This study addresses the issue of the relation between the number of response categories used in survey questions and the quality of measurement. Several hypotheses, derived from relevant theory and research, are tested through a comparison between 7- and 11-category rating scales used in the 1978 Quality of Life Survey. One hypothesis derived from information theory, that rating scales with more response categories transmit a greater amount of information and are therefore inherently more precise in their measurement, is strongly supported. A second hypothesis, that questions with greater numbers of response categories are more vulnerable to systematic measurement errors or shared method variance, is rejected. This study supports the conclusion that questions with more categories are both more reliable and more valid.

Feeling Thermometers Versus 7-Point Scales

Which Are Better?

DUANE F. ALWIN
University of Michigan

Rating scales are widely used in sample surveys and social research generally (see, e.g., Sudman, Bradburn, and Schwarz 1996). For example, the typical *agree-disagree* Likert-type question with four or five response categories can be thought of as a rating scale with either explicit or implicit numeric quantities and associated verbal labels. Other types of rating scales are pervasive in social research, especially in the measurement of attitudes and other subjective phenomena (Jabine, Straf, Tanur, and Tourangeau 1984; Turner and Martin 1984). Despite their widespread use, there is little agreement about the optimal number of response categories or scale points to use. Arguments in favor of a particular form of rating scale

AUTHOR'S NOTE: *This article was prepared in part through the support of grants from the National Institute on Aging (R01-AG04743-06 and R01-AG09747-02). Data were obtained through the Inter-University Consortium for Political and Social Research. I acknowledge the research assistance of Lynn Dielman, Frank Mierzwa, and Rebecca Bahlibi.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 25 No. 3, February 1997 318-340

run the gamut from those advocating the use of 2-category response forms (McKennell 1974) to those favoring 3 categories (Benson 1971; Jacoby and Matell 1971; Lehmann and Hulbert 1972); to those supporting 4-or 5-category Likert-type scales (Converse and Presser 1986), 6- or 7-point semantic differential scales (Green and Rao 1970; Heise 1969b; Osgood, Suci, and Tannenbaum 1957; Peabody 1962), 7-category response forms (Alwin and Krosnick 1991; Andrews and Withey 1976); and even to rating scales including 9, 10, 11, or more categories. However, most of the research that has addressed issues of response forms in survey measurement of subjective phenomena has essentially ignored this issue (e.g., Bradburn and Danis 1984; Converse and Schuman 1984). The exception is Schuman and Presser's (1981) research on the use of "middle categories" in the measurement of attitudes (i.e., 3-category scales), but that research was limited to a comparison of 2- and 3-category answer categories, and their analysis focused only on the comparison of marginal distributions rather than on the quality of measurement.

This article suggests that one way in which to evaluate the effectiveness of various forms of survey questions is in terms of their reliability and validity of measurement (see Alwin 1989, 1992; Alwin and Jackson 1979; Andrews 1990; Scherpenzeel 1995). Relevant theory regarding the transmission of information in the survey interview provides a basis for several hypotheses regarding the linkage between different numbers of response categories and the quality of measurement. Against the background of this theoretical discussion and a review of pertinent research evidence, I develop several specific hypotheses regarding the relation between the number of response categories and quality of measurement. I then present some new results from the analysis of measures of "life satisfaction" from the 1978 Quality of Life Survey of U.S. households conducted by the University of Michigan's Institute for Social Research (Campbell and Converse 1980), which employed a multitrait-multimethod (MTMM) measurement design. The design presents a comparison of "feeling thermometers" involving 11 response categories to more traditional 7-point rating scales. Using statistical estimation techniques for the analysis of MTMM data (see Alwin 1974; Browne 1984; Saris and van Meurs 1990), I examine the extent to which reliability is related to theoretical predictions regarding the maximum amount of informa-

tion transmitted by response scales of differing lengths. This approach also allows an examination of the extent to which reliability is due to "valid" measurement of common factors representing trait variation, or to common factors reflecting "invalid" method variance, that is, to systematic errors such as response sets associated with scales of different lengths.

RESPONSE CATEGORIES, INFORMATION TRANSMISSION, AND THE RELIABILITY OF MEASUREMENT

The objective of social measurement is to obtain information that is valid from the point of view of the purpose of the research, or the requirements of the theory, and that is obtained using procedures which are reliable in the sense that the results can be reproduced under replication of the procedure. In survey measurement, these objectives are served by obtaining answers to questions, either through interviews or self-completed questionnaires. Survey questions take a number of different forms, depending in part on the nature of the phenomenon of interest. In the case of measuring subjective variables (e.g., attitudes, self-assessments), many survey questions involve what are essentially *rating tasks*, in which respondents are given a stimulus in the form of a question or statement and are asked to choose a category or point along a response continuum corresponding to their attitudes, opinions, and/or feelings.[1] Such questions often take the form of *agree-disagree* questions, as in the Likert scale tradition, where the respondents are asked whether they *strongly agree*, *agree*, *disagree*, or *strongly disagree* with the position or statement contained in the question.[2]

Rating tasks are used to obtain a variety of kinds of information and can take many different forms. As noted at the outset, response scales can vary widely in the number of categories used to quantify the responses. Normally, questions with more than 4 or 5 response categories require visual aids, or "showcards," for the respondents to be able to deal with such large numbers of categories. Such rating scales usually incorporate visual labels that help quantify the meanings of the scale categories. Some have argued that the quantifiers used in such survey questions are inherently vague (Bradburn and Danis

1984), but one can imagine how much more vague such response categories would be without labels (see Saris 1988). Although a minimum of labeling clearly is important for information transmission in the survey interview or self-administered questionnaire, the extensiveness of the *labeling* of response categories does not appear to affect the reliability of measurement (Alwin and Krosnick 1991; Andrews 1984).

