Journal of Official Statistics, Vol. 20, No. 2, 2004, pp. 379–391

How Good is Good? Comparing Numerical Ratings of Response Options for Two Versions of the Self-Assessed Health Status Question

Barbara Foley Wilson[1], *Barbara M. Altman*[1], *Karen Whitaker*[1], *and Mario Callegaro*[2]

Eight sets of disability questions were tested to assess their relative value for measuring impairments, functioning, and behavior. Two versions of a self-assessed health status question (SAHS) appeared among the eight sets. While both had 5-point response option scales containing "Good" and "Very good", the rank order differed within the scales. A small experiment was conducted to explore whether "Good" and "Very good" meant the same thing to respondents when they were presented within the two different response option scales. Participants wrote each response option set on two Visual Analog Scales at different times. "Good" received a lower numerical rating (5.4 on a scale from 1 to 10) when it was the third option after "Excellent" and "Very good" than when it was the second option (7.3) after "Very good." Findings are presented in the context of past research on SAHS questions and rating scales. The results are relevant for making cross-survey comparisons.

Key words: Cognitive testing; disability measures.

1. Introduction

In the summer of 2002, eight sets of disability questions were tested in the Questionnaire Design Research Laboratory (QDRL) to assess their relative value for measuring impairments, functioning, and behavior of persons with disability. Two versions of the self-assessed health status question (SAHS) appeared among the eight sets of questions. One, used in the World Health Organization Disability Assessment Schedule (WHODAS), asks "In general, how would you rate your health today. . . "Very good," "Good," "Moderate," "Bad," "Very Bad?" Thus, the WHODAS question has a bipolar scale with five response options that are balanced around the midpoint "Moderate" and has "Very good" as the highest response option and "Very bad" as the lowest. The bipolar WHODAS question also specifies "today."

The other question, used in Australia, Canada and many U.S. surveys, asks, "In general, would you say your health is. . . Excellent, Very good, Good, Fair, or Poor?" This version of the SAHS question also has "Good" and "Very good" as options but has a unipolar scale with "Excellent" as the highest response option and "Poor" as the lowest. Another

[1] National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782-0003, U.S.A. Email: bfw3@cdc.gov
[2] University of Lincoln, Nebraska, 200 North 11th Street, Lincoln, NE, U.S.A. Email: mca@unlserve.unl.edu
Acknowledgments: We are grateful to our colleague Paul Beatty and to the anonymous referees and the associate editor for their comments and suggestions. Our work also benefitted from correspondence and data from Dr. Robert Hauser, Vilas Research Professor of Sociology, University of Wisconsin-Madison.

difference between the questions is that the bipolar WHODAS question asks the respondent to "rate" his or her health while the unipolar question just asks how "is" his or health. In spite of these differences, we know from comments made during cognitive testing in a previous QDRL project that the questions seem the same to some respondents.

The design of our study, asking eight sets of questions that were similar but not exactly the same gave us the opportunity to explore whether "Good" and "Very good" mean the same thing to respondents when they are presented in the two different scales. This pertains to the broader issue of how the labeling of scale points changes the meaning of the scale, even though there may be terms in common. We wondered in this case whether the QDRL participants made their choices based on the connotation of a word or the rank order of its presentation.

1.1. International comparisons

1.1.1. Bipolar WHODAS question

The issue of whether "Good" and "Very good" mean the same thing may have some relevance to researchers who use different questionnaires to make international comparisons. The bipolar WHODAS question is included in a survey of 70 countries. A slightly different version of it has been used in the United Kingdom on the Health Survey for England (HSE) since 1993 and the Omnibus survey in 1997 (Sturgis et al. 2001): "How is your health in general? Would you say it was very good, good, fair, bad, or very bad?" (This version does not use the word "rate," uses "fair" rather than "moderate" as the midpoint, and has the term "in general" at the end rather than the beginning of the question.) In 1997, 77 percent of U.K. men and 74 percent of U.K. women self-reported "good" or "very good" health.

