Strube, G. (1987). Answering survey questions: The role of memory. In H.J. Hippler, N. Schwarz, and S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 86-101). New York: Springer Verlag.

Sudman, S., Bradburn, N., and Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Wagenaar, W.A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology, 18*, 225-252.

Withey, S.B. (1954). Reliability of recall of income. *Public Opinion Quarterly, 18*, 31-34.

Optimizing Brief Assessments in Research on the Psychology of Aging: A Pragmatic Approach to Self-Report Measurement

Jon A. Krosnick
Stanford University
and

Allyson L. Holbrook
University of Illinois at Chicago
and

Penny S. Visser
University of Chicago

INTRODUCTION

In a great deal of research on the psychology of aging, as in most empirical social science, data are routinely collected from research participants via questionnaires. In the September 2004 issue of *Psychology and Aging*, for example, 16 articles appear in print, and 10 of them employed questionnaires assessing self-reports of stereotypes, confidence, depression, anxiety, social network structure, subjective well-being, positive affect, stress, and much more. And as is true of most questionnaire-based research, research on older adults often employs long batteries of questions, each intended to precisely measure a particular psychological or behavioral construct. These batteries, which have often been validated in prior research and used broadly in identical or nearly identical ways by various researchers on aging, often entail administering a dozen or more questions to measure a single construct. As is also true of most questionnaire-based research, other, nonbattery items are often in the questionnaires, routinely formatted in ad hoc ways, varying considerably from study to study and even within a single study. For example, Pruchno and Meeks' (2004) article in *Psychology and Aging* employed the Philadelphia Geriatric Center Positive and Negative Affect Scales developed by Lawton, Kleban, Dean, Rajagopal, and Parmelee (1992) as well as a mix of other questions offering rating scales with two, three, four, and five points.

The notion that social and psychological constructs can be measured

precisely only by using batteries of many questionnaire items has a solid conceptual and theoretical justification. Any research participant's report of a past behavior, mood state, action tendency, hope, attitude, or goal will, of necessity, contain some random measurement error, both because of ambiguity in the memories and other internal psychological cues consulted when making the rating, and because of ambiguity in the meanings of the words used in the question and the words in the offered answer choices. As documented by the Spearman-Brown prophecy formula, the greater the number of questions asked to tap a construct, the more effectively the random measurement error in each item is cancelled out by that in the others, yielding a precise assessment of the variance shared across the items. This is why the measurement of a personality attribute, an attitude, and any other such construct is routinely accomplished by asking respondents to answer a remarkably large set of questions tapping the same thing. This practice is frustrating to some researchers, because participants can only answer a limited number of questions within the time frame of a study's data collection budget, so the more questions that tap a single construct, the fewer constructs can be gauged in a single study. The enhanced precision of assessment has routinely been preferred by researchers at the expense of breadth of construct sets and at the expense of participant fatigue in answering what may appear to be the same question over and over.

However, the above logic ignores an important fact of assessment: systematic measurement error. It has long been recognized that any measuring instrument may be biased, so error in its assessments may not all be random. And if the same bias is present in a set of questions all measuring the same construct, combining responses will not yield a canceling out of the error. Indeed, combining responses will cause the shared error to represent an increasingly prominent proportion of the variance of the final index, as the number of items combined increases and the amount of random error decreases. Thus, averaging or adding together responses to large sets of items to create indices is not the solution to all measurement problems.

Ironically, much of the shared bias in questions used to build indices is created unwittingly by the researchers themselves who seek to minimize measurement error. Even more strikingly, there appears to be a remarkably simple and practical solution to these problems that will make researchers and participants happier with the process and outcomes of their efforts. By avoiding the use of question formats that create random and systematic measurement error, researchers may be able to replace long batteries with sets of just two or three items that are well written, clear in meaning, and easy to answer, yielding psychometrics comparable to or better than those of long batteries while allowing for the measurement of a much broader array of constructs in a single questionnaire.