INFORMATION THEORY

It is a time-worn assumption of information theory that a greater number of response categories conveys more information about the underlying variable of interest (Garner 1960; Garner and Hake 1951; Shannon and Weaver 1949). In the general case, the "bits of information" transmitted by a response continuum is a logarithmic function of the number of categories defined by the points on that continuum (Woelfel and Fink 1980). For example, in attitude measurement, a 2-category response scale communicates only one piece of information: attitude direction (e.g., *agree* vs. *disagree*, *approve* vs. *disapprove*, *favor* vs. *disfavor*). By contrast, a 3-category response format potentially conveys more information than does a 2-category format because it includes a category for a neutral or "middle" position. Three category scales, then, in theory communicate two pieces of information: neutrality and direction. Similarly, a 4-category format can be argued to be more informative because it conveys *intensity* or *strength* as well as direction, although it provides no neutral point. A 5-category response format is even better because it allows communication of information pertaining to both direction and intensity and also provides a neutral category and two degrees of intensity, that is, three pieces of information. Thus, if one assumes that attitudes exist on a continuum that has direction, intensity, and a region of neutrality, then a minimum of 5 categories would be necessary to communicate attitudes. This would be more likely to distinguish weak positive and negative attitudes from neutrality. This was Likert's (1932) recommended strategy for measuring attitudes.

Obviously, what we learn from information theory can be extended beyond 5 categories. One could argue, for example, that in the measurement of attitudes, 7-category response scales are better than

5-category scales because they not only permit the measurement of direction and neutrality but can distinguish three levels of attitude intensity as well (see Alwin and Krosnick 1991). Thus they provide for the communication of more information, which increases monotonically with the number of scale points. There may be some practical limitations to the use of large numbers of categories in survey measurement, and the uncritical extension of this line of reasoning may not be warranted. It should raise an interesting set of concerns for those surveys that rely primarily on 2- and 3-category scales in the measurement of attitudes, especially if one believes that attitudes should be conceptualized as having direction, intensity, and a region of neutrality.[3]

The assumption of this line of reasoning is that by using more response categories, the investigator provides a more effective framework for information transmission and respondents can more accurately communicate internal states (e.g., attitudes, feelings, beliefs). Put another way, the information obtained using more response categories ensures greater reliability of measurement.

From the point of view of effective measurement, then, the question is whether there is an optimal number of response categories, or scale points, that can be suggested to improve accuracy *regardless of survey content*. Does measurement precision increase monotonically with more response categories? Or, is there a point beyond which no further improvements in measurement quality are possible?

Due to the type of information theoretic logic referred to in the preceding, historically the conclusion has been drawn that higher reliability is a salutary consequence of using more response categories (e.g., Champney and Marshall 1939; Ferguson 1941; Guilford 1954; Jahoda, Deutsch, and Cook 1951; Murphy and Likert 1938; Symonds 1924). There is, however, a lack of consensus on this issue. Cognitive theorists would suggest that there may be some practical upper limit on the number of response categories people can handle. Certainly, given the potential cognitive difficulties that most people have in making discriminations along a scale with numerous categories, it seems plausible to argue that the quality of measurement will improve up to some point, say up to 7 categories, but beyond that information actually will be lost because the scale points tend to mean

less. Indeed, one famous article on this topic argued on the basis of a set of experiments using magnitude estimates of judgments of a number of physical stimuli that the optimal number of categories was 7 ± 2 (Miller 1956). Despite the importance of these results, it is not clear whether Miller's (1956) conclusion can be generalized to survey measures; however, it clearly is worth finding out (see, e.g., Alwin 1992).

In addition, motivational theorists also would argue against survey questions with large numbers of response categories on the grounds that respondents may not be sufficiently motivated to take the questions seriously if they are bombarded with the difficult task of making meaningful discriminations along a continuum with a large number of categories. It has been suggested that when faced with such complex tasks, many respondents may tend to "satisfice" rather than "optimize" (Alwin 1991; Krosnick and Alwin 1989; Tourangeau 1984). Cox (1980, p. 409) noted, for example, that as the number of response categories increases beyond some hypothetical minimum, response burden also increases. Cox argued that the result of more categories, on the basis of psychometric theory, is an increasing number of discrepancies between the true score and the observed score. Thus, although it may appear that the "information carrying capacity" of the scale is improved by increasing the number of response categories or scale points, it actually may be possible that reliability of measurement is lowered by using more categories.

FEELING THERMOMETERS

One prominent example of a scaling approach that uses many response categories is the application of magnitude estimation in surveys (e.g., Saris 1988). Here the respondents are given an opportunity to make fine-grained distinctions using what is essentially a continuous scale, as with a thermometer. This may present an aura of misplaced precision in that the concepts in social science often may in reality not be as continuous as we conceive temperature to be. On the other hand, subjective variables, such as attitudes, may be usefully conceptualized as latent continua reflecting predispositions to re-

spond. Although there are some social-psychological theorists who define attitude continua in terms of discrete categories, as "latitudes" or "regions" on a scale (e.g., Sherif, Sherif, and Nebergall 1965), most subjective variables can perhaps be thought of in terms of continua that reflect direction and intensity and perhaps even have a "zero" or neutral point.

One of the most notable examples of the use of this approach is the feeling thermometer used in the University of Michigan National Election Studies (NES) (Miller 1982; Weisberg and Miller n.d.).[4] The appendix gives an example of the approach used in the NES surveys. This example is one in which attitudes or feelings toward social and political figures are assessed on what is essentially a 9-point scale. This approach, and related applications, takes the point of view that the respondents' feelings or "affects" toward a political candidate, or some other attitude object, can best be registered on a response continuum with many scale points and that the analog to the thermometer provides a useful tool for measurement. Commenting on this approach, Converse and Presser (1986) noted that "despite its length, the thermometer question seems to be clear to respondents because of the familiar image of this measuring device" (p. 86). Because it involves the image of a thermometer, survey measures such as that shown in the appendix sometimes are thought of as having as many response categories, or scale points, as there are on a thermometer. My analysis of the NES feeling thermometer data indicates that rarely do more than 3% to 5% of respondents give responses other than the 9 labeled numeric options shown in the appendix (see Alwin 1992). Although analyses of such data may preserve the coding of finer gradations when they exist, such scales still should be considered to have a finite set of response categories.

