Journal of Official Statistics
Vol. 8. No. 3, 1992, pp. 251–275
© Statistics Sweden

Quality of Survey Measures: A Structural Modeling Approach

Willard L. Rodgers, Frank M. Andrews, and A. Regula Herzog[1]

Abstract: Estimates of the quality of about one hundred survey measures, broadly representative of those commonly used in survey research, were obtained by specifying several multitrait-multimethod matrices and estimating the parameters of measurement models. Specifically, the total variance of each measure in such a matrix was allocated to (1) a concept factor (to provide an estimate of construct validity); (2) a method factor; and (3) the residual. These estimates of data quality were then analyzed to identify characteristics of survey designs that are associated with more accurate responses (e.g., with higher loadings on the concept factors). The most important measure characteristic was the number of response alternatives offered. The "unfolding" of response alternatives was also associated with a higher level of construct validity, as was the explicit provision of a frame of reference for the question. Differences in data quality between subgroups of respondents defined by age, education, and a variety of other characteristics were statistically non-significant for the measures examined. Data are from face-to-face interviews conducted with an area probability sample of about 1,500 persons.

Key words: Data quality; construct validity; response biases.

1. Introduction

Survey researchers design their interview instruments with many considerations for maximizing the quality of the data to be collected. Some of these considerations are based on empirical evidence, but many derive only from anecdotal suggestions. The research to be reported here was designed to test systematically the effects of a number of survey question features by using multitrait-multimethod data matrices and structural equation models to estimate the quality of the data. Survey researchers also suspect that particular subgroups of the population – such as the elderly or those with less education – may provide survey responses of lesser quality. The research to be reported here also addresses subgroup differences in data quality.

1.1. Survey measure characteristics

Research has been done on the relationship between numerous characteristics of questions and quality of data, but it is fragmentary and tends not to be conclusive. For example, one of the most salient characteristics of a question is its substantive content, but remarkably few generalizations can be

[1] Willard L. Rodgers, Frank M. Andrews, and A. Regula Herzog, Survey Research Center, The University of Michigan, Ann Arbor, MI 48106-1248, U.S.A.

Acknowledgements: This research was supported by Grant No. AG02038 from the National Institute on Aging. The authors are grateful to Lynn Dielman for assistance with data analysis.

supported with respect to the relative quality of answers obtained to survey questions about different topics. Research on the length of the reporting period specified in the question suggests that there are generally two counteracting processes at work: underreporting due to failure to recall relevant events and overreporting due to the respondent erroneously including an event that happened before the specified time period – i.e., telescoping (Neter and Waksberg 1964; Sudman and Bradburn 1974). The effects of both processes become more substantial as the reporting period is extended, but it is generally assumed that the net effect is to increase underreporting (Sudman and Bradburn 1974). Some work suggests that longer questions produce better answers (Blair, Sudman, Bradburn, and Stocking 1977; Marquis and Cannell 1971), although Sudman and Bradburn (1974) could not generally confirm such an effect in their meta-analysis of the existing literature. The effect of question length remains an important research issue.

Research on the form of response scales suggests that choice from a total of seven to nine different response alternatives is optimal for the general population because this represents a manageable task, yet captures most of the variance, and so provides higher reliability than scales with fewer categories (Alwin, in press; Alwin and Krosnick 1991; Bollen and Barb 1981; Cochran 1968; Cox 1980). Scales using various graphical devices for rating concepts are less language-bound and thereby perhaps easier to use than response scales with verbal categories (at least in face-to-face interviews – use of graphical devices is possible with telephone interviews only if these devices are provided to the respondent in advance of the interview). Examples are ladders on which concepts can be placed higher or lower according to how they are evaluated, or faces showing various degrees of happy-unhappy expressions, one of which is chosen by the respondent as a representation of his or her subjective well-being. Andrews and Withey (1976) reported good measurement qualities for some of the non-verbal scales with which they experimented.

Research on the provision of an explicit "I don't know" (DK) response category has shown that this leads to a much higher proportion of respondents who choose the DK answer than if no such answer is explicitly mentioned (Schuman and Presser 1981); presumably it signals that DK is an acceptable answer. If respondents are left to believe that they cannot answer with DK, some will either respond according to an undifferentiated attitude or make a random guess, and thereby add either systematic bias or random error to the data. This is the theory, but empirical information about the quality of the data resulting from response scales containing or not containing a DK category is limited. Alwin and Krosnick (1991) and McClendon and Alwin (1990) found little evidence that deleting respondents on the basis of a DK "filter" question affected the quality of the data.

Finally, survey questions that use the same response scale are often presented to the respondent in the form of a battery of questions asked one after another. Recent evidence suggests that such a presentation may reduce the quality of the answers. For example, teenagers display an increased tendency to respond in a set manner to questions presented in a battery, particularly when they are possibly fatigued by a lengthy questionnaire (Herzog and Bachman 1981). This negative effect may be caused by the strong "Gestalt" that such a set of questions assumes when presented in close contiguity and identical format.

While a body of research addresses these and other issues of survey design, as reflected in the excellent reviews by Bradburn (1983) and by Sudman and Bradburn (1974), rarely have they been investigated in a coordinated and comprehensive fashion that permits comparative evaluations of all major aspects of survey measure design. In the most directly relevant study, Andrews (1984) investigated many survey design characteristics using a multitrait-multimethod approach. He found that higher numbers of response categories, the inclusion of an explicit DK response category, avoidance of long batteries, and the specification of a frame of reference enhanced the quality of the data. The study to be presented here is an extension of the earlier work by Andrews.

1.2. Survey respondent characteristics

It remains a popular notion that certain types of respondents provide systematically better or worse information in interviews than others. For example, concerns are often raised about older respondents and their ability to provide accurate information. Or, persons who want approval by others are expected to report falsely on aspects of their lives that they think others would judge desirable or undesirable (Crowne and Marlowe 1964). Bradburn (1983) concluded after a comprehensive review of the relevant literature that personal characteristics of respondents are responsible for only a very minor part of the variation in data quality. Andrews's (1984) investigation of the effects of respondent characteristics on data quality essentially confirmed Bradburn's conclusion, yielding generally small differences associated with personal characteristics. In his work, the characteristic associated with the clearest differences was age of the respondent, and it was investigated in more detail in a separate paper (Andrews and Herzog 1986): data quality was lower among those 55 years of age or older than among younger respondents. Alwin and Krosnick (1991) found that the reliability of responses in panel surveys is higher for more educated respondents.

1.3. Multitrait-multimethod approach

Our approach to the assessment of the validity of survey data relies on the measurement of multiple concepts by multiple methods, as suggested by Campbell and Fiske (1959). Multitrait-multimethod (MTMM) data can be analyzed systematically by structural equation models with unmeasured variables to decompose response variances into valid, correlated error, and residual error components. This approach has been used previously by Andrews (1984) and Rodgers, Herzog, and Andrews (1988), and those articles contain more detail on the general rationale. In the first stage of the analysis to be reported below, such data quality estimates were obtained for more than one hundred survey measures included in several multitrait-multimethod models. In the second stage, quality estimates for those survey measures were analyzed in terms of characteristics of the measures and of the respondents in a form of meta-analysis.

This paper builds on Andrews's (1984) in several important respects. Improvements in the methodology are noted at appropriate points in the following section. A design feature that distinguishes this study is that whereas the data analyzed by Andrews were collected from standard samples of adult populations, the sample for the present study included a disproportionate number of older adults, with the objective of comparing the quality of data of younger and older adults more precisely than could be done by Andrews. There is a widespread opinion among both producers and consumers of survey research that the quality of data obtained from elderly respondents is lower than that of data from younger adults, and we have already noted that Andrews found evidence of an age-related decline in validity coefficients (Andrews 1984; Andrews and Herzog 1986).

