Journal of Official Statistics, Vol. 23, No. 2, 2007, pp. 163–184

Measuring Disability in Surveys: Consistency Over Time and Across Respondents

Sunghee Lee[1]*, Nancy A. Mathiowetz*[2]*, and Roger Tourangeau*[3]

We conducted an experiment that compared different versions of a set of disability questions (including the questions included on the Census 2000 Long Form). Disability items are prone to a number of methodological problems, including inconsistency over time and self-proxy differences. Our experimental versions were designed to reduce these problems. The different versions constituted attempts to simplify the questions in various ways; in addition, some of the questions used a five-point response scale rather than the yes-no format commonly used in survey items on disability. We also varied whether the data were collected from the sample persons themselves or from a proxy from the same household. We administered the questions in a telephone interview to a national sample of households that included at least two adults 40 years old or older. In each cooperating household, we selected two adults in this age range and interviewed one of them about themselves and the other sample adult; we attempted to interview the sample households (again by telephone) a second time two weeks after the initial interview. We found that the wording of the disability questions had a major effect on the apparent prevalence of disability in this population, but, despite our efforts to simplify it, the wording of the questions had little effect on the *consistency* of responses across interviews or on *self-proxy differences*. Answers were more consistent across interviews when the same person answered the questions both times – whether that person was a self-respondent or a proxy. Only about two-thirds of those classified as having a disability in the first interview were classified as having a disability in the second. Disability is a complex concept and different respondents may have different views about whether a given person has a disability. As a result, changes in respondents may lead to changes in the disability classification of the target person. On the other hand, changing the wording to simplify the judgment seemed to have only modest effects on consistency across interviews or consistency across respondents.

Key words: Measurement of disability; self-proxy differences; decomposition; dichotomous items versus scales.

[1] UCLA Center for Health Policy Research and Department of Biostatistics, 10960 Wilshire Boulevard, Suite 1550, Los Angeles, CA 90024, U.S.A. Email: slee9@ucla.edu
[2] University of Wisconsin – Milwaukee, Department of Sociology, P O Box 413, Bolton 750, Milwaukee, WI 53201-0413, U.S.A. Email: nancym2@uwm.edu
[3] University of Maryland, Joint Program in Survey Methodology and Institute for Social Research, University of Michigan, 1218 LeFrak Hall, College Park, MD 20742, U.S.A. Email: rtourang@survey.umd.edu
Acknowledgments: The research reported here was supported by the Michigan Centers for Excellence in Health Statistics, which is funded under a cooperative agreement from the National Center for Health Statistics, Centers for Disease Control and Prevention (UR6/CCU517481-04). We gratefully acknowledge CDC's support. The contents of the article are the responsibility of the authors and do not represent the views of CDC. We also thank Paul Guerino and Elisha Smith for their work on the analysis of the data; Darby Miller Steiger for overseeing the data collection and contributing to the design of the study; and Molly McNeeley and Liberty Greene for their help in planning the study. An earlier version of this article was presented at the Annual Conference of the American Association for Public Opinion Research, St. Pete's Beach, FL, May 16–19, 2002.

1. Introduction

Many surveys include items designed to measure disability, sometimes as the focus of the survey and sometimes simply as a key variable. In the United States, the latter category includes the Census 2000 Long Form, the American Community Survey, and the National Health Interview Survey. In addition, Congress has mandated the inclusion of questions concerning disability on the National Crime Victimization Survey and an Executive Order calls for the addition of items on disability to the Current Population Survey, the source of the monthly employment statistics in the United States. Disability is, thus, on its way to becoming a standard "demographic" item, like race or gender, routinely included on national surveys. This development is consistent with the importance of disability as a topic in its own right and with the relation of disability to a number of entitlement programs (e.g., Social Security Disability Insurance and Supplemental Security Income) and other policy issues (e.g., the effect of the Americans with Disabilities Act of 1990).

1.1. Measuring Disability in Surveys

Surveys generally assess disability using a short series of questions that focus on limitations in one or more daily activities. For example, the Census 2000 Long Form included the following items:

16. Does this person have any of the following long-lasting conditions:
    a. Blindness, deafness, or a severe vision or hearing impairment?
    b. A condition that substantially limits one or more basic physical activities such as walking, climbing stairs, reaching, lifting, or carrying?

17. Because of a physical, mental, or emotional condition lasting six months or more, does this person have any difficulty in doing any of the following activities:
    a. Learning, remembering, or concentrating?
    b. Dressing, bathing, or getting around inside the home?
    c. Going outside the home alone to shop or visit a doctor's office?
    d. Working at a job or business?

Each of these questions asked for a yes or no response.
Some of the disability questions on the Long Form cover several related possibilities. For example, question 16a asks about both hearing and vision problems and 17b covers three different activities of daily living. Further, all of the items in question 17 ask respondents to judge not only whether the person has difficulties but also whether these difficulties can be traced to a "physical, mental, or emotional condition" that has lasted (or will last) six months or more. In part, the wording of these items reflects the constraints imposed by the format of the Long Form, a mail questionnaire where space is at an extreme premium. Complicated questions like these can overwhelm working memory capacity, particularly for the older respondents who are most likely to have a disabling condition (see Salthouse 1996 and Schwarz et al. 1998, on the relation between age and working memory capacity).

We thought that answers to the disability questions might be more reliable over time and more consistent across reporters if the questions were simplified somehow – if each question covered only a single limitation, say, or if the judgment regarding the presence of a difficulty were separated from the judgments of its cause and duration. In our study, we developed items that simplified the requisite judgments in various ways and compared these questions to the items used in the Census 2000 Long Form (and in the American Community Survey). Breaking complex judgments down into simpler constituents often increases the accuracy of the judgments (see Armstrong, Denniston, and Gordon 1975, and Tversky and Koehler 1994, though Belli et al. 2000 report an exception). In the survey methods literature, this tactic is sometimes referred as "decomposition," although some researchers use this term more narrowly to mean breaking a broad category (such as shopping) into narrower subcategories (shopping at department stores, grocery stores, other stores, and on-line). Armstrong, Denniston, and Gordon (1975), who introduced the term "decomposition," used it in the broader sense of breaking complicated judgments into simpler components and we follow their lead here. Our experiment examined whether variations of the decomposition strategy increased the reliability of responses to a series of disability items and reduced the differences between the answers obtained from self- and proxy respondents. In an earlier article (Lee, Mathiowetz, and Tourangeau 2004), we examined self- versus proxy reports about a summary item on disability status ("Do you consider [Target Person] to have a disability?"); here we examine responses to more detailed questions, like those in the 2000 Census, that ask about more specific disabilities (such as difficulties seeing and hearing or getting around the house).