In this article, I pose the question of whether such approaches to magnitude measurement in survey questionnaires exceed the somewhat more standard 7-category format in terms of measurement effectiveness. I address this issue by comparing sample estimates of reliability of each type of response format using an MTMM measurement design. I also address the question of the relative contributions of trait and method variance to the reliability of measurement. First, however, I briefly review past research on the linkage between the number of response categories and errors of measurement.

NUMBER OF RESPONSE CATEGORIES, RELIABILITY, AND VALIDITY OF MEASUREMENT

I argue that one way in which to address the question of the optimal approach to measuring subjective variables in surveys is in terms of the *reliability* of measurement. For these purposes, it is essential to be able to estimate the reliability of single items (Alwin 1989). Unfortunately, virtually all research examining this question has instead looked at the reliability of composite scores involving items with different numbers of response categories (e.g., Birkett 1986; Finn 1972; Jenkins and Taber 1977; Komorita 1963; Komorita and Graham 1965; Lissitz and Green 1975; Matell and Jacoby 1971, 1972; Ramsay 1973). This is problematic because composite reliability is affected by a number of factors, not just the reliability of measurement.

There is, however, a small literature that has focused on the reliability of individual survey questions. Andrews and Withey (1976), for example, compared 7- and 3-category response scales in the measurement of subjective well-being, noting that 7-point scales provide more sensitive indications of respondents' feelings and have higher reliabilities than do 3-point scales. Similarly, Andrews (1984, p. 20) compared response scales with 2, 3, 4-5, 7, 9-19, and 20+ categories in terms of reliability, concluding that the quality of the data improved with increases in response categories. Although reliabliity generally increased with more categories, he also found that 3-category scales were less reliable than scales with 2 or 4-5 categories.[5]

These findings are supported by other recent research as well. Based on Heise's (1969a) approach to reliability estimation using 3-wave panel data, my own research (Alwin 1992) found that for attitude measures in the NES and the General Social Surveys, with the one exception noted earlier (i.e., 3-category response scales), reliability generally is higher for attitudes measured using more response categories. These results are reproduced for 99 attitude measures in these studies in Figure 1. These results clearly show that there is a relationship between the number of bits of information conveyed by a set of response categories and the reliability of attitude measurement.[6] This is true, at least for 4 or more response categories. However, in contradiction to the suggestion of information theory, reliability is relatively higher when attitudes are assessing using 2-category rather

than 3-category response scales. I argued that this is because 2-category scales unambiguously assess attitude direction *only*, whereas 3-category scales confuse direction, neutrality, and intensity (Alwin 1992). Reliability improves using 4 or more categories because, I argued, these scales provide for more accurate transmission of information. The rate of increase in information transmitted with more categories occurs at a decreasing rate, which appears also to be true of reliability.

The 9-category scales from the NES data shown in Table 2 are from feeling thermometers of the sort depicted in the appendix, and thus the comparison between 7- and 9-category scales is of critical importance to our present set of concerns. In the research reported in the following, I compare 7-category scales to 11-point feeling thermometers in the measurement of life satisfaction. Clearly, the gain in reliability shown in Figure 1 from 7- to 9-category scales is relatively small, but it does conform to the prediction of the theory and was found for several different sets of studies from the NES (the 1950s, 1970s, and 1980s panels). On the basis of these findings for attitude measures, as well as the information theoretic argument advanced heretofore, I hypothesize that 11-category measurement will be more reliable than 7-category response scales in the measurement of life satisfaction.

However, the apparent advantages of reliability from higher order scales may not be due entirely to increased precision of measurement. Systematic errors may, in fact, contribute to these results (Alwin 1992, p. 112). As Cronbach (1950) pointed out more than 45 years ago, an exclusive concern with the reliability of measurement as a criterion for evaluating the optimal number of response categories ignores the possibility that some of the reliable variance is really "invalid" in the sense that it represents something about respondents' use of the scale. The use of batteries of questions in survey questionnaires, all of which use the same format, may encourage response sets and systematic responding (see Alwin and Krosnick 1985). The use of greater numbers of response categories may provide more of a basis for such rating biases in that more response categories allow for more idiosyncratic use of rating scales by respondents, and this could result in higher levels of reliability due to shared method variance. This may mean that respondents' ratings tend to fall within a restricted range of the available scale points (Feather 1973), that there may be a tendency toward differential anchoring at one extreme or the other (Hamilton