2. Methods

2.1. Data base: Survey design and sample

The data analyzed here were collected as part of the Study of Michigan Generations project conducted by the Survey Research Center at the University of Michigan. Face-to-face interviews lasting an average of 90 minutes were conducted with adults randomly selected from a multi-stage stratified area probability sample of households in the Detroit metropolitan area (Wayne, Oakland, and Macomb counties). The population of this area is similar to that of the entire United States on many demographic characteristics including sex, age, education, and income. The data were collected in 1984 from 1,491 respondents. Because one of the project's goals was to examine how respondent age is related to measurement quality, older respondents were sampled with higher probabilities than younger people; 1,016 of the respondents were age 60 or older. Weights designed to take account of differences in the probability of observation were used in estimating the variance-covariance matrices used in the analyses reported in this paper.

The survey was designed to include measures of a wide range of concepts of interest to social scientists. In each of nine MTMM matrices, a small number of concepts ("traits") were assessed using several different question and answer formats ("methods"). It was the incorporation of these MTMM matrices within the larger survey that made possible the use of the structural equation model described below to generate estimates of measurement quality.

2.2. Data analysis: Stage 1 – obtaining measurement quality estimates

Measures for which quality estimates were obtained. A "measure" is an assessment of a particular concept using a particular measurement method. In all, there were 123 measures for which we generated quality estimates. The typical MTMM matrix included three to five concepts, three to four methods, and a total of about 14 measures (range: 8 to 29). An Appendix presents the concepts, the methods, and the number of measures in each of the nine MTMM matrices. The concepts varied widely, as did the methods. The methods included self-reports using different response scales and certain non-survey sources such as reports by interviewers, map distances, and information from the U.S. census. (Note, however, that although measures based on non-survey data were included in the Stage 1 analyses (i.e., in generating measurement quality estimates), only the estimates for 95 survey-based measures were used in the Stage 2 analysis, to be described below.)

Nature of the measurement model. The measurement quality estimates were derived from a structural model of the measurement process. This model is based on a set of causal assumptions grounded in classic measurement theory (Heise and Bohrnstedt 1970; Lord and Novick 1968). The integration of classic measurement theory with modern structural modeling to produce the kinds of parameters we shall discuss is a recent and powerful innovation. In accord with classic measurement theory (and with what seems to be intuitively reasonable), we assume that a respondent's recorded answer

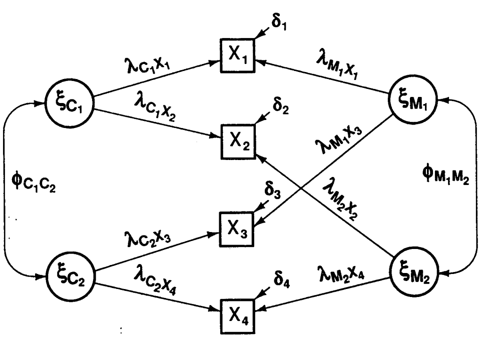

Fig. 1. Simplified structural equation model with substantive and method latent variables

to a survey question (labeled x_j for the response to the jth question; the subscript for the individual respondent is omitted to simplify the notation) is subject to three types of influences (see Figure 1): (a) the respondent's true level on concept c, labeled ξ_c (in Figure 1 two concepts are shown, labeled ξ_{C_1} and ξ_{C_2}); (b) the way that the respondent reacts to and uses the method (m) used for obtaining the data (particularly, in our case, the response scale), labeled ξ_m (in Figure 1 two methods are shown, labeled ξ_{M_1} and ξ_{M_2}); and (c) everything else that might produce variation in a recorded answer (e.g., lapses of memory by the respondent, misunderstanding by the interviewer, etc.), labeled as δ_j. These simple ideas can be represented by a causal model of the measurement process

$$x_j = \lambda_{cj}\xi_c + \lambda_{mj}\xi_m + \delta_j$$

where λ_{cj} is the loading of question j on factor c and λ_{mj} is the loading of question j on factor m. (We will refer to the λ_{cj} parameters, collectively, as the concept lambdas, and to the λ_{mj} parameters as the method lambdas.) The covariance of the two concept factors in Figure 1 is labeled as $\phi_{C_1C_2}$, and the covariance of the two method factors as $\phi_{M_1M_2}$. It is assumed that the method factors are unrelated to the concept factors. It is also assumed that the residual term, δ_j, is unrelated to either concept or method or to the residual terms for responses to other questions; the variance of this term is referred to as θ_j (collectively, these will be referred to as the theta parameters).

Some of the parameters in such a model can be interpreted as measurement quality assessments because they indicate (a) the extent to which variation in a given measure reflects differences among people with regard to the underlying concept, an approximation to construct validity (the concept lambdas); (b) the extent to which variation in the measure is influenced by differences among people in the way they react to the measurement method, a major source of correlated error (the method lambdas); and (c) the extent that variation in the measure reflects other influences, primarily random measurement error (the theta parameters). The phi parameters can be interpreted as estimates of the true relationships among the concepts and relationships among the method factors. Alwin (1974) provides a description of the method. Examples of its application and further discussion of its rational appear in Andrews and Withey (1976), and Andrews (1979, 1984). Rodgers et al. (1988) provide a detailed description of one of the nine MTMM matrices included in the present analysis.

Designating some of the parameters of this model (standardized λ_c values) as estimates of *construct validity* is in accord with established terminology (American Psychological Association 1974; Cronbach and Meehl 1955; Heise and Bohrnstedt 1970; Zeller and Carmines 1980). Construct validity is the (product moment) correlation between an observed measure and the theoretical construct (not directly observable) that the measure is intended to reflect. Construct validity is different from criterion validity (which is the correlation between the fallible measure and an observed but error-free criterion) and from any of several forms of reliability (all of which involve correlations between fallible measures).

Method of estimation. The process of estimating the structural model begins with a matrix of variances and covariances among the measures. (These were estimated using sample weights to take account of differences in the probabilities of observation of population elements.) We then sought estimates of the set of strengths for the causal linkages (the lambda parameters), correlations among latent concepts and among method factors (the phi parameters), and variances for the random error sources (the theta parameters) that would maximize the likelihood of the observed variances and covariances among the actual responses. These parameter values were obtained using the LISREL computer program (Jöreskog and Sörbom 1989). The variances of the factors are not estimable and, for convenience, were all set equal to 1.0 (this is merely a way of setting the units of the factors, and is equivalent to fixing the value of a lambda coefficient). Moreover, for the meta-analysis reported in this paper, the factor loadings (i.e., the lambdas) have been standardized by dividing each by the standard deviation of the measure so that each lambda value can be interpreted as the correlation between the measure and the standardized corresponding concept or method factor. In particular, the concept lambdas are estimates of the construct validities of the measures.

Because nine separate MTMM sets of data were used in the analyses reported here, nine distinct models were estimated – one for each MTMM matrix. The primary output was a set of three measurement quality estimates (estimates of valid variance, correlated error variance, and random error variance) for each of the 127 measures.

Evaluation of the parameter estimates. An important question that must be addressed before one considers the specific parameter estimates obtained from such a measurement model is how well the model works in general. Several considerations apply. First, the model must be theoretically reasonable. The same general form of model was specified for each of the nine MTMM data sets, and as indicated earlier, this model is a direct implementation of classic measurement theory. Second, the parameter estimates generated by the model should fall within legitimate ranges. This criterion was well, but not perfectly, met in the nine models estimated on the total set of respondents. Across these nine models, there were 480 parameters that were free to be estimated, and 99% of them (all but 6) had values within the legitimate ranges. Furthermore, none of the six out-of-range values was far out. (The "worst" was a random error variance estimated at −0.35.)