Our experimental versions of the detailed disability questions examined another issue. Recent conceptual models of disability stress the fact that disability is a matter of degree rather than a dichotomy (Jette and Badley 2000), yet most surveys use dichotomous yes-no items in assessing disability. The dichotomous response format of the census items may reflect earlier official definitions of the construct. For example, the World Health Organization offers the following definition:

> In the context of health experience, a disability is any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being (WHO 1980, p. 28).

We thought that this mismatch between the graded underlying construct (at least as recently articulated by researchers like Jette and Badley) and the dichotomous survey items used to measure it might create measurement problems. Within the context of attitude measurement, several studies suggest that responses to dichotomous items are rarely as reliable as responses to graded scales (Alwin and Krosnick 1991; Krosnick and Fabrigar 1997). Although disability is not an attitudinal construct, we thought the judgment process might be similar enough to that required in attitude items that a graded scale would yield more reliable responses from a single reporter (and more consistent responses across reporters) than a yes-no format. Our experiment, thus, compares two response formats – yes/no answers and answers on a five-point response scale. We compared these two formats in terms of their reliability over time and the level of self-proxy differences they produce.

1.2. Evidence of Measurement Problems in Disability Items

Relatively little methodological work has been done on the measurement of disability in survey settings (see Mathiowetz 2000 for a review), but much of the existing work suggests that there is cause for concern about the quality of disability data from surveys. For example, McNeill (1993) analyzed disability data from the 1990 Census and the Content Reinterview Survey (CRS) conducted five to nine months later. He found numerous discrepancies between the classification of persons based on their Census data and their classification based on the CRS data. Of the 298 persons classified as having difficulty getting around outside based on census responses, only 146 (49.0 percent) were classified as having mobility difficulties based on CRS. The pattern was even more striking for questions on difficulty in taking care of personal needs; of the 415 persons classified as having difficulty in the Census, only 69 (16.6 percent) were reported to have a similar difficulty in the CRS. As Mathiowetz (2000, pp. 44–50) pointed out in her discussion of these results, the 1990 Census procedures differed from those used in the CRS in several ways, including the context of the disability questions, the use of self- and proxy responses, and the mode of data collection. All of these methodological differences as well as real changes in disability status during the intervening time period could have contributed to the differences between the Census and CRS classifications. Still, McNeill (1998) found similar discrepancies in disability classifications in two successive rounds of the Survey of Income and Program Participation, even though the same items and the same respondents were used both times; these findings suggest that, along with real change, sheer unreliability may be a big contributor to the inconsistency across interviews.

All three of the factors that Mathiowetz (2000) cited as potential contributors to the apparent inconsistency of disability classifications over time – question context, self-proxy differences, and the mode of data collection – do in fact seem to affect reports about disability. Todorov (2000) analyzed data from the 1994 and 1995 National Health Interview Survey (NHIS), which demonstrated a context effect on disability reports. The NHIS for those years included several disability items, one of which asked whether the sample person had "serious difficulty seeing." The proportion of persons reported to have serious vision problems increased by about 10 percent (from 2.25 percent to 2.50 percent) when the disability items followed a list of 24 medical conditions involving hearing, vision, or other sensory problems compared to when they followed one of five other condition lists involving bodily systems unrelated to sight, such as the circulatory system. Although the effect of the prior items was not large in absolute terms (and may have little practical importance for most policy purposes), it was statistically significant and apparent in both years (Todorov 2000). In addition, the condition lists came considerably earlier in the questionnaire than the item on difficulty seeing, which would probably serve to reduce their effect on answers to the later disability item.

Two studies also showed differences in disability rates depending on whether the data were collected from the person himself or herself or from a proxy respondent (Mathiowetz and Lair 1994; Rodgers and Miller 1997). Both of these studies focused on items assessing difficulties in carrying out activities in daily living (ADL). Mathiowetz and Lair (1994) examined responses to ADL items administered as part of the 1987 National Medical Expenditure Survey. This was a longitudinal household survey with five rounds of data

collection. The ADL items were administered toward the end of the Round 1 and Round 4 interviews and covered difficulties in bathing, dressing, using the toilet, feeding, and getting in and out of bed. Self-response was associated with lower rates of reported ADL difficulties in both rounds. This was not an experimental design, but the difference between self- and proxy reports remained significant after Mathiowetz and Lair introduced extensive controls to adjust for any actual differences in functional status. Rodgers and Miller (1997) also examined the differences in reported level of ADL difficulties among self- and proxy respondents in the Asset and Health Dynamics among the Oldest Old Survey, which focuses on the population of 70 years old or older. Self-respondents were found to report fewer ADL difficulties than proxies. Again, the self-proxy difference remained significant in a model that included a number of control variables designed to adjust for differences in actual health status.

Finally, McHorney, Kosinski, and Ware (1994) conducted an experimental comparison of telephone and self-administration of the SF-36 (a 36-item health inventory) and found that data collection by telephone led to lower reports of chronic conditions than self-administration. All of these results suggest that disability reports are subject to considerable variation reflecting the conditions of the survey measurement.

Our experiment attempted to find disability items that were less prone to such variation across survey conditions. We compared several different versions of the disability questions. Our major hypothesis was that breaking down complex judgments into their simpler constituents would improve the consistency of the answers over time and across reporters. Some versions of the questions attempted to simplify the judgments – by asking separately about vision and hearing or by asking separate questions about the presence, source, and duration of a difficulty – whereas other versions asked for multiple judgments at once. Our second hypothesis concerned the format of the responses. We compared yes-no items with items that asked for graded judgments (on a five-point scale). Again, we thought that the five-point scales would improve consistency over time and across reporters because such scales fit the graded character of disability better than the standard yes-no format.

As in our earlier article (Lee, Mathiowetz, and Tourangeau 2004), we compared self-reports with proxy responses; self-proxy differences were one of our major outcome measures. In addition, we thought that self-respondents would be more consistent over time than proxies. A key feature of the design is that the data were collected in two interviews, allowing us to assess the reliability of responses over time. Thus, we can compare the different versions of the questions in terms of their reliability as well as of their susceptibility to differences across reporters.

2. Method

2.1. Overview

The Gallup Organization conducted the initial interviews with members of 1,002 sample households and reinterviewed 800 of them about two weeks later. Both interviews were conducted by telephone. Each interview gathered information about two household members who were at least 40 years old. The sample households were identified by

screening a larger telephone sample selected via random-digit dialing (RDD) for the presence of two or more members who met our age cutoff. Each questionnaire included either the disability items that appeared in the Census 2000 Long Form or one of the experimental versions we developed. Any given household received the same version of the questions in both interviews. We randomly determined whether the second interview was conducted with the same respondent as the initial interview or with the other person selected from the same household (see the Appendix for a schematic summary of the overall study design). There were separate experiments involving the questions designed to measure sensory disabilities and those designed to assess nonsensory disabilities.