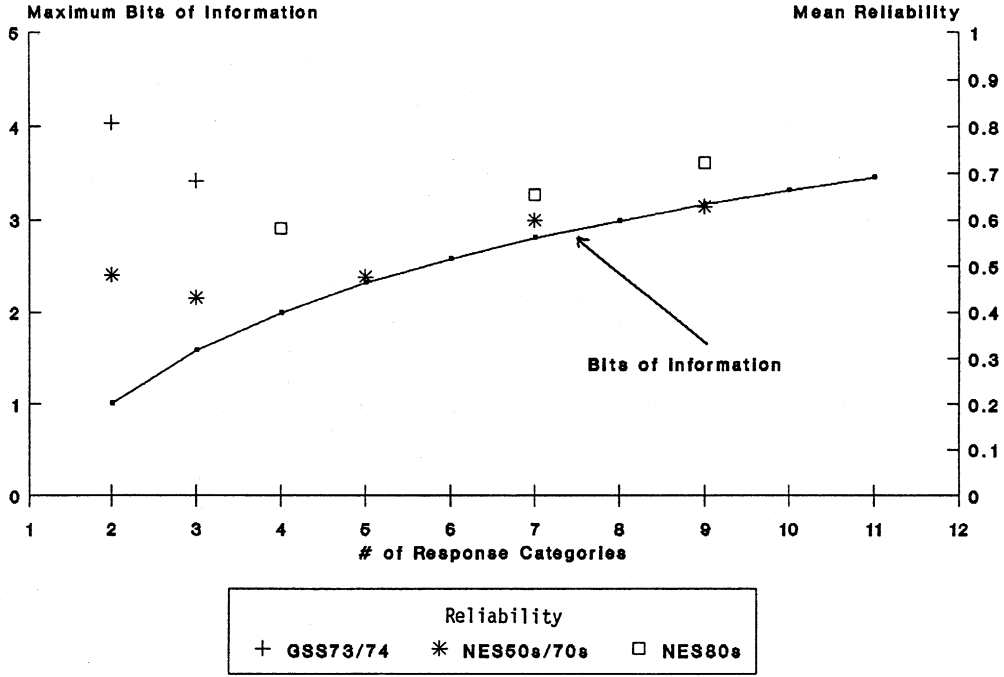

Figure 1: Relationship of Number of Response Categories, Information Transmission, and Reliability of Rating Scales
SOURCE: Alwin (1992). Reprinted with permission.

1968), that individuals may interpret the meanings of judgment categories differentially (Cronbach 1946, 1950; Messick 1968), or that subgroups may differ in the nature of their response sets (Cunningham, Cunningham, and Green 1977). Variations across persons in such response tendencies lead to correlated response patterns, or what Costner (1969) referred to as "differential bias," producing spuriously positive correlations among measures due to the common method of measurement. As a consequence, it becomes important to attempt to separate reliability into two components: reliable trait variation and reliable method variation. This can be accomplished via methods of MTMM design and analysis (see Alwin 1974; Andrews 1984; Campbell and Fiske 1959; Saris and van Meurs 1980).

MEASUREMENT MODEL

One method that has been proposed for obtaining information on the interpretation of the reliability of measurement is based on Campbell and Fiske's (1959) famous MTMM matrix. This approach, summarized by Alwin (1974) and Browne (1984), augments a basic assumption of classical true score theory (Lord and Novick 1968) by positing *two* sources of "true" variation in each measure, one representing *trait* variation and one representing *method* variation. A confirmatory factor analytic model for the MTMM matrix, attributable to Jöreskog (1971) and Werts and Linn (1970), has been developed for this purpose. This model may be written as follows for a set of measures in which each measure is a unique combination of a given trait and a given method:

$$y_{ij} = \lambda_i \tau_i + \lambda_j \eta_j + \varepsilon_{ij},$$

where y_{ij} is the measure of the i^{th} trait by the j^{th} method, τ_i is the i^{th} trait factor being measured, η_j is the j^{th} method factor measured, and ε_{ij} is the random error component associated with the measurement of a particular combination of trait and method in y_{ij}. The λ coefficients are factor pattern coefficients, that is, regression coefficients linking the y_{ij} measure to the i^{th} trait and j^{th} method factors. Following the termi-

nology suggested by Heise and Bohrnstedt (1970), I refer to the standardized λ coefficient linking a given measure to reliable trait variation (i.e., τ_i) as a *validity* coefficient and to the standardized λ coefficient representing reliable method variation (i.e., η_j) as an *invalidity* coefficient. In other words, both valid and invalid sources of variation contribute to reliability, and the purpose of this modeling strategy is to separate these two components. Unless noted otherwise, I discuss the standard form of the model throughout the remainder of the article.

Given this model, along with the assumption that the three components are independent, the observed variance of a particular measure y_{ij} may be decomposed into components of reliability, validity, and invalidity as follows:

$$\sigma^2_{y_{ij}} = \lambda^2_i\sigma^2_{\tau_i} + \lambda^2_j\sigma^2_{\eta_j} + \sigma^2_{\varepsilon_{ij}},$$

where $\lambda^2_i\sigma^2_{\tau_i}$ represents the estimated true trait variance in the variable being measured, $\lambda^2_j\sigma^2_{\eta_j}$ represents the estimated true variance due to the method of measurement, and $\sigma^2_{\varepsilon_{ij}}$ represents the estimated random measurement error variance. Under this model, the *reliability* of a given measure is equal to the sum of the true trait and method variance expressed as a proportion of the total observed response variance. Thus the decomposition of the population variance of a given measure (i.e., a particular trait-method combination in some population of interest) into components of reliability, validity, and invalidity can be written as follows:

$$1.0 = \frac{\sigma^2_{y_{ij}}}{\sigma^2_{y_{ij}}} = \overbrace{\underbrace{\frac{\lambda^2_i\sigma^2_{\tau_i}}{\sigma^2_{y_{ij}}}}_{\text{Validity}} + \underbrace{\frac{\lambda^2_j\sigma^2_{\eta_j}}{\sigma^2_{y_{ij}}}}_{\text{Invalidity}}}^{\text{Reliability}} + \overbrace{\frac{\sigma^2_{\varepsilon_{ij}}}{\sigma^2_{y_{ij}}}}^{\text{Unreliability}}.$$

The objective of the following data analysis is to estimate the reliability of measures and its components within methods of measurement.

DATA AND MEASURES

The primary results reported here consist of a comparison of 7- and 11-category rating scales in the measurement of life satisfaction. This comparison is based on a 1978 national survey in which 17 domains of life satisfaction were measured using these two types of rating scales. The 1978 Quality of Life Survey consists of a probability sample (N = 3,692) of persons 18 years of age or older living in households (excluding those on military reservations) within the coterminous United States. Interviews were conducted during June through August 1978. The original sample of approximately 4,870 occupied housing units, comprising two independently chosen multistage area probability samples, was used to represent the noninstitutionalized adult population of the United States. The overall completion rate was approximately 76%. Sampling and other procedural details are given in Campbell and Converse (1980).