Third, the obtained parameters must be able to account for the variances and covariances actually observed among the measures; that is, the model must fit. These models fit the survey data well as shown by the summary statistics in Table 1. One useful indicator of model fit is the Jöreskog-Sörbom (1989) Adjusted Goodness-of-Fit Index (AGFI). High values of the AGFI indicate good fit, with 1.00 indicating perfect fit. These models produced an average AGFI of 0.964 (range: 0.921 to 0.993). A second indicator of fit, Hoelter's (1983) critical number (CN), averaged 504 across the nine models (range: 219 to 1,068). In every model, the CN exceeded Hoelter's recommended criterion of 200, in most cases by very substantial amounts. Hence this is another indicator of the satisfactory fits. Two additional goodness-of-fit indices

Table 1. Goodness-of-fit statistics for each of nine measurement models.

Model	χ^2	df	AGFI	CN	T-L	Δ_2
1. Political attitudes	199.71	39	.959	406.03	.901	.931
2. Frequencies of activities	89.72	32	.989	763.43	.991	.995
3. Everyday functioning	183.63	27	.949	324.18	.905	.944
4. Life quality	1243.97	248	.915	342.91	.947	.956
5. Distances	400.63	43	.924	219.32	.955	.971
6. Voting	123.06	36	.992	599.38	.992	.996
7. Morale	551.89	117	.921	386.45	.960	.969
8. Health and income	25.04	10	.985	1067.72	.996	.998
9. Neighborhood characteristics	433.26	100	.993	427.64	.866	.889

χ^2 — Chi-square goodness-of-fit statistic
df — Degrees of freedom
AGFI — Adjusted goodness-of-fit index (Jöreskog and Sörbom 1989, p. 27)
CN — Critical n (Hoelter 1983)
T-L — Tucker and Lewis incremental fit index (Tucker and Lewis 1973)
Δ_2 — Bollen's incremental fit index (Bollen 1989)

shown in Table 1, one proposed by Tucker and Lewis (1973) and the other by Bollen (1989), are both above 0.9 in value for all except one of the nine models. (The exception is a model for which the effective sample size is relatively small because it has to do with neighborhood characteristics. For a more detailed description of Stage 1 analyses see Rodgers et al. 1988.)

2.3. Data analysis: Stage 2 – linking measurement quality to measure and respondent characteristics

General strategy. The goal of the Stage 2 analysis was to account for variation in each of the three measurement quality estimates derived in Stage 1 by variation in survey and respondent characteristics. Specifically, we asked for which survey characteristics and for which types of respondent estimated validity was higher or lower than average, and what percentage of the variation in the validity estimates can be explained by those survey and respondent characteristics. Parallel questions were investigated for the estimates of both correlated and random error components. The general strategy was to use the measurement quality estimates as dependent variables and characteristics of the survey measures as independent variables. Because characteristics of respondents only accounted for about 1% of the variance in any of the quality estimates, this part of the analysis will be summarized rather briefly.

It is important to note that our analysis strategy involved a shift of cases between Stages 1 and 2. When the measurement quality estimates were being *generated*, the "cases" were individual respondents – as is conventional for most survey analyses. When the quality estimates were being analyzed, the "cases" were survey measures. As such, the cases that constituted the primary data base for the Stage 2 analysis consisted of 95 survey-based measures as answered by the entire group of 1,491 respondents. Associated with each of those cases were three estimates of its measurement quality (its valid, correlated error, and random error components) that formed the dependent variables and 15 variables describing the characteristics of the survey measure that formed the independent variables. Because of the way in which these data quality indicators were estimated, usual assumptions about the independence of their error terms are probably incorrect, but this is ignored in our analysis because there is no straightforward way to take it into account. The consequence may be that standard errors are underestimated.

Measure characteristics. The following characteristics of the survey measures were distinguished. (For reasons discussed later, not all were actually used.) Details about how these variables were coded are available from the authors.

Question characteristics:

- Whether question asks for a fact, an attitude, or something in between.
- Frame of reference (absolute, relative to the past or to other people).
- Recall period (past few weeks, past year, all other – including present conditions and unspecified time frame).
- Sensitivity of the topic to social desirability bias effects (low to high).
- Whether question wording includes an explicit attempt to reduce expected social desirability bias.
- Length of introduction to the question (number of words).
- Length of the question itself (number of words).

Response scale characteristics:

- Number of response categories.
- Whether an explicit "don't know" category was included.
- Whether response categories were fully labeled, partly labeled, or not labeled.
- Whether response categories were read to respondent, read in a sequential manner (unfolding scale), or presented in written form in the respondent booklet.

Context of question:

- Position in interview (number of questions in the interview preceding this question).
- Whether in a battery of questions (with common response scale) or not.
- Position in battery (number of questions in the series preceding this question).

The additivity assumption. We expected to find two distinct first-order interaction effects, and we performed a more general check for other non-additivities. Based on our expectations, two pattern variables were included as predictors in the Stage 2 analysis. One expected interaction involved the two predictors that reflected lengths – length of introduction to the question and length of question itself. We expected that highest validity would occur in the medium/medium combination or in a combination close to this, and that the short/short and long/long combinations would produce less good data. The second expected interaction involved the two social desirability predictors. If a measure was not subject to social desirability bias, then whether an attempt to ameliorate that bias was present should have no effect, but given high susceptibility social desirability bias, amelioration attempts were expected to have positive effects on validity.

The linearity assumption. For two of the intervally scaled predictor variables we expected non-linear relationships. We expected validity would rise but then reach an asymptote as the number of answer categories increased, and would rise and then decline as the measure was located farther back in the questionnaire. Accordingly, a quadratic term was included for each one to let us check our expectation.

Analysis procedure. In our analysis of the relationship of the measure characteristics to the measurement quality estimates, we assumed that the data quality estimates (i.e., the dependent variables) were intervally scaled, and examination of the distributions showed that they had little skew. We took cognizance of differences in the precision of the data quality estimates by using weighted least squares (WLS) procedures, specifying the reciprocals of the estimated standard errors of the measurement quality parameters as weights. These standard errors were produced as part of the output of the LISREL runs in the Stage 1 analysis. In effect, this procedure gave greater emphasis in the Stage 2 analysis to those questions for which we had evidence that their measurement quality was more precisely estimated compared to questions whose measurement quality was less reliably estimated.

2.4. Respondent characteristics

Although investigation of the effects of certain respondent characteristics on data quality was an important part of this research, we found, as noted earlier, that the effects were small, and hence do not stress them here. The fact that they were small, however, is itself an extremely interesting and potentially important result. Brief mention of the method by which they were investigated seems desirable.

The general strategy for investigating the relationships between respondent characteristics and measurement quality is to develop estimates of measurement quality for specific subgroups of respondents and then to compare them. The Stage 1 analysis described above was repeated for subgroups defined by the following characteristics: (1) age; (2) education; (3) sex; (4) race; (5) self-rated health; (6) amount of help provided by interviewers; (7) scores on a measure based on the Crowne-Marlowe (1964) social desirability scale; (8) scores on a memory test; and (9) scores on a measure of cognitive rigidity. (Age was used to define three subgroups; each of the other characteristics defined two.) In addition, we examined subgroups defined by *combinations* of age and the eight other characteristics in order to see whether age effects, if any, would differ according to (or could be explained by) the level of one or another of these other variables. A total of 5,435 measurement quality estimates for survey-based measures were obtained for these subgroups, and these were examined in a series of Stage 2 analyses parallel to those described above for the entire sample and provide the basis on which we can report that respondent characteristics have very little systematic effects on measurement quality indicators.

