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

Jon A. Kronick
Stanford University

Allyson L. Holbrook
University of Illinois at Chicago

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 questions and the 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 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 schedule, 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 important 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, then, 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 num-
ber 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 di-
agnostic items up 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 Krom-
nick and Fahrenhrig, 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., Mastell and
Jacob, 1971). In contrast, the reliability and validity of unipolar scales,
with a zero point at one end (e.g., running from no importance to a very
high importance), seem to be optimized at somewhat shorter lengths, approxi-
mately 5 points long (e.g., Wickman and Wamery, 1950).

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 Kromnick 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 ques-
tions 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 offer-
ing 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., Kromnick
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, 1983). 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

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., Cougch 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 presum-
bably more fatigued (see Kromnick, 1991). A number of studies now demon-
strate how acquiescence can distort the results of substantive investigations
(e.g., Jackman, 1973; Winkler, Eronou, and Ware, 1982), and in a par-
ricularly powerful example, acquiescence undermined the scientific value
of The Authoritarian Personality's extensive investigation of fascism and
anti-Semitism (Adorno, Frankel-Brunswik, 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
either the opposite to the other half (called "item reversals"). This ap-
proach 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 assertions in the original items. To
ensure this balance entails extensive prestating 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 presum-
bably 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 but instead answered the items
thoughtfully, their final index scores would presumably be more valid than
placing them at the midpoint.

Most importantly, answering an agenda/disagree question always in-
volves answering a comparable rating question in one’s mind first. For
example, respondents asked to report their extent of agreement or disagree-
ment 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 is 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., 1992 Agree/disagree and rating scales; e.g., 1982 Mivosevic et al., 1982). 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 person's 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 vote carefully 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"). This results in increasing each question carefully and answering optimally (see Krosnick, 1991; Krosnick and Abelson, 1988). Because different scales 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.

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 research.
REFERENCES