2.2. Sample Design and Data Collection

The target population was households in the U.S. with two or more members who were 40 years old or older. The purpose of restricting the survey to this end of the age range was to increase the proportion of persons classified as having a disability. The purpose of restricting it to households with two age-eligible members was to allow us to compare self- and proxy reports within the same households. Computer assisted telephone interview with list-assisted RDD was used as a cost-effective means of representing this population. Because we used listed-assisted RDD to select the sample, our sample is restricted to the portion of that population accessible by landline telephones with a number in a bank of 100 consecutive possible numbers with at least one residential listing.

The sample households were identified through screening questions administered to a national sample. The initial interviews were carried out from August 21 to November 13, 2000, and the second interviews between September 5 and November 27. To the extent possible, the second interview was to be carried out two weeks after the first. We used this short time frame to minimize real changes in disability status between the two interviews. From each household, one of the age-eligible members was selected to complete the first interview, answering questions both about himself or herself and about the other sample person selected within that household. The last birthday method was used to select the respondent for the initial interview. In households with more than two age-eligible members, we randomly selected the second sample person. The second person we selected in each household was generally the spouse of the first (87.7 percent of the time); most of the remainder were other relatives (most often parents or children) of the first person (6.7 percent). As part of the experimental design, we randomly determined whether the respondent for the first interview or the other sample person from that household would complete the second interview two weeks later.

Gallup fielded 8,012 numbers for the initial interview. They were randomly generated from a random sample of banks of 100 consecutive potential numbers (e.g., the numbers 301 314-7900 – 301 314-7999); sampling was restricted to 100-banks that included at least one residential telephone number. This is a standard telephone sampling method (see Brick et al. 1995; Casady and Lepkowski 1993) that yields a relatively high proportion of working residential numbers (WRNs). Of the 8,012 potential telephone numbers fielded, we estimate that 58.8 percent, or a total of 4,711, were WRNs. About a third (33.2 percent) of the households that completed the screening questions had two or more eligible members (implying a total of 1,564) and 1,002 of them completed the initial interview.

According to the American Association for Public Opinion Research formula RR3, this represents a response rate of 64.1 percent. (This calculation assumes that the observed rate of WRNs and eligible households applies to the full sample of 8,012 numbers fielded.) Of the 1,002 responding households in the initial wave, 800 completed the second interview for a response rate of 79.8 percent. The overall response rate across the two interviews was 51.2 percent (that is, 64.1 percent × 79.8 percent).

2.3. Experimental Design

The experiment varied three factors. The first was whether we interviewed the same person in both interviews or had different respondents in the two interviews. (In fact, during the waning days of the field period, we allowed interviewers to carry out the second interview with whichever sample person they could persuade to participate. In total, 34 of the reinterviews were not done with the assigned respondent for the second wave, but with the other sample person from the household. Except where noted, the analyses are based on all 800 sample households that completed both interviews, with cases classified by their actual, not their intended, treatment. The results do not change appreciably if we drop the data from the 34 households in which the wrong person was interviewed in the second wave.)

The second experimental variable was the wording of the questions on sensory disability. We compared four different versions of these questions:

A. Do you have any of the following long-lasting conditions . . . blindness, deafness, or a severe vision or hearing impairment?
B1. (When you wear your glasses or contacts,) Do you often have difficulty seeing road signs?
B2. (When you wear your glasses or contacts,) Do you often have difficulty seeing words and letters the size of ordinary newspaper print?
B3. (When you wear your hearing aid,) Do you often have difficulty hearing what is said in normal conversation?
C1. Do you have any of the following long-lasting conditions . . . blindness or a severe vision impairment?
C2. Deafness or a severe hearing impairment?
D1. (When you wear your glasses or contacts,) How much difficulty do you have seeing road signs?
D2. (When you wear your glasses or contacts,) How much difficulty do you have seeing words and letters the size of ordinary newspaper print?
D3. (When you wear your hearing aid,) How much difficulty do you have hearing what is said in normal conversation?

Version A was taken from the Census 2000 Long Form. Versions B and D separated vision impairments from hearing impairments and provided a concrete standard for what constitutes an impairment (e.g., the inability to read newsprint). As it turned out, the concrete standards we chose were far too lenient, yielding rates of disability about three times higher than the other two versions of the sensory disability questions. We therefore focus on Versions A and C in the rest of this article. Version C was quite similar to the

census question and retained the yes-no format but asked separately about vision and hearing problems.

The third experimental variable was the wording of the remaining disability items. We also compared four different versions of these questions. (A fifth version of the questions was identical to Version D, but added a follow-up question asking how the person carried out the activity – without any assistance, with the assistance of another person, or with the help of special equipment or medication. We found no differences between this version and Version D and combined the two in the analyses.) We illustrate the differences across the four versions of these questions with the items on basic physical limitations:

A. Do you have a long-lasting condition that substantially limits one or more basic physical activities such as walking, climbing stairs, reaching, lifting, or carrying?
B. The next question is about basic physical activities, such as walking, climbing stairs, reaching, lifting, or carrying. To what extent does a long-lasting condition affect your ability to carry out one or more of these basic physical activities?
C1. Are you limited in your ability to perform one or more basic physical activities such as walking, climbing stairs, reaching, lifting, or carrying?
C2. Is this because of a health or physical problem?
C3. How long have you been limited in your ability to perform one or more of these basic physical activities?
C4. Do you expect to be limited in your ability to perform one or more of these basic physical activities three months from now?
D1. How much difficulty do you have performing basic physical activities such as walking, climbing stairs, reaching, lifting, or carrying?
D2. Is this because of a health or physical problem?
D3. How long have you been limited in your ability to perform one or more of these basic physical activities?
D4. Do you expect to be limited in your ability to perform one or more of these basic physical activities three months from now?

Version A was again taken from the Census 2000 Long Form. Version B closely follows the wording of Version A but used a five-point response scale rather than a yes-no format. Versions C and D both break the overall judgment required in Version A into separate judgments about the presence of a difficulty, its source, and the expected duration. (The follow-up items in Versions C and D were skipped for respondents who reported no difficulty at all or very little difficulty with the set of activities in question.) These last two versions differ in offering two response options (Version C) or five (Version D). The five options offered by both Versions B and D were "No difficulty at all," "A little difficulty," "Some difficulty," "A lot of difficulty," and "Completely unable" to perform the activity in question. Five response options seemed the most that would be practical in a telephone interview. Aside from basic physical activities, the questionnaire asked about difficulties in learning, remembering, or concentrating; dressing, bathing, or getting around inside the home; going outside the home; and working at a job or business. The four versions of the items assessing each of these nonsensory disabilities formed a two by two design, contrasting two response formats (yes-no vs a five-point scale) and the use of a single item vs multiple items for each nonsensory disability.