3. Findings

3.1. Overview of concept, method, and residual effects

An overall summary across all nine MTMM models concerning the sources of variance in the measures examined in this study is shown in Table 2. The first column provides data for the validity coefficients of these measures; that is, the standardized loadings of the items on the concept factors, as estimated by the lambda coefficients in LISREL analyses. The mean value of these concept lambda coefficients is 0.74 (0.73 if the estimates are weighted by the reciprocals of their standard errors) and their standard deviation is 0.18 (both weighted and unweighted). The mean proportion of explained variance, which is obtained by squaring the loadings of the concept lambdas, indicates that somewhat over half of the variance in the survey responses is related to the concepts.

The second column of Table 2 provides comparable information with respect to the loadings of these measures on the method

Table 2. Summary statistics for measurement quality estimates

	Concept lambdas	Method lambdas	Thetas
Unweighted estimates:			
Mean of estimates	.74	.28	.55
Mean proportion of explained variance	.57	.10	.32
Weighted estimates[1]:			
Mean of estimates	.73	.28	.56
Mean proportion of explained variance	.57	.10	.32

[1] Estimates of the measurement quality parameters were weighted by the reciprocal of the estimated standard errors.

factors (again, these are standardized lambda coefficients obtained from LISREL analyses). The average value of these loadings is 0.28 (with a standard deviation of 0.14), suggesting that overall about 10% of the variability in survey responses must be attributed to the characteristics of the response scales.

The average value of the standard deviation of the residual term (i.e., the square root of the thetas) is 0.55, and the standard deviation of these values is 0.18. This analysis suggests that about one-third of the variance in the measures is attributable to residual variance.

3.2. Survey characteristics: Variation in concept, method, and residual effects

The findings are summarized in Tables 3 and 4. The organization of those tables is as follows. Table 3 includes estimates of the total explanatory power of each predictor on each of the three types of measurement quality parameters. Table 4 provides greater detail about the form of the relationships, by showing bivariate and multivariate regression coefficients for each variable and each set of dummy variables. The entries in the left-hand side of Tables 3 and 4 refer to analyses in which the concept lambdas are the dependent variable; those in the middle refer to analyses of the method lambdas; and those in the right hand side refer to analyses of the thetas. In Table 3, the first column in each of these three sets (labeled "eta") contains measures of the strength of bivariate relationships for each question characteristic: correlation coefficients (for linear versions of variables treated as intervally scaled) or multiple correlation coefficients (for sets of dummy variables, or for linear plus quadratic terms to represent intervally scaled variables). The second column (labeled "beta") contains comparable measures of the strength of associations for each question characteristic, but now controlling on the other question characteristics included in the table. (For linear predictors, "beta" is the standardized multiple regression coefficient; for both linear and other predictors, beta-squared is the proportion of variance in the dependent variable that is explained by that predictor after taking account of the other predictors.) Where more than one procedure was explored to take account of a question characteristic (e.g., linear versus categorical), a single procedure was selected for this multiple regression analysis. Likewise in Table 4, the entries in the first column (of each set of four columns) are bivariate regression coefficients; those in the third column are multiple regression coefficients. (These coefficients, however, differ from usual regression coefficients for sets of dummy variables; they indicate deviations from the mean rather than from an omitted category.) Entries in the second and fourth columns are the standard errors of the regression coefficients in the column to their left.

In these models we had constrained the factor loadings on each method factor to be equal for all measures using that method. This means that we did not obtain separate estimates of method loadings for each of the 95 measures used in the measurement models, but only for each of 23 methods. Because of this small number of independent estimates, few measure characteristics have statistically significant bivariate relationships to the method lambdas, and *none* of the multiple regression coefficients is statistically significant. Our analysis has such low power that even strong predictors of method loadings may not show up as statistically significant.

Overall, the survey question characteristics account for over half of the variance in the concept lambdas and the thetas, suggesting that a majority of the variability in data quality can be accounted for by aspects of the instrument design. Only a few percentage points of the variance in the method lambdas are accounted for by question characteristics.

A large amount of information is contained in Tables 3 and 4, and space constraints do not allow detailed discussion of most of the entries. Our strategy is to consider one predictor – the number of response alternatives (which is the most powerful predictor of the data quality measures) – in some detail, but to focus only on the highlights of other predictors.

Number of response alternatives. As shown in Table 4, the average number of categories of the 95 measures used in the measurement models is 5.2: 15 measures with *just two* alternatives (e.g., "Yes" or "No"), 34 with *seven or more* alternatives, and the remaining 46 with *four to six* alternatives. By itself, this variable explains about 22% of the variance in the concept lambdas (i.e., 0.47^2 from Column 1 in Table 3). Moreover, the relationship between the number of categories and the loadings on the concept factors is positive and close to linear (within the range of categories investigated here) – note that the curvilinear and categorical variables add little or nothing to the explained variance (Table 3). Every additional response alternative is associated with an increase of 0.037 in the expected value of the validity coefficient, as shown by the bivariate regression coefficient (Column 1, Table 4).

The multiple regression coefficient for the number of response categories with respect to the concept lambdas is shown in the third column of entries in Table 4 to be 0.064, almost twice as large as the bivariate regression coefficient. (Only the linear term is included among the predictors, because as noted before the quadratic term added no explanatory power and the linear term alone has almost as much explanatory power as the set of dummy variables.) This implies that the bivariate relationship between these two variables is partially suppressed by their relationships with the other measure characteristics; we take the multiple regression coefficient as the better indicator of the expected consequence of increasing the number of response alternatives on measurement validity while holding other measure characteristics constant. The absence of a significant coefficient for the quadratic term suggests that this pattern holds at least across the range of this variable (2 through 10) in our measurement models. We strongly suspect, however, that increasing the number of categories much beyond 10 would have

Table 3. Estimates of bivariate etas and multivariate betas

	Concept Lambda		Method Lambda		Theta	
	Eta	Beta	Eta	Beta	Eta	Beta
Question characteristics:						
Type of question (categorical)	.20*	.00	.21	.21	.30**	.13
Frame of reference (categorical)	.15*	.07	.00	.00	.07	.08
Recall period (categorical)	.18*	.48*	.51**	.97	.25**	.63**
Social desirability:						
Question	.23**		.00	.00	.13	
Attempt to reduce	.13		.25	.00	.11	
Additive	.32**		.19		.22**	
Interactive	.29**	.38	.41		.24*	.54**
Length of introduction and question						
Introduction	.05	.23**	.00	.00	.00	
Question	.19**	.30**	.00	.00	.00	
Additive	.16		.00		.00	
Interactive	.35**		.00		.18	.24**
Response scale characteristics:						
Number of categories						
Linear	.47**	.83**	.00	1.09	.35**	.88**
Linear + curvilinear	.47		.11		.37**	
Categorical	.49**		.00		.43**	
Explicit “Don’t Know” (categorical)	.43**	.29*	.00	.00	.36**	.28*
Response labels (categorical)	.19*	.43**	.00	.76	.28**	.29**
How presented (categorical)	.39**	.18	.20	.00	.24**	.00

Table 3 (cont).

