey Measurement? Assessing the Reliability and Validity of Survey Measures.” Pp. 405–34 in *Handbook of Survey Research*, edited by Peter V. Marsden and James D. Wright. Bingley, UK: Emerald Group.

Alwin, Duane F. 2011. “Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach.” Pp. 265–93 in *Question Evaluation Methods: Contributing to the Science of Data Quality*, edited by Jennifer Madans, Kristen Miller, Aaron Maitland, and Gordon Willis. Hoboken, NJ: John Wiley.

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(1):139–81.

Andrews, Frank M. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 48(2): 409–42.

Belson, William A. 1981. *The Design and Understanding of Survey Questions*. Aldershot, UK: Gower.

Biemer, Paul P., Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, eds. 1991. *Measurement Errors in Surveys*. New York: Wiley.

Bradburn, Norman M., and S. Sudman, and Associates. 1979. *Improving Interviewing Methods and Questionnaire Design: Response Effects to Threatening Questions in Survey Research*. San Francisco, CA: Jossey-Bass.

Brislin, Richard W. 1986. "The Wording and Translation of Research Instruments." Pp. 137–64 in *Field Methods in Cross-cultural Research*, edited by W. J. Lonner and J. W. Berry. Newbury Park, CA: Sage.

Brown, R. L. 1989. "Using Covariance Modeling for Estimating Reliability on Scales with Ordered Polytomous Variables." *Educational and Psychological Measurement* 49(2):385–98.

Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 6(1): 81–105.

Cannell, Charles F., Kent H. Marquis, and Andre Laurent. 1977. "A Summary of Studies of Interviewing Methodology." *Vital and Health Statistics*. Series 2, No. 69, March.

Cannell, Charles F., Peter V. Miller, and Lois Oksenberg. 1981. "Research on Interviewing Techniques." *Sociological Methodology* 12:389–437.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: Sage.

Fowler, Floyd J. 1992. "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly* 56(2):218–31.

Galton, Francis. 1893. *Inquiries into the Human Faculty and Its Development*. London: Macmillan.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley.

Hardy, Melissa A. 1993. *Regression with Dummy Variables: Quantitative Applications in the Social Sciences*. Newbury Park, CA: Sage.

Heise, David R. 1969. "Separating Reliability and Stability in Test-retest Correlation." *American Sociological Review* 34(1):93–191.

Holbrook, Allyson, Young Ik-Cho, and Timothy Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70(4):565–95.

Jöreskog, Karl G. 1970. "Estimating and Testing of Simplex Models." *British Journal of Mathematical and Statistical Psychology* 23:121–45.

Jöreskog, Karl G. 1990. "New Developments in LISREL: Analysis of Ordinal Variables Using Polychoric Correlations and Weighted Least Squares." *Quality and Quantity* 24(4):387–404.

Jöreskog, Karl G. 1994. "On the Estimation of Polychoric Correlations and Their Asymptotic Covariance Matrix." *Psychometrika* 59(3):381–89.

Kalton, Graham, and Howard Schuman. 1982. "The Effect of the Question on Survey Response: A Review." *Journal of the Royal Statistical Society* 145(1):42–73.

Kaplan, Abraham .1964. *The Conduct of Inquiry*. San Francisco, CA: Chandler.

Knauper, Barbel, Robert F. Belli, Daniel H. Hill, and A. Regula Herzog. 1997. "Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality." *Journal of Official Statistics* 13(2):181–99.

Krosnick, Jon A., and Duane F. Alwin. 1987. "Satisficing: A Strategy for Dealing with the Demands of Survey Questions." General Social Survey Technical Report No. 74; Methodological Report No. 46. Chicago: National Opinion Research Center.

Krosnick, Jon A., and Leandre R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." Pp. 141–64 in *Survey Measurement and Process Quality*, edited by Lars E. Lyberg, Paul P. Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: John Wiley.

Krosnick, Jon A., and Stanley Presser. 2010. "Question and Questionnaire Design" Pp. 263–313 in *Handbook of Survey Research*, edited by Peter V. Marsden and James D. Wright. Bingley, UK: Emerald Group.