2.4. Questionnaires

Both the initial interview and reinterview questionnaires followed the same general organization. The initial questionnaire began with questions about the respondents' own sensory and nonsensory disabilities, using one of the experimental versions described earlier. After these items, the questionnaire included an overall question about the respondents' perception of themselves as having a disability ("Do you consider yourself to have a disability?") and other people's perception of them ("Would other people consider you to have a disability?"). Then, the questionnaire asked parallel items about the other sample person from that household, using the same versions of the disability questions for the other sample person as for the respondent. The questionnaire concluded with a series of demographic questions asking about the employment status, educational attainment, and race/ethnicity of the respondent and then of the other sample member; and finally a couple of questions asking about the respondent's mood and the weather. At the end of the first interview, the interviewers thanked the respondents and told them "We are interested in how people's views about their health changes over time and it is important for us to get the perspectives of different people." The interviewer then asked when it would be convenient to recontact the household for a second interview.

The second interview questionnaire began by asking whether the respondent or other adult for whom information had been collected in the original interview had experienced a significant change in health since the previous interview; the rest of the questionnaire (including the key disability questions) was identical to the questionnaire used in the initial interview. On average, the initial interview took about eleven minutes to complete; the reinterview, about ten minutes.

3. Results

Our analysis focused on three major outcomes. First, we examine the estimated prevalence of disability by the different question versions and by whether the data were provided by self- versus proxy reporters. These analyses determine whether the basic survey estimates would differ depending on the version of the questions and source of the report and allow us to examine the differences between self- and proxy reporters. Next, we analyze the consistency of disability classifications across interviews as a function of the experimental variables. In examining both the rates of reported disability and consistency over time, we present the results first for the sensory items and then for the nonsensory items. We also created a pooled measure of nonsensory disability that combined responses to all five sets of nonsensory items; this composite classified a person as having a disability if he or she reported having difficulty (or was reported to have difficulty) on *any* of the five sets of nonsensory questions. The analyses of the consistency data help us determine which version of the questions yields more reliable data. Finally, we briefly examine responses to the perception items (the respondents' perceptions of themselves as having a disability and the perceptions of other people) and the relationship of these responses to classification on the individual disability items.

The analyses we present are unweighted, since we are more interested in comparing the different experimental groups than in making population estimates. Still, because of the clustering of observations by household, we used PROC SURVEYLOGISTIC in SAS and

RLOGIST in SUDAAN to carry out the analyses. Both programs provide adjusted standard errors and significance tests that reflect the correlation between the two observations within each household.

3.1. Estimated Rates of Disability

3.1.1. Sensory Disability

Table 1 shows the proportion of people classified as having a sensory disability by the version of the items and by whether the report was from the sample person or a proxy respondent; the data are from the initial interviews. Two-way logit analyses revealed no effect of the version of the questions on reported rates of sensory difficulties but showed a marginally significant self-proxy difference ($\chi^2 = 3.54$, df $= 1, p < .07$) and a marginally significant interaction between the version of the items and the respondent variable ($\chi^2 = 3.54$, df $= 1$, $p < .07$). Overall, proxies were marginally more likely to report sensory difficulties (11.4 percent vs 8.2 percent for self-reporters) but this self-proxy difference was apparent only with the version that asked about seeing and hearing in a single question, the version used in Census 2000. Among those who got that version of the disability question, proxies were almost twice as likely as self-respondents to report a sensory disability (13.3 percent vs 6.8 percent).

3.1.2. Nonsensory Disability

Table 2 shows the rates of reported difficulties in response to the various nonsensory items. It is obvious how to classify people who got the yes-no versions of the items, but less clear how to categorize those who reported their difficulties on a five-point scale. It turned out that no cutoff point on the scale yielded results close to those from the yes-no items. If we counted only those said to have "a lot of difficulty" with the activity in question or to be "completely unable" to perform it as having a disability, the proportions classified as having a disability were significantly lower than the proportions based on the yes-no questions. (The results tabulated in Table 2 follow this criterion.) If we also included those said to have "some difficulty," the proportions were significantly higher than those based on the yes-no responses. To simplify the discussion, we report results only for the classifications based on the stricter criterion. The final row in the table, labeled "Any limitation," is a summary measure in which a person was classified as having a disability if

Table 1. Rates of Reported Sensory Difficulties (and Sample Sizes), by Version of the Questions and Self/Proxy Respondent

	Single item (Version A)	Hearing and vision separated (Version C)	A vs C	Self vs proxy
Self	6.8 (266)	9.6 (261)	$\chi(1)^2 = 0.00$ (ns)	$\chi(1)^2 = 3.54$ (ns)
Proxy	13.3 (264)	9.6 (261)		
Combined	10.0 (530)	9.6 (522)		

Note: Data are from the initial interviews only.

Table 2. Rates of Reported Nonsensory Disabilities, by Activity, Version of the Questions, and Reporter

Activity	Single items/ Yes-No (Version A)	Single items/ 5-point scale (Version B)	Multiple items/ Yes-No (Version C)	Multiple items/ 5-point scale (Version D)	Chi-square values		
					A vs C	B vs D	Self vs proxy
Walking, climbing, reaching							
Self	20.8 (259)	9.9 (253)	11.5 (235)	5.5 (254)	5.50*	0.05	1.84
Proxy	20.0 (260)	7.5 (253)	16.7 (234)	11.1 (253)			
Total	20.4 (519)	8.7 (506)	14.1 (469)	8.3 (507)			
Learning, remembering, concentrating							
Self	7.0 (259)	2.8 (253)	3.0 (233)	0.8 (254)	7.72*	1.82	4.84*
Proxy	9.2 (260)	3.6 (253)	3.9 (234)	3.2 (254)			
Total	8.1 (519)	3.2 (506)	3.4 (467)	2.0 (508)			
Dressing, bathing, getting around inside the home							
Self	3.5 (259)	2.4 (253)	2.6 (235)	1.2 (254)	0.18	0.01	3.71 ($p < .06$)
Proxy	5.0 (260)	2.0 (253)	5.1 (235)	3.5 (254)			
Total	4.2 (519)	2.2 (506)	3.8 (470)	2.4 (508)			
Going outside the home							
Self	6.2 (260)	4.8 (252)	2.1 (235)	3.5 (254)	5.10*	0.00	7.33**
Proxy	8.9 (260)	5.1 (253)	6.0 (235)	6.7 (253)			
Total	7.5 (520)	5.0 (505)	4.0 (470)	5.1 (507)			
Working at a job or business							
Self	11.3 (256)	12.1 (207)	6.7 (209)	6.2 (210)	3.66 ($p < .06$)	4.00*	0.01
Proxy	12.9 (256)	8.2 (219)	8.9 (217)	5.6 (195)			
Total	12.1 (512)	10.1 (426)	7.8 (412)	5.9 (405)			
Any limitation							
Self	24.6 (256)	14.7 (245)	14.2 (233)	8.5 (248)	9.45**	0.18	2.14
Proxy	25.6 (258)	10.7 (243)	18.0 (233)	15.7 (249)			
Total	25.1 (514)	12.7 (488)	16.1 (466)	12.1 (497)			

Note: Data are from the initial interviews only. Numbers in parentheses represent cell sample sizes. All the χ^2 values have 1 degree of freedom.

he or she was classified as having a disability on any of the five sets of nonsensory questions.