	Concept Lambda		Method Lambda		Theta	
	Eta	Beta	Eta	Beta	Eta	Beta
Context of question:						
Position in interview						
Linear	.12	.32**	.00	.00	.33**	.39**
Linear + curvilinear	.23**		.00		.32	
Categorical	.00		.12		.21**	
Position in battery						
Linear	.00	.00	.17	.00	.00	.04
Categorical	.07		.00		.15	
Number of estimates		95		23		94
Adjusted R-square		.55		0.038		0.62
F-statistic		6.53		1.05		6.95
Degrees of freedom		21, 73		17, 5		26, 67
p-value		0.00		0.53		0.00

**p < .05, *p < .10

Notes: 1. All entries are adjusted for degrees of freedom.
2. For specific categories distinguished on qualitative variables, see Table 4.

Table 4. *Bivariate and multiple regression coefficients*

	n/	Concept Lambdas				Method Lambdas				Thetas			
		Bivariate		Multiple		Bivariate		Multiple		Bivariate		Multiple	
	mean	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E	*b*	S.E.	*b*	S.E.
Question characteristics:													
Type of question													
Factual	21	.05*	.03	.05	.10	.05	.06	−.10	.18	−.11**	.03	−.06	.11
Mixed	41	.01	.02	−.01	.04	−.06*	.03	.09	.10	.03*	.02	−.00	.05
Attitudinal	33	−.05**	.02	−.01	.04	.05	.04	−.05	.09	.03	.03	.05	.05
Frame of reference													
Absolute, or none given	88	−.01*	.01	−.01	.00	.01	.01	.00	.01	.01	.01	.01	.01
Comparative	7	.10*	.06	.07	.05	−.06	.11	−.03	.14	−.07	.06	−.08	.06
Recall period													
Present	60	−.03**	.01	−.04*	.02	.03	.02	.03	.06	.02	.01	.08**	.03
Weeks	18	.06*	.03	.16**	.08	−.16**	.05	−.25*	.12	.04	.04	−.21**	.09
One year or more	17	.03	.04	−.06	.10	.08	.06	.20	.19	−.11**	.04	−.07	.11
Social desirability of topic	8.0	.018**	.007			−.003	.014	−.009	.029	−.014	.009		
Attempt to reduce soc. des.	5.9	−.010	.006			.017	.011	.012	.029	.011	.007		
Additive effects:													
Social desirability		.023**	.007			−.009	.014			−.019**	.009		
Attempt to reduce		−.016**	.006			.019	.011			.016**	.008		
Interactive effects,													
soc. des./attempt to reduce													
Low/Any	16	−.10**	.04	−.01	.06	−.02	.07			.07	.04	−.02	.07
Med/Low	21	.05	.03	−.04	.05	−.08	.06			.02	.04	.10*	.06
Med/Med	17	.06*	.03	−.03	.06	.04	.06			−.08**	.03	.06	.06
Med/High	15	−.06*	.04	.08	.05	.13**	.06			−.03	.03	−.17**	.05
High/Low	11	.06	.04	−.07	.06	−.14*	.07			.02	.05	.11	.07
High/Med	6	−.00	.07	−.11	.08	.13	.11			−.03	.08	.12	.08
High/High	9	−.04	.05	.17*	.09	−.01	.08			.12**	.06	−.09	.12
Length of introduction	29.4	.0008	.0007	−.0017**	.0008	−.0009	.0013	−.0009	.0027	.0004	.0007		

Table 4. *(cont)*

	n/	Concept Lambdas				Method Lambdas				Thetas			
		Bivariate		Multiple		Bivariate		Multiple		Bivariate		Multiple	
	mean	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E	*b*	S.E.	*b*	S.E.
Length of question	18.6	−.0047**	.0022	−.0070**	.0027	.0007	.0040	.0028	.0069	.0010	.0025		
Additive effects:													
Length of introduction		−.0000	.0008			−.0011	.0016			.0008	.0009		
Length of question		−.0047*	.0026			−.0013	.0049			.0025	.0030		
Interactive effects, length of introduction/question													
None/15-20	7	.09	.06			−.03	.13			−.05	.07	.00	.05
None/21+	12	−.09**	.04			.15*	.08			−.03	.04	−.08	.05
1-34/ < 15	5	.02	.07			.11	.13			−.08	.06	−.15**	.06
1-34/15-20	10	−.01	.05			.07	.10			.02	.05	−.02	.05
1-34/21+	16	−.11**	.04			−.01	.07			.12**	.05	.03	.06
35+/ < 15	24	.01	.03			−.03	.05			.03	.03	.03	.03
35+/15-20	15	.09**	.04			−.10	.07			−.06	.04	.07**	.04
35+/21	6	.02	.06			−.09	.12			.02	.07	.10	.06
Response Scale Characteristics:													
Number of categories													
Linear alone	5.2	.037**	.007	.064**	.017	−.009	.016	−.077	.042	−.028**	.007	−.068**	.018
Curvilinear													
Linear		.039**	.008			−.016	.016			−.023**	.008		
Quadratic		−.0022	.0025			.0078	.0056			−.004	.003		
Categorical													
2	15	−.11**	.03			.10	.07			.02	.03		
4	23	−.09**	.03			−.01	.05			.15**	.04		
5-6	23	.04	.03			−.06	.05			.00	.03		
7 or more	32	.08**	.02			.00	.05			−.08**	.02		
Explicit DK category													
Yes	18	−.16**	.03	−.11*	.06	.05	.07	−.10	.11	.16**	.04	.13*	.08
No	77	.03**	.01	.02*	.01	−.01	.01	.02	.02	−.03**	.01	−.02*	.01

Table 4. (cont)

		Concept Lambdas				Method Lambdas				Thetas			
	n/	Bivariate		Multiple		Bivariate		Multiple		Bivariate		Multiple	
	mean	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E	*b*	S.E.	*b*	S.E.
Category labels													
All categories labeled	76	−.02**	.01	.02	.03	−.02	.02	−.06	.05	.03**	.01	.01	.04
Some categories labeled	8	.09	.05	.08	.11	.04	.10	.21	.22	−.08	.05	−.17	.11
No categories labeled	11	.07	.04	−.19	.12	.10	.09	.27	.20	−.10**	.04	.07	.11
How presented													
Read to respondent	35	−.09**	.02	−.03	.05	.06	.04	−.04	.11	.06**	.02	.02	.06
Unfolding format	15	.02	.04	.10*	.05	−.05	.07	.03	.13	.01	.05	−.05	.07
Respondent booklet	45	.06**	.02	−.00	.03	−.04	.03	.02	.09	−.05**	.02	−.00	.05
Context of Question:													
Position in interview													
Linear alone	119	.0004	.0003	.0009**	.0004	.0004	.0005	.0009	.0009	−.001**	.0003	−.001**	.000
Curvilinear													
Linear		.0003	.0003			.0005	.0005			−.001**	.0003		
Quadratic		.0000**	.0000			−.0000	.0000			−.000	.0000		
Categorical													
Early	13	−.02	.04			−.09	.08			.11**	.05		
Mid to late	82	.00	.01			.01	.01			−.02**	.01		
Position in battery													
Linear	3.2	.0055	.0077	−.0078	.0085	−.017	.013	−.0044	.0214	.007	.009	.011	.009
Categorical													
1st (or not in battery)	33	−.03	.02			.04	.05			.03	.03		
2nd to 4th	37	.03	.02			.01	.04			−.04**	.02		
5th to 9th	25	.00	.03			−.05	.05			.04	.03		

**p < .05, *p < .10

diminishing effects on the validity of the answers.

The bivariate relationship of the number of response categories with the method lambdas is statistically non-significant, regardless of how it is entered: linear, curvilinear, or categorical (Column 3, Table 3). The bivariate regression coefficient is faintly negative, but the multiple regression coefficient of -0.077, while not statistically significant because of the small sample size, is relatively large suggesting that increasing the response alternatives may decrease method effects.