The design of this survey included multiple measures of several domains of life satisfaction. This design permits the analysis of both trait and method components of variation in each measure given that each domain was assessed using multiple response formats (see Alwin 1989; Andrews 1984; Saris and van Meurs 1990). The following 17 domains of satisfaction were assessed: satisfaction with the respondents' community, neighborhood, place of residence (dwelling unit), life in the United States today, education received, present job (for those respondents who were employed), being a housewife (for unemployed women respondents), ways in which to spend spare time, personal health, family's present income, standard of living, savings and investments, friendships, marriage (for married respondents), family life, self as a person, and life as a whole.[7] All of these measures were assessed using both 7- and 11-point response scales. Three of these domains—place of residence, standard of living, and life as a whole—were rated using three separate scales. The three methods used were (a) a 7-point *satisfied-dissatisfied* scale, (b) a 7-point *delighted-terrible* scale, and (c) an 11-point feeling thermometer. The order of presentation of measurement approaches was, unfortunately, the same across topics and across respondents. This confounds question context and measurement format, and the possibility of respon-

dent conditioning could affect the results. Also, the methods of measurement differed in the extent of verbal labeling of response options. The 7-point *satisfied-dissatisfied* scale labeled only the end points and the midpoint, the 7-point *delighted-terrible* scale presented a fully labeled set of response options, and the 11-point feeling thermometer labeled only the end points (in contrast to that shown in the appendix). Thus the issues of number of scale points and extensiveness of verbal labeling also are confounded in the results presented in the following. It is nonetheless worth devoting attention to these data given the potential contribution to knowledge about the link between reliability and number of response categories. I return to a consideration of these issues, however, in the discussion and interpretation of the results.

RESULTS

The data described in the preceding section were analyzed using the LISREL computer program (Jöreskog and Sörbom 1986). The covariance matrix among the variables was analyzed using the model described earlier, wherein the sample response variances of the variables were partitioned into three parts: (a) reliable trait variance, (b) reliable method variance, and (c) unreliable variance. Because, as noted earlier, some of the satisfaction questions were asked only of subsets of the sample (specifically those employed [job satisfaction], those doing housework [housework satisfaction], and those married [marital satisfaction]), it was necessary to explore several sets of results. I first estimated the model for the total sample for those questions asked of everyone. The results for such questions are based on the total sample.[8] Then I performed parallel analyses for all questions asked of the married subsample, all questions asked of the employed subsample, and all those asked the housework question. I used the results from these subsamples to report the coefficients for the measurement of job, marital, and housework satisfaction. A detailed comparison of the results for the other items in these subgroups indicated virtually identical patterns to what is reported for the total sample. I therefore feel comfortable with the coefficients presented subsequently as representative of the measurement quality of 7- and 11-point scales across these 17 domains.

TABLE 1: Reliability, Validity, and Invalidity of Life Satisfaction Measures: 1978 Quality of Life Survey (N = 3,692)

	7-Point Satisfaction Scale			*7-Point Delighted-Terrible Scale*			*11-point Thermometer*		
Concept	ρ	λ_τ	λ_η	ρ	λ_τ	λ_η	ρ	λ_τ	λ_η
Community	.681	.774	.290				.770	.674	.577
Neighborhood	.765	.816	.312				.891	.648	.685
Dwelling	.676	.762	.289	.756	.777	.375	.794	.761	.447
United States	.455	.613	.295				.551	.686	.292
Education	.447	.592	.328				.987	.963	.254
Health	.526	.709	.246				1.056	1.012	.208
Time	.582	.612	.465				.756	.800	.336
Friends	.533	.656	.346				.734	.765	.365
Family	.537	.656	.346				.785	.845	.264
Income	.673	.698	.442				.820	.860	.386
Standard of living	.697	.681	.489	.732	.643	.559	.791	.807	.365
Savings	.786	.773	.403				.814	.856	.277
Life	.619	.635	.497	.579	.628	.431	.741	.812	.285
Self	.608	.698	.394				.768	.845	.224
Job	.619	.772	.165				.882	.900	.341
Housework	.564	.702	.296				.853	.922	.036
Marriage	.608	.759	.233				.886	.941	.089

Table 1 presents a summary of the model estimates. Three coefficients are presented for each trait/method combination: (a) the estimated *reliability* for each item, (b) the coefficient of *validity* (the standardized factor pattern coefficient linking that item to the trait factor in question), and (c) the coefficient of *invalidity* (the standardized factor pattern coefficient linking that item to the method factor involved). In the table, these are denoted as ρ, λ_τ, and λ_η, respectively.

The key results indicate that in the comparison between 7- and 11-point scales, the latter have higher reliabilities in every case. This strongly supports the hypothesis, based on information theory, that questions with more response categories permit the transmission of information more reliably. Moreover, in 14 of 17 cases, 11-point scales have higher validity coefficients, indicating that traits measured using more response categories are more highly correlated with the underlying trait than is the case with those measured by 7-point scales. Finally, in all but 5 of the 17 cases, the 11-point scales clearly have lower coefficients of invalidity, indicating that they are affected less,

rather than more, by method variance. This runs counter to the hypothesis suggested earlier, which argued that questions with more response categories are more vulnerable to the effects of systematic response error such as response sets. The results decidedly support the use of 11-point scales over 7-point scales in the measurement of life satisfaction.