We analyzed the rates in Table 2 by fitting fully saturated logit models that included the following three experimental variables: (1) whether the estimate was based on answers to a single question or multiple questions that separately assessed the level of difficulty, the source of the difficulty, and its duration; (2) whether the questions used a yes-no response format or a five-point scale; and (3) whether the data were obtained from self- or proxy reports. (The chi-square values in the table for the A vs C and B vs D comparisons are based on contrasts within those logit models.) Again, to simplify the discussion we focus on the "any limitation" composite that combines responses across the five sets of nonsensory disability questions.

The disability rates tended to be lower when respondents had to answer multiple questions about each set of activities than when they answered a single question covering the presence, source, and duration of each difficulty. There are significant main effects for the single vs multiple item variable on the composite measure and on two of the five sets of questions about specific activities (walking/climbing/reaching and working at a job or business). When a single question was used (collapsing across Versions A and B), 19.1 percent of the sample persons were classified as having some type of nonsensory disability versus 14.4 percent when multiple questions were used (collapsing across Versions C and D) — $\chi^2 = 5.41$, df $= 1$, $p < .05$ for the main effect of the number of items variable.

Versions C and D may have lowered the estimated rates of disability by filtering out respondents who have some difficulty with an activity but not one resulting from a chronic health condition. The multiple-item version of the questions forces respondents to explicitly take into account the source and duration of their difficulties. We recalculated the rates for the multi-item versions disregarding answers to the follow-up items on the source and duration of the difficulty. The differences between these initial answers and the final estimates presented in Table 3 represent the effect of the follow-up items on the proportion of persons classified as having a disability. Table 3 provides these estimates for the two sets of items (walking/climbing/reaching and working at a job or business) for which the single and multiple question versions differed significantly. As can be seen from the table, there is little difference between the estimates based on responses only to the initial question and those based on answers to both the initial and follow-up questions – very few respondents reported a difficulty that was *not* due to a health condition lasting six more months or more. These findings suggest that it was not multiple questions per se that led to lower estimates but rather the change in the wording of the initial question. It is possible that the somewhat longer combined question led to fuller recall of nonsensory problems (cf. Sudman and Bradburn 1982, who recommend longer questions for gathering factual information). Another possibility is that mentioning the source of the difficulty (a "long-lasting condition") in the initial question triggers better recall of any difficulties linked to that condition. A final possibility is that the phrasing used in Versions A and B ("a long-lasting condition that substantially limits/affects your ability") leads respondents to construe the relevant concept more broadly than the phrasing used in Versions C and D ("Are you limited in your ability/How much difficulty do you have. . ."). We had anticipated that the latter phrasing would elicit more reported difficulties than the former.

Table 3. Rates of Reported Nonsensory Disabilities by Activity and Version of the Questions, Taking into Account (Final) or Ignoring (Initial) Responses to the Follow-up Items

	Multiple items/Yes-No (Version C)		Multiple items/5-point scale (Version D)	
	Initial	Final	Initial	Final
Walking, climbing, reaching				
Self	12.8 (235)	11.5 (235)	5.9 (254)	5.5 (254)
Proxy	18.8 (234)	16.7 (234)	11.1 (253)	11.1 (253)
Combined	15.8 (469)	14.1 (469)	8.5 (507)	8.3 (507)
Working at a job or business				
Self	6.7 (209)	6.7 (209)	6.7 (210)	6.2 (210)
Proxy	9.4 (203)	8.9 (203)	5.6 (195)	5.6 (195)
Combined	8.0 (412)	7.8 (412)	6.2 (405)	5.9 (405)

Note: Data are from the initial interviews only. Numbers in parentheses represent cell sample sizes

As we already noted, the use of a five-point scale affected the estimated rate of disability as compared to the yes-no format. For three of the five activities as well as the pooled measure, the estimate of persons with disabilities is significantly lower with the five-point scale. The effect is substantial. The estimate of persons with any nonsensory disability is 20.8 percent among those offered the yes-no format (averaging across Versions A and C) as compared to 12.4 percent among those offered the five-point response scale (averaging across Versions B and D) — $\chi^2 = 18.8$, df $= 1$, $p < .001$. (If we also classified individuals who reported "some difficulty" with an activity as having a disability, these findings were reversed – that is, items using five-point scales yielded *higher* estimates of disability rates than the yes-no items.) The five-point scale we used clearly does not map readily onto the yes-no format that is typically used to measure disability in surveys.

Although there was no main effect of the self-proxy variable for the pooled nonsensory disability measure, three of the five sets of items on individual activities show an effect for this variable. For those three sets of items (which asked about learning, remembering, and concentrating; dressing, bathing, and getting around inside the house; and going outside alone), proxy reports yielded higher rates of disability than self-reports did (all p's $< .05$ for the reporter main effect).

From our perspective, the key findings involve the interactions between the use of single versus multiple questions and the self-proxy variable and the interactions between the response format variable and who the respondent was. The self-proxy difference is clearly larger when multiple questions are used to assess disability than when a single question is used ($\chi^2 = 7.93$, df $= 1$, $p < .001$). Figure 1 displays this interaction. There are similar significant interactions involving the item on walking, climbing, and reaching and the item on getting around outside the house (data not shown). None of the interactions between the response format and the self-proxy variable were significant. The self-proxy differences are similar for the two response formats.

In summary, reported rates of nonsensory disability tended to be higher when a single combined question was asked than when separate questions were used to assess the presence, source, and duration of a difficulty, when the questions used a yes-no format

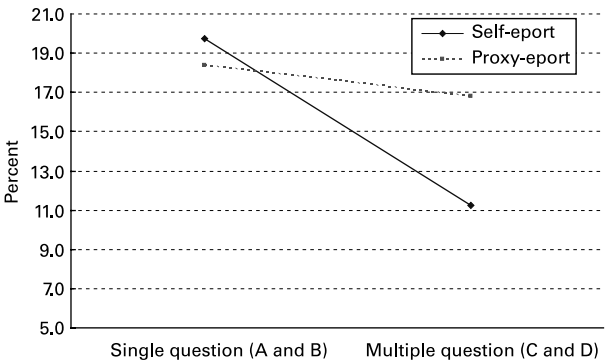


Fig. 1. Rate of Any Reported Nonsensory Disability by Reporter and Single vs Multiple Questions

than when they used a five-point scale, and when the data were obtained from proxy respondents than when they were obtained from the sample persons themselves. All three are features of the questions used in Census 2000 and in the American Community Survey.