Increasing the number of response alternatives also has a negative effect on the thetas (i.e., it reduces residual error) and in this case the bivariate and multiple regression coefficients are both statistically significant. As with the concept and method lambdas, the multiple regression coefficient is larger (in absolute value) than the bivariate regression coefficient. The bivariate pattern using dummy variables suggests that random error is greatest for measures with four response alternatives, but this departure from linearity is not statistically significant when other measure characteristics are controlled (not shown).

Type of question. We now comment on the remaining measure characteristics and their estimated effects on the measurement quality indices. The first set of measure characteristics listed in Tables 3 and 4 consists of those related to the content of the questions, starting with whether the question asks for factual information or for an attitude or feeling. The bivariate regression coefficients indicated that answers to attitudinal questions are less valid and more strongly affected by method factors than are factual or "intermediate" questions, but this pattern diminishes upon controlling for the other measure characteristics in the multiple regression analysis. Factual measures do appear to have somewhat smaller proportions of residual variance, significantly so in the bivariate analysis, non-significant but in the same direction for the multiple regression analysis.

Frame of reference. Most of the questions did not provide the respondents with any explicit frame of reference, either because the respondents were expected to answer on an absolute scale rather than comparatively or because the respondents were allowed to adopt their own frame of reference. Seven of the questions, however, included explicit frames of reference: the respondents were asked to compare themselves, for example, to others their own age or to themselves five years previously. For our sample the latter measures have higher validity coefficients, are less affected by method factors, and have less residual variance than do the former.

Recall period. This characteristic has three categories: (1) "Present," which for many measures means that no specific time referent is stated (the questions often refer to the rather amorphous "extended present"); (2) "Weeks," which means that the question asks about events or feelings that have occurred specifically in the last few weeks; and (3) "1+ Years," typically for questions that ask about events that have occurred in the last 12 months. The relationships are not all statistically significant, but measures asking about recent weeks appear to be considerably better than other measures: their validity estimates tend to be higher, they are less affected by method factors, and they have a smaller proportion of residual variance.

Social desirability effects. Questions differ in the extent to which certain responses may be judged to be more socially desirable than other responses. Contrary to our expectations, however, we found no evidence that measures judged to be more subject to social desirability effects were necessarily any less valid than other measures, nor that attempts to counteract such effects always improved validities. In fact, the bivariate regression coefficients are statistically significant, and in the opposite direction from those expectations. However, because we expected that these two characteristics would interact in their effect on the quality measures, we also created the pattern variable shown in Table 4. The interaction is not statistically significant, but an interesting trend is noticeable across the multiple regression coefficients: measures of high and medium level of social desirability show lower validities (concept lambdas) and higher random errors (thetas), *unless* they were accompanied by an attempt to mitigate the social desirability demand characteristic.

Length of question and length of introduction. In the context of the present study, longer (more wordy) questions and longer introductions to sets of questions are associated with *lower* validities than questions with as few as 15 words, especially after controlling on other measure characteristics.

Explicit "Don't Know" response alternative. One characteristic of response scales is whether or not the respondents were explicitly reminded that "Don't Know" (DK) is an alternative to the substantive choices. Among the 95 measures used in the measurement models for this study, 18 had explicit DK alternatives. Our analysis indicates that measures with the explicit DK option have *lower* validities than do other measures, and *higher* measure-specific variances. These bivariate differences diminish (and are no longer statistically significant) after controlling on the other measure characteristics, but the direction of the differences persists.

Labeling of response alternatives. Whether response alternatives were given verbal or graphic labels is not significantly correlated to validity when examined by itself, but in the multivariate analysis measures with all graphic scales appear to be less valid than measures with verbal labels for some or all of the alternatives. Although not statistically significant, measures with verbal labels for all alternatives have lower loadings on method factors than do other measures. Also, according to the bivariate coefficients, measures with all non-verbal alternatives have the lowest proportion of random error, but this pattern disappears upon controlling for other characteristics.

Manner of presentation of response scale. The last of the set of four characteristics of response scales that we examined is the manner in which those scales are presented to the respondents. For 35 of the measures, the response alternatives were simply read by the interviewer as an extension of the question itself. These measures had below-average validity estimates, suggesting that the respondents may have trouble retaining the content of the question while processing the response scale. Almost half (45) of the measures were accompanied by response scales that were printed in a booklet placed in the hands of the respondents. The bivariate pattern suggested above average validities for these measures, but this apparent advantage largely disappears in the multiple regression analysis. The remaining 15 measures had response scales that were presented by "unfolding": that is, a series of questions with just two alternatives each were read to the respondents and in this way the differentiation in the responses was built-up in a sequential manner. The multiple regression coefficient for this type of measure indicates that this technique produces the highest validities, but none of those patterns is statistically significant.

Position in interview. The third set of measure characteristics listed in Tables 3 and 4 consists of those referred to as contextual.

The first in this set is the position of the question in the interview, measured in terms of number of questions that preceded a particular question. The questions used in the various measurement models had positions that ranged from 8th to 205th, with an average value of 117 and a standard deviation of 65. The regression coefficients in Table 4 indicate that answers to questions asked later in the interview are *more* valid, and have a lower proportion of residual variance than answers to earlier questions.

Position in battery. The other contextual characteristic of measures that we considered is whether or not each measure was part of a battery of parallel items using the same response scale, and if so what its position was in that battery. No statistically significant differences were found, and in particular there is no evidence that placing a question in a battery has a detrimental effect on the quality of answers to that question.

4. Discussion

The findings presented in this paper have several implications, including suggestions for the design of surveys, implications for the quality of the data obtained from specific subgroups of respondents, and further development of a methodology to investigate construct validity. We will discuss each in turn.

4.1. Implications for survey design

This research showed that over half of the variance in answers to survey questions, obtained through face-to-face interviews, can be considered valid (or at least internally consistent), because it can be accounted for by the nature of the concepts that are being measured. The mean (standardized) loading of the measures on the concept factors was 0.74. The different concepts that were examined in this research were familiar ones to social scientists such as health and functioning, activities, attitudes, satisfactions, and economic resources.

Another 10 to 15% of the variance in the responses is attributable to characteristics of the question and response scales and consistent ways in which individuals answer the question. While this is not an inordinate amount, it reminds us that similarity of question formats may introduce a positive bias component into estimates of relationships among variables. Whereas it is commonly thought that measurement error is purely random and thus introduces a negative bias into estimates of relationships among variables, which can be compensated for by correcting for attenuation, one needs to recall that a positive correlation between errors in measures of two variables will tend to inflate the relationship between the underlying variables (if that relationship is also positive).

The amount of residual error – about one-third of the variance in responses to survey questions – is likely to contribute a negative bias component because it is not obviously related to the error structure of other questions and their responses.

In sum, these findings indicate that a sizeable portion of the variance in survey responses is systematically related to the concept being measured but that correlated and residual error are not negligible and should be considered when interpreting results obtained with such measures. Correlated error and residual error are likely to exert opposing influences on relationships among survey variables. In one of the models on which our meta-analysis is based, we found that the negative bias from residual error slightly outweighed the positive bias from correlated error and thus that the (uncorrected) intercorrelations among measures were somewhat lower than the (corrected) intercorrelations among concept factors (Rodgers et al. 1988). The direction and magnitude of the net bias will depend on the specific concepts and measures involved in a particular analysis. Zeller and Carmines (1980) present a general formula to adjust observed correlations for validity of measures as well as for correlated error between them. Andrews (1984) provides further discussion of this adjustment and specific examples.