OPTIMIZING QUESTIONNAIRE DESIGN

Undoing the damage done in building large batteries must begin with an understanding of the principles of optimal questionnaire design. The structure, wording, and order of questions for questionnaires has traditionally been viewed as "an art, not a science," in the words of Princeton University psychologist Hadley Cantril (1951, p. vii) over five decades ago. And in his book *The Art of Asking Questions,* published the same year, Stanley Payne (1951) cautioned that "the reader will be disappointed if he expects to find here a set of definite rules or explicit directions. The art of asking questions is not likely ever to be reduced to some easy formulas" (p. xi). Thirty years later, Sudman and Bradburn (1982) agreed, saying that "no 'codified' rules for question asking exist" (p. 2). Sampling and data analysis are indeed guided by such rules that are backed by elaborate theoretical rationales. But questionnaire design has been thought of as best guided by intuition about how to script a naturally flowing conversation between a researcher and a respondent, even if that conversation is sometimes mediated by an interviewer.

Experienced questionnaire designers have followed some conventions over the years, but those conventions varied enough from individual to individual and from discipline to discipline to suggest that there are few universally accepted principles. If a questioning approach seems to work smoothly when respondents answer a questionnaire, then many researchers presume it will probably yield sufficiently useful data.

In recent years, however, it has become clear that this is an antiquated view that does not reflect the accumulation of knowledge throughout the social sciences about effective question asking. To be sure, intuition is a useful guide for designing questions, and a good questionnaire yields conversations that feel natural and comfortable to respondents. However, intuition can sometimes lead us astray, so it is useful to refine intuition with scientific evaluation. Fortunately, a large body of relevant scientific studies has now accumulated, and when taken together, their findings clearly suggest formal rules about how best to design questions to maximize the reliability and validity of measurements made by individual questions. However, this work has been scattered across the publication outlets of numerous disciplines, and this literature has not yet been comprehensively and integratively reviewed in a central place. Doing so is the goal of a forthcoming book, *The Handbook of Questionnaire Design* (Krosnick and Fabrigar, forthcoming), and we summarize some of the book's conclusions here, with a focus on optimal design of rating scales for efficient and effective brief assessments of social and psychological constructs in research on aging.

Number of Scale Points

When designing a rating scale, one must begin by specifying the number of points on the scale. As is true in Pruchno and Meeks' (2004) article in *Psychology and Aging* and in a great many social and psychological studies, rating scales vary considerably in their length, ranging from dichotomous items up to scales of 101 points (see, e.g., Robinson, Shaver, and Wrightsman, 1999). A large number of studies have compared the reliability and validity of scales of varying lengths (for a review, see Krosnick and Fabrigar, forthcoming). For bipolar scales, which have a neutral or status quo point in the middle (e.g., running from positive to negative), reliability and validity are highest for about 7 points (e.g., Matell and Jacoby, 1971). In contrast, the reliability and validity of unipolar scales, with a zero point at one end (e.g., running from no importance to very high importance), seem to be optimized at somewhat shorter lengths, approximately 5 points long (e.g., Wikman and Warneryd, 1990).

Presenting a 7-point bipolar rating scale is easy to do visually but is more challenging when an interviewer must read the seven choices aloud to a respondent, who must hold them in working memory before answering. Fortunately, research by Krosnick and Berent (1993) shows that such 7-point scales can be presented in easier ways without compromising data quality. Specifically, such scales can be presented in sequences of two questions that ask first whether the respondent is on one side of the midpoint or the other or at the midpoint (e.g., "Do you like bananas, dislike them, or neither like nor dislike them?"). Then, an appropriately worded follow-up question can ascertain how far from the midpoint the respondents are who settle on one side or the other (e.g., "Do you like bananas a lot or just a little?"). This branching approach takes less time to administer than offering the single 7-point scale all at once, and measurement reliability and validity are higher with the branching approach as well.