3.2. Consistency of Classifications Across Interviews

Our next set of analyses used the data from both interviews to examine the degree of consistency in the classification of individuals across waves. We had hoped that our experimental versions of the questions would improve across-interview consistency. We focused on reports about sensory difficulties, difficulties with basic physical activities (walking, climbing stairs, and so on), and any nonsensory difficulties (pooling responses to all the items on nonsensory difficulties). The remaining items yielded very few sample persons with disabilities (see Table 2). Again, we carried out logit analyses, this time examining the proportion of sample persons classified consistently in the two interviews by the experimental variables. We examined the rates of consistency across interviews within the entire sample of 1,600 persons for whom data were collected in both interviews; in addition, we examined consistency just for those persons classified as having a disability for that particular activity in the initial interview. For example, we examined the proportion classified as having a sensory disability in the *second* interview among those classified as having a sensory disability in the first.

3.2.1. Effect of Question Format

For both the sensory and nonsensory items, we observed no consistent pattern of differences in the rates of consistency across interviews by the different versions of the disability questions. Within the entire sample, there are some scattered significant effects for the version of the questions on consistency across interviews, but no systematic pattern. In general, the same versions of the questions that produced the *highest* levels of reported disability also tended to produce the *lowest* levels of consistency across interviews. This probably reflects the fact that the rates of consistent classification are much higher among those who were not classified as having a disability in the initial

interview than among those who were (as is apparent in Table 4). Because inconsistencies were much more likely for those classified as having a disability in the initial interview, the same versions of the questions that classified more sample persons as having a disability were prone to produce inconsistent reports across interviews. When we restricted the analysis to those classified as disabled in the first interview, we found no significant differences in consistency across interviews across the different versions of the questions.

3.2.2. Effect of Reporter

By contrast, in virtually every analysis of consistency across interviews, who it was that provided the data mattered a great deal. Whether we examine the entire sample or just those sample persons classified as having a disability in the first wave, the classification was more likely to be consistent across the two interviews when the same respondent provided the data both times. Table 5 shows the proportion of people classified consistently across interviews by the combination of reporters in the two rounds (for example, data from self report in both rounds of interviewing). In all six rows of the table, the proportion classified consistently is higher when the same respondent provided the data both times than when different reporters were interviewed in the different rounds. Four of these differences are statistically significantly, according to the logit analyses (all p's $< .05$), and an additional one is marginally significant. (The two nonsignificant results involve sensory disabilities, where the pattern is the same but the sample sizes are smaller, since we only examine data from two versions of the questions.)

Another pattern to emerge from these analyses is that, when the same respondent completed both interviews, proxies were somewhat more reliable than self-reporters. Although the differences are rarely significant, they are nonetheless consistent across the various sets of items. In all six comparisons in Table 5, the rates of consistent classification are higher for the proxy-proxy combination than for the self-self (compare the first two columns of Table 5). In part, this difference may be due to the use of relatively stable information in the formulation of responses about others as opposed to more labile

Table 4. Percentage of Sample Persons Classified Consistently Across Waves, by Initial Interview Disability Status

Activity	Classification based on Round 1 interview			Total sample
	Having a disability	Not having a disability	Difference	
Seeing or hearing	64.6 (82)	96.0 (780)	$\chi(1)^2 = 68.1$	90.5 (862)
Walking, climbing, reaching	64.3 (210)	94.3 (1379)	$\chi(1)^2 = 101.4$	90.3 (1589)
Any nonsensory limitation	72.9 (255)	93.4 (1284)	$\chi(1)^2 = 75.0$	90.0 (1539)

Note: Numbers in parentheses represent cell sample sizes. Figures for seeing and hearing are based on Versions A and C of the sensory difficulty questions. The χ^2 values are from logit models including Wave 1 status, respondent combination, and version of the questions as predictors; all p's $< .001$.

Table 5. Percentage of Persons Classified Consistently Across Waves by Reporter in Each Wave, for Full Sample and Those Classified as Having a Disability in Initial Interview

Activity	Entire sample					
	Self-self	Proxy-proxy	Self-proxy	Proxy-self	Same respondent	Different respondents
Seeing and hearing	93.8 (194)	94.9 (195)	90.7 (237)	93.2 (236)	94.3 (389)	92.0 (473)
Walking, climbing, reaching	91.7 (374)	94.1 (371)	90.0 (421)	88.7 (423)	92.9 (745)	89.3 (844)
Any nonsensory limitation	92.0 (361)	95.3 (365)	89.3 (411)	87.6 (402)	93.7 (726)	88.4 (813)
Activity	Sample persons classified as having a disability in initial interview					
Seeing and hearing	62.5 (16)	69.2 (26)	55.0 (20)	70.0 (20)	66.7 (42)	62.5 (40)
Walking, climbing, reaching	66.7 (54)	83.3 (54)	56.8 (44)	59.3 (59)	75.0 (108)	58.2 (103)
Any nonsensory limitation	76.2 (63)	88.1 (67)	66.1 (56)	70.4 (71)	82.3 (130)	68.5 (127)

Note: Numbers in parentheses represent cell sample sizes. Figures for seeing and hearing are based on Versions A and C of the sensory difficulty questions.

information (that may be more sensitive to wording effects) in the formulation of responses about oneself (Schwarz and Wellens 1997).

3.2.3. Effect of Health Change

The second wave questionnaire included an item asking whether either of the sample persons had experienced a change in health since the initial interview. Seventy-eight sample persons (out of 1,600 for whom data were collected in both interviews) were reported to have undergone a change in health. When changes were reported, the consistency of disability classifications is somewhat lower than when no health change was reported, but the difference was not dramatic. More importantly, even when the cases with a health change are excluded from the analyses, our main conclusions about consistency do not change. The most important variable in determining the consistency remains whether the same reporter provided the information in both interviews.

3.2.4. Effect of Respondent Characteristics

The key factor determining the level of consistency across interviews is whether the same person answered the questions rather than which version of the questions they answered. Do particular respondent characteristics predict consistent reporting? We analyzed age (55 and under versus 56 and older), educational attainment (high school diploma or less versus some college or more), gender, and race (white versus non-white) of the reporter as possible predictors of consistency across waves. We tested the relation of these variables to consistency across interviews in logit analyses, restricting the analyses to the cases in which the same reporter responded both times. The respondent was classified as a consistent reporter if the person for whom he or she reported received the same disability classification in both interviews. Age and education were related to consistent reporting ($p < .001$). Younger respondents are more likely to provide consistent reports than older respondents (86.7 versus 63.1 percent), and more educated respondents are more likely to be consistent reporters than less educated ones (83.0 versus 65.0). (Separate models for self-self and proxy-proxy reports supported the same conclusions.) When we restricted this analysis further to sample persons classified as having at least one disability at the time of the first interview, only age remained a significant predictor of consistency across interviews, with younger respondents being more consistent reporters than older ones.