Lee, Sik-Yum, Wai-Yin Poon, and Peter M. Bentler. 1990. "A Three-stage Estimation Procedure for Structural Equation Models with Polytomous Variables." *Psychometrika* 55(1):45–51.

Lord, Frederick M., and Melvin L. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Madans, Jennifer, Kristen Miller, Aaron Maitland, and Gordon Willis, eds. 2011. *Question Evaluation Methods: Contributing to the Science of Data Quality*. Hoboken, NJ: John Wiley.

Marquis, Kent H., Charles F. Cannell, and Andre Laurent. 1972. "Reporting Health Events in Household Interviews: Effects of Reinforcement, Question Length, and Reinterviews." *Vital and Health Statistics*. Series 2, No. 45.

Marquis, M. Susan, and Kent H. Marquis. 1977. *Survey Measurement Design and Evaluation Using Reliability Theory*. Santa Monica, CA: RAND.

Moser, Claus A., and Graham Kalton. 1972. *Survey Methods in Social Investigation*. 2nd ed. New York: Basic Books.

Muthén, Bengt O. 1984. "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators." *Psychometrika* 49(1):115–32.

Payne, Stanley L. 1951. *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.

Ruckmick, Christian A. 1930. "The Uses and Abuses of the Questionnaire Procedure." *Journal of Applied Psychology* 14(1):32–41.

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

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

Saris, Willem E., and A. van Meurs. 1990. *Evaluation of Measurement Instruments by Meta-analysis of Multitrait Multimethod Studies*. Amsterdam, the Netherlands: North-Holland.

Schaeffer, Nora Cate, and Jennifer Dykema. 2011. "Questions for Surveys: Current Trends and Future Directions." *Public Opinion Quarterly* 75(5):909–61.

Schaeffer, Nora Cate, and Stanley Presser. 2003. "The Science of Asking Questions." *Annual Review of Sociology* 29:65–88.

Scherpenzeel, Annette C. 1995. "A Question of Quality: Evaluating Survey Questions by Multitrait-multimethod Studies." PhD dissertation, Department of Methodology, University of Amsterdam.

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(3):341–83.

Schuman, Howard, and Graham Kalton. 1985. "Survey Methods." Pp. 634–97 in *Handbook of Social Psychology*, 3rd ed., edited by Gardner Lindzey and Eliott Aronson. New York: Random House.

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

Sudman, Seymour, and Norman M. Bradburn. 1974. *Response Effects in Surveys*. Chicago: Aldine.

Sudman, Seymour, and Norman N. Bradburn. 1982. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco, CA: Jossey-Bass.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass.

Tanur, Judith M., ed. 1992. *Questions about Questions—Inquiries into the Cognitive Bases of Surveys*. New York: Russell Sage.

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

Turner, Charles F., and Elizabeth Martin. 1984. *Surveying Subjective Phenomena*. Vol. 1. New York: Russell Sage.

van der Zouwen, Johannes. 1999. "An Assessment of the Difficulty of Questions Used in the ISSP-Questionnaires, the Clarity of Their Wording, and the Comparability of the Responses." *ZA-Information* 46(1):96–114.

Wiley, David E., and James A. Wiley. 1970. "The Estimation of Measurement Error in Panel Data." *American Sociological Review* 35(1):112–17.

Yan, Ting, and Roger Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience, and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22(1):51–68.

Author Biographies

Duane F. Alwin is the inaugural holder of the Tracy Winfree and Ted H. McCourtney Professorship in Sociology and Demography at Pennsylvania State University, where he directs the Center for Life Course and Longitudinal Studies. He is also emeritus research professor at the Survey Research Center, Institute for Social Research, and emeritus professor of sociology, University of Michigan, Ann Arbor. In addition to survey methodology, his research interests focus on the integration of demographic and developmental perspectives in the study of human lives. His current research work includes the study of socioeconomic inequality and health, parental child-rearing values, children's use of time, and social factors in cognitive aging. He has published extensively on these and related topics and is the recipient of numerous prestigious awards, grants, and special university honors.

Brett A. Beattie is a 2015 PhD recipient from the Department of Sociology and Criminology, Pennsylvania State University, specializing in family demography and quantitative methodology.