DISCUSSION AND CONCLUSIONS

The results reported here represent an improvement over previous approaches in several respects. First, this investigation has focused on the reliability of single survey questions and has avoided the complexities involved in formulating this set of issues within the framework of internal consistency estimates of reliability (e.g., coefficient alpha), which depend not only on the extent of item-level reliability but also on the number of items and the unidimensionality of scale components (Cronbach 1951). Second, this study has made the comparison of response scale length with respect to one domain of content, self-assessed satisfaction with various aspects of life, and has not confounded the assessment of the relation between the number of response categories and reliability with the types of concepts being measured (cf. Andrews 1984). Third, although this study has compared only two scale lengths, 7 and 11 categories, the comparison focuses on a critical issue in the debate about whether it is efficacious to extend more traditional approaches. The 7-category scale is by far the longest type of response scale used in most survey measurements of subjective variables. Scales with more than 7 categories may be impractical in many situations, and there is an issue of whether such decisions can be justified in terms of reliability of measurement. Of course, the choice of response scales also should be based on construct validity, and the theoretical appropriateness of a particular type of measure may be an overriding concern. Increasingly, matters of practical significance are important given the relationship of survey costs to the amount of information obtained in surveys.

As noted earlier, there are two methodological issues that detract from the clarity of these results. The first is that the order of presentation of measurement forms was the same across topics across respon-

dents. The 7-point satisfaction scales were given first, the thermometers were given second (somewhat later in the questionnaire), and the *delighted-terrible* scales were given last (near the end of the questionnaire). The nonbalanced nature of the design does not permit distinguishing between form of measurement and measurement context in that there may be some form of conditioning effect. For example, respondents may have the answers to the earlier questions still relatively available in memory, and this may affect their responses on the later tasks, namely the thermometer. This would work to increase the true correlations among measures in the monotrait-heteromethod block, and the net result would be stronger trait correlations for the later set of measures and less influence of method factors. However, in this case the very last set of measures employed are the 7-point *delighted-terrible* scales, and the hypothesis is disconfirmed in that case; that is, they do not have lower method components or higher validity coefficients. In light of this evidence, this explanation for our main results hardly seems credible.

A second problem of interpretation in this study involves the fact that the three forms of measurement included different degrees of labeling. The 7-point *satisfied-dissatisfied* scale labeled the end points and midpoint, the 11-point thermometer labeled only the end points, and the 7-point *delighted-terrible* scale presented a fully labeled set of response options. There clearly is confounding between the number of response categories and the nature of the verbal labels attached. However, within the present context, it is extremely difficult to sort out. Past research may be the only basis for guidance. Past research, however, has shown that the extensiveness of labeling does not appear to be related to the reliability of measurement (Alwin and Krosnick 1991; Andrews 1984).

If these alternative explanations can be ruled out in the present case, then the results suggest that in the measurement of satisfaction with various domains of life, 11-point scales clearly are more reliable than comparable 7-point scales. Moreover, measures involving 11-point scales have higher correlations with underlying trait components and lower correlations with method components underlying the measures. In addition to being more reliable, then, 11-point scales are *no more* vulnerable to response sets, as conceptualized here in terms of shared method variance, when compared to 7-point scales. These findings

support the conclusion that questions with more response categories may be preferable to those with fewer response categories, when feasible, in that they produce measures that are both more reliable and more valid. Reductions in measurement errors in this sense can result in more powerful statistical decision making, less biased correlation and regression estimates, and greater confidence in the usefulness of the data.

APPENDIX
Example of the Michigan Feeling Thermometer

I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of the person and I'd like you to rate that person using this feeling thermometer. You may use any number from 0 to 100 for a rating. Ratings between 50 and 100 degrees mean that you feel favorable or warm toward the person. Ratings between 0 and 50 degrees mean that you don't feel too favorable toward the person. If we come to a person whose name you don't recognize, you don't need to rate that person. Just tell me and we'll move on to the next one. If you do recognize the name, but don't feel particularly warm or cold toward the person, you would rate that person at the 50-degree mark.

— 100° Very warm or favorable feeling
— 85° Quite warm or favorable feeling
— 70° Fairly warm or favorable feeling
— 60° A bit more warm or favorable than cold feeling
— 50° No feeling at all
— 40° A bit more cold or unfavorable feeling
— 30° Fairly cold or unfavorable feeling
— 15° Quite cold or unfavorable feeling
— 0° Very cold or unfavorable feeling
— Don't recognize name

NOTES

1. It is important to distinguish between ratings and rankings in this context, both because of their inherent differences and because ratings are far more pervasive in surveys. In contrast

to ratings, rankings ask respondents to order a set of objects or their attributes according to some criterion of preference or judgment (see Alwin and Krosnick 1985).

2. There is some confusion among practitioners as to what constitutes a Likert-type scale. Some use this term to refer to *any* agree-disagree format (e.g., Converse and Schuman 1984; Schuman and Presser 1981). Likert (1932) actually used a 5-category *agree-disagree* scale including a "neutral" point and the explicit offering of a *don't know* response.

3. It is noteworthy in this regard that one of the most widely used attitude surveys in the social sciences, the General Social Survey (GSS) (Davis and Smith 1995), relies heavily on 2- and 3-category scales. The National Election Study (Center for Political Studies 1994), by contrast, primarily uses 7-category scales and feeling thermometers.

4. This method has broad appeal and certainly is not limited to Michigan surveys. I refer here to the Michigan studies in part because of their visibility and in part because this measurement approach distinguishes them from other prominent surveys, for example, the GSS (Davis and Smith 1995).

5. Unfortunately, although Andrews's (1984) analysis represents an important set of findings, his results with respect to number of response categories may be less informative for present purposes than might otherwise be desirable. Andrews analyzed a pool of survey questions measuring a wide range of content including subjective variables as well as reports of factual content (e.g., number of days per month on which certain behaviors occurred). The obvious linkage between behavioral versus subjective content and the tendency for greater numbers of response categories to be used with behavioral reports may, therefore, confound content with the number of response categories in assessing reliability.

6. "Bits of information" is computed as follows: $H = \log_2(x) = \ln(x)/\ln(2)$. See Woelfel and Fink (1980, pp. 20-21).