3.3. Perceptions of Disability

The questionnaire included items asking whether each sample person considered himself or herself to be disabled and whether other people considered the sample person to be disabled. Our analysis focused on the data provided by the respondents about themselves in the initial interview (see Lee, Mathiowetz, and Tourangeau 2004 for an analysis of the consistency across interviews of responses to these questions). We used logistic regression to examine the relation between their answers to the self- and other people's perception

items and their disability status on each set of activities. Respondents were significantly more likely to see themselves as having a disability if they reported difficulties seeing or hearing, working, or with basic physical activities (walking, climbing stairs, and so on) than if they did not report these difficulties. Problems with the other activities (e.g., getting around outside the house) had no significant relation to self-perception of disability. Similarly, respondents were significantly more likely to say that other people considered them disabled if they had reported difficulties in these same three domains or with learning, concentrating, or remembering.

4. Discussion

We examined three experimental variables – simplification of the questions, the response format, and the respondent. How did each of these variables affect disability reporting?

4.1. Effect of Question Simplification

The different versions of the questions often produced different estimates of disability rates. For the items on difficulties seeing and hearing, separating the questions on seeing and hearing tended to produce somewhat (though nonsignificantly) higher rates of reported difficulty (see Table 1). For the items on nonsensory difficulties, asking separately about the presence, source, and duration of the problem reduced the apparent prevalence of difficulties (see Table 2). This difference between the single and multiple question versions, which was significant for three of the five sets of nonsensory items as well as for the pooled composite, does not appear to reflect the effect of the follow-up questions asking about the source and duration of the difficulty. The initial items asking about the presence of a difficulty already displayed lower rates of reported problems than the single-item versions of these questions (see Table 3).

We thought that simplifying the questions would improve the quality of the reports but were less sure how it would affect the rates. Questions that call for several judgments at once impose a greater burden on working memory and may be more error-prone than those that ask for a single judgment. Is there any evidence about which versions of the disability questions produced better reports? The separate items on seeing and hearing seemed to be somewhat less susceptible to self-proxy differences than the version that asked about both simultaneously, but this finding (which is apparent in Table 1) was only marginally significant and the two versions of the sensory disability questions produced responses that were equally consistent across interviews. By contrast, the items that asked about the presence, source, and duration of a nonsensory difficulty in a single question showed smaller self-proxy differences than the versions that asked separate questions to elicit each component judgment (see Figure 1). The single- vs multiple-item versions of the nonsensory questions did not differ in terms of the consistency of classifications across interviews. Thus, there is not much evidence in our results that simplifying the questions had a marked or consistent effect on the quality of the answers. It could be that chronic disabling conditions and their effect on everyday activities form a coherent conceptual package for most people; uncoupling the judgments about different components of the package does not seem to

make them any easier or more reliable. It is also possible that some other method of decomposing these questions would have led to better results.

4.2. Effect of Response Format

Although disability would clearly seem to be a matter of degree, it is usually measured as a dichotomous variable in surveys. Our use of a five-point scale allowed us to compare the results from the scale with those from the yes-no format. The five-point scale does not easily map into the two-point dichotomy. Restricting the classification to respondents who have at least "a lot of difficulty" with an activity resulted in lower estimates of the prevalence of nonsensory disabilities than yes-no items. On the other hand, classifying individuals who report having "some difficulty" as having a disability produced much higher prevalence estimates than those based on yes-no items (data not shown). Whatever cutoff respondents use spontaneously when they are asked yes-no questions about disabilities apparently involves a level of difficulty that lies somewhere between those represented by "some difficulty" and "a lot of difficulty." It might have been better to collect the data in two steps, first asking whether the sample person had any difficulty with a given activity and then asking for a rating of the severity of any reported difficulty.

In any case, the items we developed that used a five-point scale produced classifications that were just as susceptible to self-proxy differences and no more consistent across interviews than the yes-no items. So long as the goal is to estimate the proportion of the population with a disability, yes-no items seem to work as well as items that allow more graded assessments. Of course, the question of how to collect (and report) disability rates is substantive not methodological. From a methodological perspective, we thought questions that fit the naïve conception of disability more closely (incorporating the notion that disability is continuum rather than a categorical state) would yield more reliable answers and more agreement between reporters. That prediction did not receive much support. (See Tourangeau et al. 2006, on the importance of the fit between the concepts assumed by the questions and those held by the respondents.)

4.3. Effect of Reporter

As has been observed in the past, proxies tend to report higher levels of disability than self-respondents. One of the major problems with reports about disability is their instability over time (e.g., McNeil 1993). In our study, only about two thirds of those classified as having a disability at the time of the initial interview were classified the same way two weeks later (see Table 4). This low level of consistency occurred despite our using the same interviewers, the same mode of data collection, and the same questionnaire in both interviews. The variable that had the clearest and most consistent effect on consistency across interviews was whether the same person answered the questions in both interviews (see Table 5). Who we asked mattered far more than which version of the questions we asked them. The classifications were also somewhat

more consistent when the data were obtained from the same proxy reporter in both interviews than from the self-reporter.

Several studies have suggested that proxy reports may be more reliable than self-reports (O'Muircheartaigh 1991; see Moore 1988 for a review). In the case of disability, proxy reporters may be more attuned to the usual state of the sample person than to temporary fluctuations in his or her condition, leading to more stable, though not necessarily more accurate, reports. A number of studies suggest that proxy reports are based on generic rather than episodic information – that is, information about the usual pattern rather than details about specific incidents (Schwarz and Wellens 1997; Blair et al. 1991). Proxies may base their answers to disability questions on relatively stable information about the target and underreport actual variations in his or her abilities. Aside from whether the reporter was a self-respondent or a proxy, we found that younger and more educated respondents were more likely to provide consistent answers than their older and less educated counterparts.