Scale Point Labeling

A number of studies show that data quality is better when all points on a rating scale are labeled with words than when only some are labeled thus and the others are labeled with numbers or are unlabeled (e.g., Krosnick and Berent, 1993). Furthermore, respondents are more satisfied when more rating scale points are verbally labeled (e.g., Dickinson and Zellinger, 1980). When selecting labels, researchers should strive to select ones that have meanings that divide up the continuum into approximately equal units (e.g., Klockars and Yamagishi, 1988). For example, "very good, good, and poor" is a combination that should be avoided, because the terms do not divide the continuum equally: the meaning of "good" is much closer to the

meaning of "very good" than it is to the meaning of "poor" (Myers and Warner, 1968).

A very common set of rating scale labels used in questionnaires these days was initially developed by Rensis Likert (1932) and assesses the extent of agreement with an assertion: strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, strongly disagree. Yet a great deal of research shows that this response choice set is problematic because of acquiescence response bias (see, e.g., Couch and Keniston, 1960; Jackson, 1967; Schuman and Presser, 1981). Some people are inclined to agree with any assertion, regardless of its content, and content-free agreement is more common among people with limited cognitive skills, for more difficult items, and for items later in a questionnaire, when respondents are presumably more fatigued (see Krosnick, 1991). A number of studies now demonstrate how acquiescence can distort the results of substantive investigations (e.g., Jackman, 1973; Winkler, Kanouse, and Ware, 1982), and in a particularly powerful example, acquiescence undermined the scientific value of *The Authoritarian Personality*'s extensive investigation of fascism and anti-Semitism (Adorno, Frankel-Brunswick, Levinson, and Sanford, 1950).

It might seem that the damage done by acquiescence can be minimized by measuring a construct with a large set of items, half of them making assertions opposite to the other half (called "item reversals"). This approach is designed to place acquiescing respondents in the middle of the final measurement dimension but will do so only if the assertions made in the reversals are as extreme as the statements in the original items. To ensure this balance entails extensive pretesting and is therefore cumbersome to implement. Furthermore, it is difficult to write large sets of item reversals without using the word "not" or other such negations, and evaluating assertions that include negations is cognitively burdensome and error laden for respondents, thus increasing both measurement error and respondent fatigue (e.g., Eifermann, 1961; Wason, 1961). Finally, acquiescers presumably end up at the midpoint of the resulting measurement dimension, which is probably not where most belong on substantive grounds. That is, if these individuals were induced not to acquiesce and instead answered the items thoughtfully, their final index scores would presumably be more valid than placing them at the midpoint.

Most importantly, answering an agree/disagree question always involves answering a comparable rating question in one's mind first. For example, respondents asked to report their extent of agreement or disagreement with the assertion "I am not a friendly person" must first decide how friendly they are (perhaps concluding "very friendly") and then translate that conclusion into the appropriate selection in order to formulate their answer ("disagree" to the original item). It would be simpler and more direct to ask respondents how friendly they are on a rating scale ranging

from "extremely friendly" to "not friendly at all." These would be called "construct-specific response options," because they explicitly state levels of the construct being measured (i.e., friendliness). In fact, every agree/disagree rating scale question implicitly requires the respondent to make a mental rating of an object along a continuous dimension representing the construct of interest, so asking about that dimension is simpler, more direct, and less burdensome. Not surprisingly, then, the reliability and validity of rating scales involving construct-specific response options are higher than those of agree/disagree rating scales (e.g., Ebel, 1982; Mirowsky and Ross, 1991; Ruch and DeGraff, 1926; Wesman, 1946). Consequently, it is best to avoid long batteries of questions in the latter format and instead to ask questions using rating scales with construct-specific response options.