7. Full details on the wording and presentation of these measures, along with the correlation matrices and setup files, can be obtained from the author on request.

8. The estimated likelihood ratio statistic, L^2, for this model is 1,800.7 with 309 degrees of freedom. Because it is directly dependent on the sample size, this statistic cannot be used to evaluate the fit of the model. Even trivially different models would be rejected because of a highly significant difference on the χ^2 probability distribution. The Bentler and Bonett (1980) *normed fit index* (NFI) permits the comparison of models free of the influence of sample size (see Alwin 1988). The NFI for the present model equals .97 compared to an NFI of .89 for a model that excludes the method factors ($L^2 = 6{,}764.97$ with 465 degrees of freedom). Thus the proposed model is considered to reflect a more than adequate fit to the data. Clearly, use of some of the 309 degrees of freedom to alter the model to achieve a better fit to the data is possible, but the comparative fit of the model can hardly be improved, and it is extremely doubtful that such model modifications would change the basic results reported.

REFERENCES

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

———. 1988. "Structural Equation Models in Research on Human Development and Aging." Pp. 71-170 in *Methodological Issues in Aging Research*, edited by K. W. Schaie, R. T. Campbell, W. Meredith, and S. C. Rawlings. New York: Springer.

———. 1989. "Problems in the Estimation and Interpretation of the Reliability of Survey Data." *Quality and Quantity* 23:277-331.

———. 1991. "Research on Survey Quality." *Sociological Methods & Research* 20:3-29.

———. 1992. "Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement." Pp. 83-118 in *Sociological Methodology 1992*, edited by P. V. Marsden. Oxford, UK: Basil Blackwell.

Alwin, D. F. and D. J. Jackson. 1979. "Measurement Models for Response Errors in Surveys." Pp. 68-119 in *Sociological Methodology 1980*, edited by P. V. Marsden. San Francisco: Jossey-Bass.

Alwin, D. F. and J. A. Krosnick. 1985. "The Measurement of Values in Surveys: A Comparison of Ratings and Rankings." *Public Opinion Quarterly* 48:409-42.

———. 1991. "The Reliability of Attitudinal Survey Measures: The Role of Question and Respondent Attributes." *Sociological Methods & Research* 20:139-81.

Andrews, F. M. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 46:409-42.

———. 1990. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." Pp. 15-51 in *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, edited by W. E. Saris and A. van Meurs. Amsterdam, Netherlands: North-Holland.

Andrews, F. M. and S. B. Withey. 1976. *Social Indicators of Well-Being: Americans' Perceptions of Life Quality*. New York: Plenum.

Benson, P. H. 1971. "How Many Scales and How Many Categories Shall We Use in Consumer Research? A Comment." *Journal of Marketing* 35:59-61.

Bentler, P. M. and D. G. Bonett. 1980. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88:588-606.

Birkett, N. J. 1986. "Selecting the Number of Response Categories for a Likert-Type Scale." In *Proceedings of the American Statistical Association* (Section on Survey Research Methods). Washington, DC: American Statistical Association.

Bradburn, N. and C. Danis. 1984. "Potential Contributions of Cognitive Research to Questionnaire Design." Pp. 101-29 in *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, edited by T. B. Jabine, M. L. Straf, J. M. Tanur, and R. Tourangeau. Washington, DC: National Academy Press.

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

Campbell, A. and P. E. Converse. 1980. *The Quality of American Life: 1978 Codebook*. Ann Arbor: University of Michigan, Inter-University Consortium for Political and Social Research.

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

Center for Political Studies. 1994. *Continuity Guide to the American National Election Studies, 1952-1993*. Ann Arbor: University of Michigan, Institute for Social Research.

Champney, H. and H. Marshall. 1939. "Optimal Refinement of the Rating Scale." *Journal of Applied Psychology* 23:323-31.

Converse, J. M. and S. Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Newbury Park, CA: Sage.

Converse, J. M. and H. Schuman. 1984. "The Manner of Inquiry: An Analysis of Survey Question Form Across Organizations and Over Time." Pp. 283-316 in *Surveying Subjective Phenomena*, Vol. 2, edited by C. F. Turner and E. Martin. New York: Russell Sage.

Costner, H. L. 1969. "Theory, Deduction, and Rules of Correspondence." *American Journal of Sociology* 75:245-63.

Cox, E. P., III. 1980. "The Optimal Number of Response Alternatives for a Scale: A Review." *Journal of Marketing Research* 17:407-22.

Cronbach, L. J. 1946. "Response Evidence on Response Sets and Test Design." *Educational and Psychological Measurement* 6:475-94.

———. 1950. "Further Evidence on Response Sets and Test Design." *Educational and Psychological Measurement* 10:3-31.

———. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16:297-334.

Cunningham, W. H., I.C.M. Cunningham, and R. T. Green. 1977. "The Ipsative Process to Reduce Response Set Bias." *Public Opinion Quarterly* 41:379-84.

Davis, J. A. and T. W. Smith. 1995. *General Social Surveys, 1972-1995: Cumulative Codebook.* Chicago: National Opinion Research Center.

Feather, N. T. 1973. "The Measurement of Values: Effects of Different Assessment Procedures." *Australian Journal of Psychology* 25:221-31.

Ferguson, L. W. 1941. "A Study of the Likert Technique of Attitude Scale Construction." *Journal of Social Psychology* 13:51-57.

Finn, R. H. 1972. "Effects of Some Variations in Rating Scale Characteristics on Means and Reliabilities of Ratings." *Educational and Psychological Measurement* 32:255-65.

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

Garner, W. R. and H. W. Hake. 1951. "The Amount of Information in Absolute Judgements." *Psychological Review* 58:446-59.

Green, P. E. and V. R. Rao. 1970. "Rating Scales and Information Recovery: How Many Scales and Response Categories to Use?" *Journal of Marketing* 34:33-39.