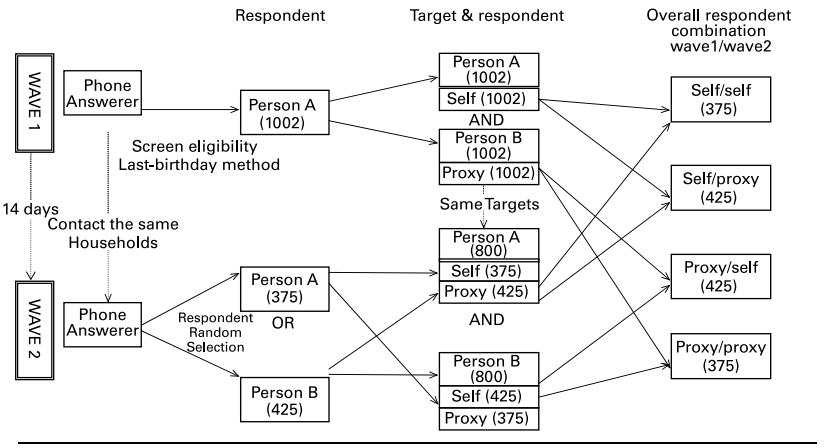
4.4. Conclusions

Our attempts to simplify the wording of the disability questions did not consistently reduce self-proxy differences or increase reliability across interviews. Clearly, there is still plenty of room for improvement in these items – two members of the same household often disagreed about whether a given sample person had a difficulty, and even for a single respondent there was considerable variation in assessments over a period as short as two weeks. The measurement of disability in surveys remains a challenge, requiring respondents to make difficult judgments. The versions of the questions that we took from Census 2000 ask for complex, dichotomous judgments and consistently yielded the highest reported rates of disability (see Tables 1 and 2).

Disability is not necessarily a stable characteristic. Perhaps it should come as no surprise that respondents disagree in making their judgments about the status of themselves or other members of their families and that they change their minds from one interview to the next. Although the Census 2000 questions (and other standard items used to measure disability) assume that disability is a stable characteristic (reflecting "long-lasting conditions"), for many respondents it may involve real fluctuations over short periods of time and self-reporters may be more aware of these changes than proxy respondents are. Thus, despite their greater reliability, proxy reports may be less valid than those obtained directly from the sample person. Unfortunately, our results examine only the consistency of the reports – over time and across reporters – and do not attempt to assess their validity.

Our study investigated items on a single topic – disability – but we believe these items have a lot in common with other *quasi-attitudinal* items (such as items asking respondents to rate their overall health, to decide who "usually" lives in their residence, or even to report their ethnic backgrounds). Such items ask respondents to reduce complex phenomena involving multiple (continuous) dimensions to a single (often dichotomous) judgment; the concepts of the respondents may or may not fit well with the concepts underlying the questions. Our experiment represents a first attempt to improve the reliability of such judgments.

Appendix. Overall Study Design



Note: The selected respondents answered questions about themselves and another household member. Numbers in parentheses represent cellsample sizes

(Source: Lee 2002)

5. References

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

Armstrong, J.S., Denniston, W.B. Jr., and Gordon, M.M. (1975). The Use of the Decomposition Principle in Making Judgments. Organizational Behavior and Human Performance, 14, 257–263.

Belli, R., Schwarz, N., Singer, E., and Talarico, J. (2000). Decomposition Can Harm the Accuracy of Behavioral Reports. Applied Cognitive Psychology, 14, 295–308.

Blair, J., Menon, G., and Bickart, B. (1991). Measurement Effects in Self vs. Proxy Responses: An Information-Processing Perspective. In Measurement Errors in Surveys, Biemer, P.P. Groves, R.M. Lyberg, L.E. Mathiowetz, N.A. and Sudman S. (eds), 145–166. New York: Wiley.

Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in List-assisted Telephone Surveys. Public Opinion Quarterly, 59, 218–235.

Casady, R.J. and Lepkowski, J.M. (1993). Stratified Telephone Survey Designs. Survey Methodology, 19, 103–113.

Jette, A.M. and Badley, E. (2000). Conceptual Issues in the Measurement of Work Disability. In Survey Measurement of Work Disability, N.A. Mathiowetz and G.S. Wunderlich (eds), 4–27. Washington, D.C.: National Academy Press.

Krosnick, J.A. and Fabrigar, L.R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), 141–164. New York: Wiley.

Lee, S. (2002). I Am Disabled On and Off! A Study of Proxy Response in a Disability Survey. Paper presented at the Joint Statistical Meetings, American Statistical Association, New York, NY.

Lee, S., Mathiowetz, N.A., and Tourangeau, R. (2004). Perceptions of Disability: The Effects of Self- and Proxy Response. Journal of Official Statistics, 20, 671–686.

Mathiowetz, N.A. (2000). Methodological Difficulties in the Measurement of Work Disability. In Survey Measurement of Work Disability, N.A. Mathiowetz and G.S. Wunderlich (eds), 28–54. Washington, D.C.: National Academy Press.

Mathiowetz, N. and Lair, T. (1994). Getting Better? Change or Error in the Measurement of Functional Limitations. Journal of Economic and Social Measurement, 20, 237–262.

McHorney, C., Kosinski, M., and Ware, J. (1994). Comparisons of the Costs and Quality of Norms for the SF-36 Health Survey Collected by Mail Versus Telephone Interview: Results from a National Survey. Medical Care, 32, 551–567.

McNeil, J. (1993). Census Bureau Data on Persons with Disabilities: New Results and Old Questions about Validity and Reliability. Paper presented at the Annual Meeting of the Society for Disability Studies, Seattle, WA.

McNeil, J. (1998). Selected 92/93 Panel SIPP Data: Time 1 = Oct.93-Jan.94, Time 2 = Oct.94-Jan.95. Unpublished Table.

Moore, J.C. (1988). Self-proxy Response Status and Survey Response Quality. Journal of Official Statistics, 4, 155–172.

O'Muircheartaigh, C. (1991). Simple Response Variance: Estimation and Determinants. In Measurement Errors in Surveys, P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds), 551–574. New York: Wiley.

Rodgers, W. and Miller, B. (1997). A Comparative Analysis of ADL Questions in Surveys of Older People. The Journals of Gerontology, 52B, 21–36.

Salthouse, T.A. (1996). The Processing-Speed Theory of Adult Age Differences in Cognition. Psychological Review, 103, 403–428.

Schwarz, N., Park, D.C., Knauper, B., and Sudman, S. (1998). Cognition, Aging, and Self-Reports. Ann Arbor, MI: Edwards Brothers.

Schwarz, N. and Wellens, T. (1997). Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Actors and Observers. Journal of Official Statistics, 13, 159–173.

Sudman, S. and Bradburn, N. (1982). Asking Questions: A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass.

Todorov, A. (2000). The Accessibility and Applicability of Knowledge: Predicting Context Effects in National Surveys. Public Opinion Quarterly, 64, 429–451.

Tourangeau, R., Conrad, F.G., Arens, Z., Fricker, S., Lee, S., and Smith, E. (2006). Everyday Concepts and Classification Errors: Judgments of Disability and Residence. Journal of Official Statistics 22, 385–418.

Tversky, A. and Koehler, D.J. (1994). Support Theory: A Non-extensional Representation of Subjective Probability. Psychological Review, 101, 547 −567.

WHO (1980). The International Classification of Impairments, Disabilities, and Handicaps – A Manual Relating to the Consequences of Disease. Geneva: World Health Organization.

Received May 2005
Revised February 2007