Nondifferentiation

The danger of mounting systematic measurement error applies not just to agree/disagree scales but also to any long battery of questions employing the same response scale. For example, the personality construct "need to evaluate" is measured by offering respondents a series of assertions (e.g., "I form opinions about everything" and "I pay a lot of attention to whether things are good or bad") and asking them to indicate the extent to which each one describes them ("extremely characteristic," "somewhat characteristic," "uncertain," "somewhat uncharacteristic," and "extremely uncharacteristic"; Jarvis and Petty, 1996). When answering many questions using the same rating scale, a substantial number of respondents provide identical or nearly identical ratings across questions as a result of survey satisficing (Krosnick, 1991). Although these respondents could devote careful thought to the response task, retrieve relevant information from memory, and report differentiated judgments in response to the various questions, they choose to shortcut this process because they lack the cognitive skills and/or the motivation required. Instead they choose what appears to be a reasonable point on the rating scale and select that point over and over (i.e., constituting "nondifferentiation"), rather than interpreting each question carefully and answering optimally (see Krosnick, 1991; Krosnick and Alwin, 1988). Because different satisficers select different points on the rating scale, they end up at different places on the final measurement continuum built with a battery of items, and this constitutes systematic measurement error as well. For this reason, it is preferable that multiple items measuring a single construct use different sets of rating scale labels.

Order Effects

A final consideration relevant to optimizing rating scale design is the fact that people's answers to rating scale questions are sometimes influenced by the order in which the response alternatives are offered. After reading the stem of most rating scale questions, respondents are likely to begin to formulate a judgment with which to answer the question. For example, the question, "How friendly are you?" would induce respondents to generate an assessment of their level of friendliness before looking at the offered response options. As satisficing respondents read or listen to the answer choices presented, they are likely to settle for the first response option they encounter that is within their "latitude of acceptance." According to Sherif and colleagues (Sherif and Sherif, 1967; Sherif, Sherif, and Nebergall, 1965), people's judgments can be located at single points on continua, but around those points are regions on the continua that people also find acceptable representations of their beliefs. The first such acceptable response option a satisficing respondent hears is the one likely to be selected, thus inducing primacy effects in ratings, which have been observed in many studies (e.g., Belson, 1966; Carp, 1974; Chan, 1991; Mathews, 1929). To prevent this phenomenon from undetectably biasing ratings in a single direction, it is best to rotate the order of response choices across respondents and to statistically control for that rotation when analyzing the data.

However, this recommendation must be modified in light of conversational conventions about word order. Linguists Cooper and Ross (1975) outlined a series of rules for ordering words in sentences, one of which says that in pairs of positive and negative words, it is conventional to say the positive or affirmative before the negative (e.g., "plus or minus," "like or dislike," "for or against," "support or oppose"). Similarly, Guilford (1954) asserted that it is most natural and sensible to present evaluative response options on rating scales in order from positive to negative (e.g., "like a great deal" to "dislike a great deal"). Holbrook, Krosnick, Carson, and Mitchell (2000) showed that measurement validity is greater when the order of answer choices conforms to this convention.

CONCLUSION

If researchers follow all of the above guidelines in designing rating scale questions for brief assessments, reliability and validity can be maximized, systematic measurement error can be minimized, and thus the number of questions needed to measure a single construct can be reduced.

We hope that this review of research on questionnaire design is encouraging to the many scholars of aging who employ questionnaires in their

work. Effective questionnaire design can be accomplished efficiently and practically in ways that minimize participant burden and maximize the breadth of constructs assessed and the accuracy of those assessments. By employing the principles of good measurement described here, researchers studying aging can move the field ahead even more successfully than they have to date.

REFERENCES

Adorno, T.W., Frankel-Brunswick, E., Levinson, D.J., and Sanford, R.N. (1950). *The authoritarian personality*. New York: Harper.

Belson, W.A. (1966). The effects of reversing the presentation order of verbal rating scales. *Journal of Advertising Research, 6*, 30-37.

Cantril, H. (1951) *Public opinion 1935-1946*. Princeton, NJ: Princeton University Press.

Carp, F.M. (1974) Position effects on interview responses. *Journal of Gerontology, 29*(5), 581-587.

Chan, J.C. (1991). Response order effects in Likert-type scales. *Educational and Psychological Measurement, 51*, 531-541.

Cooper, W.E., Ross, J.R. (1975). World order. In R.E. Grossman, L.J. San, and T.J. Vance (Eds.), *Papers from the parasession on functionalism* (pp. 63-111). Chicago: Chicago Linguistic Society.