The average results hide, of course, considerable variation across measures. In our meta-analysis we attempted to identify characteristics of measures that account for these variations. The strongest effect was noted for the number of response alternatives: scales with higher numbers of response alternatives produced responses with a higher proportion of valid variance and a lower proportion of residual error variance than scales with fewer alternatives. Several other aspects of the response scale also seemed important. Response scales that were "unfolded" step-by-step produced more valid responses than those which simply called for reading all alternatives at one time or for presenting them spelled-out in a respondent booklet. Moreover, when a "don't know" (DK) response alternative was explicitly permitted, responses of lower validity and higher residual error resulted than when respondents received no DK alternative.

Among characteristics of the question, a comparative phrasing produced more valid responses than an absolute one, and a shorter recall period less residual error than a long one. Questions with a specific recall period (of about a week) also produced data of better quality than did questions asking about the amorphous present. This may be interpreted as indicating the superiority of providing a specific frame of reference (parallel to the relative versus the absolute frame of reference) rather than a consequence of the length of the recall period.

With respect to the context of questions within the interview, questions which are placed later in the interview produced responses that are more valid and contain less residual error.

The extent to which our findings can be generalized is limited by at least two considerations: the data were collected from a local sample by face-to-face interviews at a particular time; and it was necessary to make assumptions about the data in implementing our analyses. We cannot directly assess the extent to which data from a different population, by a different mode, or in a different year would yield different findings. However, the absence of any large or statistically significant differences related to respondent characteristics suggests that the findings can probably be generalized to other populations. Second, it is possible that our findings are sensitive to the specific assumptions that we made in our analyses. For example, in our estimation of the MTMM models we assumed that method factors are unrelated to concept factors, that all items employing a particular method are equally affected by that method factor, and that residual terms are uncorrelated with one another. We found that we needed to impose these restrictions to obtain plausible estimates of the model parameters, and we are unable to assess the extent to which they may have introduced biases into our estimates.

The generality of the present findings is strengthened by their consistency with those reported by Andrews (1984), which are based on different sets of variables from different studies. There are also differences between the findings from the present study and that by Andrews, some of which can be explained by differences in procedures but others of which are not so easily explained.

The primary difference between this study and the prior study by Andrews is that a higher proportion of the measure variances in this study is explained by method factors and a correspondingly lower proportion by the concept factors. Such a difference in susceptibility could reflect nothing more than the choice of concepts or the choice of methods in the two studies. It is also possible, however, that an analytic difference may have contributed to this discrepancy: in Andrews's measurement models the method factors were constrained to be independent of one another, whereas in the present study this constraint was not imposed.

Some, though not all, of our findings with respect to measure characteristics that are related to data quality estimates are consistent with findings reported by Andrews (1984). In particular, the number of response categories emerged as the most important predictor in Andrews's study as well as in the present study and, furthermore, is consistent with a body of related research that was noted briefly at the outset of this paper. A strong recommendation for using at least five to seven (rather than two to four) response alternatives when devising response scales can therefore confidently be made. This recommendation must be underscored because response categories are often kept to a minimum in order to ease respondent burden. A casual examination of many survey instruments reveals a large proportion of questions with response scales containing four or fewer alternatives.

The recommendation with respect to number of response alternatives receives some additional support from our finding that the unfolding format produced responses of relatively high validity. This format was designed to make complex response scales more palatable. And despite what might appear to be the case, unfolding scales do not take much more time to answer because they follow a colloquial speech pattern.

Our findings also lead us to a strong recommendation for the use of relative rather than absolute question wordings. The superiority of the relative frame of reference was also reported by Andrews.

Other findings are not consistent with those by Andrews (1984) and deserve further research to clarify their effect. They include his intuitively reasonable finding that questions in the middle of the interview yielded responses of better quality than those at the very beginning or towards the end. The current investigation found response quality to improve throughout the course of the interview. We cannot explain these findings but note that they are not related to the respective lengths of the interviews. Neither did we replicate Andrews's intuitively reasonable finding about the detrimental effect of a position late in a battery. Moreover, a puzzling finding from Andrews's prior research – that fully labeled response categories resulted in lower data quality than partially labeled ones – also did not replicate and thus might be considered with caution.

Our finding that respondent characteristics were virtually unrelated to the data quality indicators is generally consistent with the report by Andrews (1984) who found that respondent characteristics explained much less variance in data quality estimates than did characteristics of the measures. The personal characteristic that was of most importance in Andrews' findings was chronological age. More specifically, Andrews and Herzog (1986) found evidence that data quality is lower for responses from older respondents. The age patterns observed in the present study are generally consistent in direction with the previous finding, but fall short of statistical significance.

4.2. *Implications for special subgroups*

The lack of any substantial variation by subgroups was somewhat surprising, although not totally unexpected in light of Andrews's previous findings and the assessment by Bradburn (1983) that respondent characteristics account for much less variance in survey response quality than do survey characteristics.

In this project we had a particular interest in the age of the respondent and how it affects response quality. We observed even smaller age differences than in our previous work which had documented small but consistent age differences (Andrews and Herzog 1986; Rodgers et al. 1988), but the pattern was consistent with our previous findings. The good news of this finding is, of course, that often-voiced concerns about difficulties experienced by older respondents in responding to surveys are unfounded. Particular note should be taken of the finding that the beneficial effect of the number of response categories is not restricted to younger respondents. Unlike what had been hypothesized previously (Lawton 1977), the validity of answers by older respondents increases equally with an increasing number of response alternatives as does the validity of answers by younger adults.

To summarize, we concur with previous observations that despite great research efforts very little evidence can be marshaled for individual differences in data quality. This assessment should be qualified, however, because of investigations in which we and others have found that sampled respondents who did not participate in the survey were systematically different from those who did. They were less healthy and less interested in the topic of the survey, both characteristics that are known to be related to response quality (Bergstrand, Vedin, Wilhelmsson, and Wilhelmsson 1983; Criqui, Barrett-Connor, and Austin 1978; Herzog 1987). If a more complete cross-section of the population was reached in standard surveys, more substantial data quality effects of individual difference variables might well be found.

4.3. *Implications for methodology*

A construct validation approach is the only form of validation possible for measures of concepts that are subjective in nature and thus have no obvious external referents for validation. The familiar forms of construct validation such as correlating the new measure of a concept with a more established measure or correlating the measure with known predictors leave the nagging question whether part of the relationship may be caused by the often similar format of the questions to be validated and of the criterion or predictor measures. This concern highlights the chief advantage of the method used here: the types of response scales are modeled as separate unmeasured "methods" and thus are controlled when construct validity is investigated. As such, our approach also represents a systematic implementation of Campbell and Fiske's (1959) call for measuring multiple concepts with multiple measures.

We believe that the method described in this paper deserves wider application for the investigation of construct validity. While in the present study a generic application was performed and measures were chosen that would be representative of measures used in many social science research projects, the method might also be used by a study focusing on a particular issue, say social relationships between parents and grown children or frequency of leisure activities. In such a study the core measures might be evaluated by collecting data on them according to a multitrait-multimethod matrix and submitting the data to the analysis performed at Stage 1 of the present study.

For studies that are unable to perform their own validity investigation, the findings from this study can be used as guidelines. This is particularly true because no effects of the content of the question – attitudes versus facts – was found and thus the results are likely to be generalizable beyond specific question contents.

Thus, when designing a new survey instrument, the results from this study should be taken into consideration. For example, questions about feelings and evaluations should be phrased in relative rather than absolute terms and response scales should contain at least five categories. When using existing questions, adjustments of bivariate and multivariate relationships may be made to take account of the validity, correlated and random error of a particular form of measurement device.