Guilford, J. P. 1954. *Psychometric Methods.* New York: McGraw-Hill.

Hamilton, D. L. 1968. "Personality Attributes Associated With Extreme Response Set." *Psychological Bulletin* 72:406-22.

Heise, D. R. 1969a. "Separating Reliability and Stability in Test-Retest Correlations." *American Sociological Review* 34:93-101.

———. 1969b. "Some Methodological Issues in Semantic Differential Research." *Psychological Bulletin* 72:406-22.

Heise, D. R. and G. W. Bohrnstedt. 1970. "Validity, Invalidity, and Reliability." Pp. 104-29 in *Sociological Methodology 1970*, edited by E. F. Borgatta and G. W. Bohrnstedt. San Francisco: Jossey-Bass.

Jabine, T. B., M. L. Straf, J. M. Tanur, and R. Tourangeau. 1984. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* (report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology, Committee on National Statistics, and Commission on Behavioral and Social Sciences Education, National Research Council). Washington, DC: National Academy Press.

Jacoby, J. and M. S. Matell. 1971. "Three-Point Scales Are Good Enough." *Journal of Marketing Research* 8:495-500.

Jahoda, M., M. Deutsch, and S. W. Cook. 1951. *Research Methods in Social Relations.* New York: Dryden.

Jenkins, G. D., Jr. and T. D. Taber. 1977. "A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability." *Journal of Applied Psychology* 62:392-98.

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

Jöreskog, K. G. and D. Sörbom. 1986. *LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood* (user's guide, Version 6). Chicago: Scientific Software Inc.

Komorita, S. S. 1963. "Attitude Content, Intensity, and the Neutral Point on a Likert Scale." *Journal of Social Psychology* 61:327-34.

Komorita, S. S. and W. K. Graham. 1965. "Number of Scale Points and the Reliability of Scales." *Educational and Psychological Measurement* 25:987-95.

Krosnick, J. A. and D. F. Alwin. 1989. "Response Strategies for Coping With the Cognitive Demands of Survey Questions." Unpublished manuscript, Institute for Social Research, University of Michigan.

Lehman, D. R. and J. Hulbert. 1972. "Are Three-Point Scales Always Good Enough?" *Journal of Marketing Research* 9:444-46.

Likert, R. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology*, No. 140.

Lissitz, R. W. and S. B. Green. 1975. "Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach." *Journal of Applied Psychology* 60:10-13.

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

Mattell, M. S. and J. Jacoby. 1971. "Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity." *Educational and Psychological Measurement* 31:657-74.

———. 1972. "Is There an Optimal Number of Alternatives for Likert Scale Items? Effects of Testing Time and Scale Properties." *Journal of Applied Psychology* 56:506-9.

McKennell, A. 1974. "Surveying Attitude Structures." *Quality and Quantity* 7:1-96.

Messick, S. 1968. "Response Sets." Pp. 492-96 in *The International Encyclopedia of the Social Sciences*, Vol. 13, edited by D. L. Sills. New York: Macmillan.

Miller, G. A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63:81-97.

Miller, W. E. 1982. *American National Election Study, 1980*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Murphy, G. and R. Likert. 1938. *Public Opinion and the Individual: A Psychological Study of Student Attitudes on Political Questions, With a Retest Five Years Later*. New York: Russell & Russell.

Osgood, C. E., G. J. Suci, and P. H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana: University of Illinois Press.

Peabody, D. 1962. "Two Components in Bi-Polar Scales: Direction and Extremeness." *Psychological Review* 69:65-73.

Ramsay, J. O. 1973. "The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values." *Psychometrika* 38:513-32.

Saris, W. E. 1988. *Variation in Response Functions: A Source of Measurement Error in Attitude Research*. Amsterdam, Netherlands: Sociometric Research Foundation.

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

Scherpenzeel, A. 1995. "A Question of Quality: Evaluating Survey Questions by Multitrait-Multimethod Studies." Ph.D. thesis, University of Amsterdam.

Schuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Shannon, C. and W. Weaver. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Sherif, C. W., M. Sherif, and R. E. Nebergall. 1965. *Attitude and Attitude Change*. Philadelphia: W. B. Saunders.

Sudman, S., N. M. Bradburn, and N. Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Symonds, P. M. 1924. "On the Loss of Reliability in Ratings Due to Coarseness of the Scale." *Journal of Experimental Psychology* 7:456-61.

Tourangeau, R. 1984. "Cognitive Sciences and Survey Methods." Pp. 73-100 in *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, edited by T. B. Jabine, M. L. Straf, J. M. Tanur, and R. Tourangeau. Washington, DC: National Academy Press.

Turner, C. F. and E. Martin. 1984. *Surveying Subjective Phenomena*. New York: Russell Sage.

Weisberg, H. F. and A. H. Miller. n.d. *Evaluation of the Feeling Thermometer: A Report to the National Election Study Board Based on Data From the 1979 Pilot Survey*. Ann Arbor, MI: Center for Political Studies, Institute for Social Research.

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

Woelfel, J. and E. Fink. 1980. *The Measurement of Communication Processes: Galileo Theory and Method*. New York: Academic Press.

Duane F. Alwin is a professor of sociology and program director in the Survey Research Center of the Institute for Social Research at the University of Michigan. In addition to measurement issues in survey research, he is interested in human development and aging, the family, social change, social inequality, and social psychology. His current work focuses on the linkage between aging and social change. Recent publications include "Taking Time Seriously: Social Change, Social Structure and Human Lives" in Linking Lives and Contexts: Perspectives on the Ecology of Human Development *(edited by P. Moen et al., American Psychological Association, 1994) and "Parental Socialization in Historical Perspective" in* The Parental Experience at Midlife *(edited by C. Ryff and M. M. Seltzer, University of Chicago Press, 1996).*