Couch, A., and Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal Social Psychology, 60*, 151-174.

Dickinson, T.L., and Zellinger, P.M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology, 65*, 147-154.

Ebel, R.L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19*, 267-278.

Eifermann, R.R. (1961). Negation: A linguistic variable. *Acta Psychologica, 18*, 258-273.

Guilford, J.P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Holbrook, A.L., Krosnick, J.A., Carson, R.T., and Mitchell, R.C. (2000). Violating conversational conventions disrupts cognitive processing of attitude questions. *Journal of Experimental Social Psychology, 36*, 465-494.

Jackman, M.R. (1973). Education and prejudice or education and response-set? *American Sociological Review, 38*(June), 327-339.

Jackson, D.N. (1967). Acquiescence response styles: Problems of identification and control. In I.A. Berg (Ed.), *Response set in personality measurement*. Chicago: Aldine.

Jarvis, W.B.G., and Petty, R.E. (1996). The need to evaluate. *Journal of Personality and Social Psychology, 70*, 172-194.

Klockars, A.J., and Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*(2), 85-96.

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

Krosnick, J.A., and Alwin, D.F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52*, 526-538.

Krosnick, J.A., and Berent, M.K. (1993) Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science, 37*(3), 941-964.

Krosnick, J.A., and Fabrigar, L.R. (Forthcoming). *Designing great questionnaires: Insights from psychology*. New York: Oxford University Press.

Lawton, M.P., Kleban, M.H., Dean, J., Rajagopal, D., and Parmelee, P.A. (1992). The factorial generality of brief positive and negative affect measures. *Journal of Gerontology: Psychological Sciences, 47*(4), P228-P237.

Likert, R. (1932). *A technique for the measurement of attitudes*. New York: McGraw-Hill.

Matell, M.S., and Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Mathews, C.O. (1929). The effect of the order of printed response words on an interest questionnaire. *Journal of Educational Psychology, 20*, 128-134.

Mirowsky, J., and Ross, C.E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2x2 index. *Social Psychology Quarterly, 55*, 217-235.

Myers, J.H., and Warner, W.G. (1968). Semantic properties of selected evaluation adjectives. *Journal of Marketing Research, 5*, 409-412.

Payne, S. (1951). *The art of asking questions*. Princeton, NJ: Princeton University Press.

Pruchno, R.A., and Meeks, S. (2004). Health-related stress, affect, and depressive symptoms experienced by caregiving mothers of adults with a developmental disability. *Psychology and Aging, 19*(3), 394-401.

Robinson, J.P., Shaver, P.R., and Wrightsman, L.S. (1999). Measures of political attitudes. In *Measures of social psychological attitudes* (vol. 2). New York: Academic Press.

Ruch, G.M., and DeGraff, M.H. (1926). Correction for chance and "guess" versus "do not guess" instructions in multiple-response tests. *Journal of Educational Psychology, 17*, 368-375.

Schuman, H., and Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Sherif, C.W., Sherif, M., and Nebergall, R.E. (1965). *Attitude and attitude change: The social judgment-involvement approach*. Philadelphia: W.B. Saunders.

Sherif, M., and Sherif, C.W. (1967). Attitudes as the individual's own categories: The social-judgment approach to attitude and attitude change. In C.W. Sherif and M. Sherif (Eds.), *Attitude, ego-involvement and change* (pp. 105-139). New York: Wiley.

Sudman, S., and Bradburn, N. (1982) *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Wason, P.C. (1961). Responses to affirmative and negative binary statements. *British Journal of Psychology, 52*, 133-142.

Wesman, A.G. (1946). The usefulness of correctly spelled words in a spelling test. *Journal of Educational Psychology, 37*, 242-246.

Wikman, A., and Warneryd, B. (1990). Measurement errors in survey questions: Explaining response variability. *Social Indicators Research, 22*, 199-212.

Winkler, J.D., Kanouse, D.E., and Ware, J.E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67*, 555-561.