Appendix

Concepts and Methods Included in the Nine MTMM Data Sets

1. *Political attitudes.* Concepts: trust in government; trust in people; negative sentiment toward aging. Methods: 4-pt. agree/disagree scale with positively worded questions; 4-pt. agree/disagree scale with negatively worded questions; yes/no scale. Measures: 11.
2. *Frequencies of activities.* Concepts: seeing physicians; attending religious services; contacting relatives, children/parents, friends; attending organizational functions. Methods: 6-pt. times in the past year; 7-pt. average frequency per week/month/year; 7-pt. average frequency per week/month/year in grid format. Measures: 11.
3. *Everyday functioning.* Concepts: ability to see, hear, move, remember. Methods: ratings by self or interviewer on 5-pt. scale of difficulty; functional test. Measures: 10.
4. *Life quality.* Concepts: assessments of health, housing, income, savings, transportation, safety from crime, friends, life-as-a-whole. Methods: 7-pt. satisfaction scale; 7-pt. sad/happy faces scale; 4-pt. worry scale; 5-pt. poor/excellent scale; 4-pt. extent of problems scale; 10-pt. ladder scale. Measures: 29 (4 of these measures also appear in #8).
5. *Distances.* Concepts: distances to nearest drugstore, grocery store, fire station, hospital. Methods: respondent estimate; mean miles estimated by all respondents in same neighborhood; map miles. Measures: 12.
6. *Voting.* Concepts: whether voted in President election in 1980, in Congressional election in 1982, for School Board election in 1983; whether registered to vote in 1980, 1982, 1983. Methods: self-report; election records. Measures: 12.
7. *Morale.* Concepts: depression; anxiety; attitude about aging; psychosomatic symptoms. Methods: 4-pt. agree/disagree (positive and negative); yes/no (positive and negative); 5-pt. frequency (positive and negative). Measures: 18.
8. *Health and Income.* Concepts: absolute level of own health; own health relative to others' health; absolute level of own income; own income relative to others' incomes. Methods: respondent's rating on 5-pt. excellent/poor scale; respondent's rating on 10-pt ladder. Measures: 8 (4 of these measures also appear in #4).
9. *Neighborhood characteristics.* Concepts: % aged 60+; % black; % with incomes >$10,000; % with incomes >$30,000.

Methods: respondent's own estimate; mean estimate from respondents living in the same neighborhood; 1980 census data; data from respondent screening operations by interviewers. Measures: 16.

5. References

Alwin, D.F. (1974) Approaches to the Interpretation of Relationships in the Multitrait-Multimethod Matrix. In Sociological Methodology 1973–74, ed. H.L. Costner, San Francisco: Jossey-Bass.

Alwin, D.F. (in press). Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement. In Sociological Methodology 1992, ed. P.V. Marsden, Oxford: Blackwell.

Alwin, D.F. and Krosnick, J.A. (1991). The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. Sociological Methods and Research, 20, 139–181.

American Psychological Association (1974). Standards for Educational and Psychological Tests. Washington, DC: the author.

Andrews, F.M. (1979). Estimating the Construct Validity and Correlated Error Components of the Rated Effectiveness Measures. In Scientific Productivity, ed. F.M. Andrews, Cambridge: Cambridge University Press/UNESCO.

Andrews, F.M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. Public Opinion Quarterly, 48, 409–442.

Andrews, F.M. and Herzog, A.R. (1986). The Quality of Survey Data as Related to Age of Respondent. Journal of the American Statistical Association, 81, 403–410.

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

Bergstrand, R., Vedin, A., Wilhelmsson, C., and Wilhelmsson, L. (1983). Bias Due to Non-Participation and Heterogeneous Sub-Groups in Population Surveys. Journal of Chronic Diseases, 36, 725–728.

Blair, E., Sudman, S., Bradburn, N.M., and Stocking, C.B. (1977). How to Ask Questions About Drinking and Sex: Response Effects in Measuring Consumer Behavior. Journal of Marketing Research, 14, 316–321.

Bollen, K.A. (1989). A New Incremental Fit Index for General Structural Equation Models. Sociological Methods and Research, 17, 303–316.

Bollen, K.A. and Barb, K.H. (1981). Pearson's *r* and Coarsely Categorized Measures. American Sociological Review, 46, 232–239.

Bradburn, N.M. (1983). Response Effects. In Handbook of Survey Research, eds. P.H. Rossi, J.D. Wright, and A.B. Anderson, New York: Academic Press.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. Psychological Bulletin, 56, 81–105.

Cochran, W.G. (1968). The Effectiveness of Adjustment by Subclassifications in Removing Bias in Observational Studies. Biometrics, 24, 295–313.

Cox, E.P. III (1980). The Optimal Number of Response Alternatives for a Scale: A Review. Journal of Marketing Research, 17, 407–422.

Criqui, M.H., Barrett-Connor, E., and Austin, M. (1978). Differences Between Respondents and Non-Respondents in a Population-Based Cardiovascular Disease Study. American Journal of Epidemiology, 108, 367–372.

Cronbach, L.J. and Meehl, P.E. (1955). Construct Validity in Psychological Tests. Psychological Bulletin, 52, 281–302.

Crowne, D. and Marlowe, D. (1964). The Approval Motive. New York: Wiley.

Heise, D.R. and Bohrnstedt, G.W. (1970). Validity, Invalidity, Reliability. In Sociological Methodology 1969, eds. E.P. Borgatta and G.W. Bohrnstedt, San Francisco: Jossey-Bass.

Herzog, A.R. (1987). Nonresponse in Sample Surveys of Older Adults. Paper presented as part of the Symposium, Interviewing Older Adults: Sources of Measurement Error, 40th Annual Scientific Meeting of the Gerontological Society of America, Washington, D.C.

Herzog, A.R. and Bachman, J.G. (1981). Effects of Questionnaire Length on Response Quality. Public Opinion Quarterly, 45, 549–559.

Hoelter, J.W. (1983). The Analysis of Covariance Structures: Goodness-of-Fit Indices. Sociological Methods and Research, 11, 325–344.

Jöreskog, K.G. and Sörbom, D. (1989). LISREL 7 User's Reference Guide. Mooresville, IN: Scientific Software, Inc.

Lawton, M.P. (1977). Morale: What Are We Measuring? In Measuring Morale, ed. C.N. Nydegger, Washington, D.C.: Gerontological Society.

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

McClendon, M.J. and Alwin, D.F. (1990). No Opinion Filters and Attitude Reliability. Paper presented at the Annual Meetings of the American Association of Public Opinion Research. Lancaster, PA.

Marquis, K.H. and Cannell, C.F. (1971). Effects of Social Reinforcement on Reporting in the Health Interview. Vital and Health Statistics, National Center for Health Statistics, DHEW Publication No. 1000, Series 2, No. 41. Washington, D.C.: U.S. Government Printing Office.

Neter, J. and Waksberg, J. (1964). A Study of Response Errors in Expenditures Data from Household Interviews. Journal of the American Statistical Association, 59, 17–55.

Rodgers, W.L., Herzog, A.R., and Andrews, F.M. (1988). Interviewing Older Adults: Validity of Self-Reports of Satisfaction. Psychology and Aging, 3, 264–272.

Schuman, H. and Presser, S. (1981). Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. New York: Academic Press.

Sudman, S. and Bradburn, N.M. (1974). Response Effects in Surveys: A Review and Synthesis. Chicago: Aldine.

Tucker, L.R. and Lewis, C. (1973). A Reliability Coefficient for Maximum Likelihood Factor Analysis. Psychometrika, 38, 1–10.

Zeller, R.A. and Carmines, E.G. (1980). Measurement in the Social Sciences. Cambridge: Cambridge University Press.

Received January 1991
Revised February 1992