In describing the purpose of the Self-Assessed Health Status (SAHS) question in the U.K., Sturgis et al. declare that the questions "appear to have been devised and included in surveys mainly on criteria of face validity and practicality" (Sturgis et al. 2001, p. 83). They found no theoretical derivation or methodological development. They are also unaware of "any clear published statements of what each of these questions is intended to measure." However, they point out that the general health question is widely used on surveys, is short, easy to answer, and easy to administer. Cognitive work done to explore respondents' interpretation of the term "health in general" found that respondents understood the term to encompass an absence of ill-health, the ability to lead a normal life, a state of mind, physical fitness, frequency of doctor visits and ability to go to work or school.

The bipolar SAHS question has been used since 1981 in the Netherlands to study trends in inequalities in self-reported health and health-related behavior. Analysts noted the widening of inequalities of health in relation to income and speculated that it might be due, in part, to policy measures that reduced coverage and generosity of disability benefits (Dalstra et al. 2002).

1.1.2. The unipolar Australian, Canadian and U.S. question

A unipolar question is used in Australia, Canada, and in many U.S. surveys. Statistics derived from the question are routinely reported. For example, according to the Australian Bureau of Statistics (2002), 83 percent of Australians reported their health status as "good,"

"very good" or "excellent" in the 1995 National Health Survey. And, according to the Centers for Disease Control and Prevention (2002), 85 percent of U.S. respondents said their health was "good," "very good," or "excellent" in 2000.

The unipolar SAHS has been shown to be a good predictor of mortality (Idler and Benyammi 1997) and functional ability (Idler and Kasl 1995). It has also proved to be highly correlated to health costs. Bierman et al. (1999) evaluated how well the question predicted the financial health of Medicare managed care plans. They found that: "Medicare expenditures had a marked inverse relation to self-assessed health rating – those who rated their health the worst spent the most on health care." Data from both the Behavioral Risk Factor Surveillance System (BRFSS) (Washington State Department of Health 2002) and the National Health Interview Survey (NHIS 2002) show inequalities in self-rated health by age, education, income and race/ethnicity.

1.2. Number of scale points and choice of bipolar or unipolar scales

Although both self-assessed general health questions we are considering here have 5-point verbal scales, a potentially important difference is polarity – that is, whether a set of response options has a clear conceptual neutral midpoint (bipolar), or whether the response options represent varying levels of some construct with no conceptual midpoint and with a zero point at one end (unipolar). In describing how to design rating scales for effective measurement in surveys, Krosnick and Fabrigar (1997) drew upon the empirical research to explore the effect of the number of scale points on reliability and validity. Their research used a variety of approaches, including secondary analyses of existing data and direct experimental comparisons. The researchers list several decisions that must be made in terms of how long the scale should be, whether there should be a midpoint, whether the labels should be numeric or verbal and, if verbal, whether all points should be labeled. They cite Matell and Jacoby (1971), who found that reliability and validity of bipolar scales are highest for about 7 points. They also cite Wikman and Wärneryd (1990), who found, in contrast, that the reliability and validity of unipolar scales seem to be optimized for scales approximately 5 points long. Culling through decades of research, Krosnick and Fabrigar further found that data quality is better and that respondents are more satisfied when all scale points are labeled with words (Dickinson and Zellinger 1980). On the basis of work by Klockars and Yamagishi (1988), Krosnick and Fabrigar advise that labels should have meanings that divide up the continuum into approximately equal units and recommend that scales be constructed to capture the differentiations people make naturally. They recommend that unipolar scales have four to seven points and that bipolar scales have seven to nine points, but also that the questionnaire designer consider whether respondents would reasonably want two or three points on each side of the midpoint.

Krosnick has also compiled a table showing numeric ratings of quality terms in which he averages the numeric ratings found in eleven studies conducted between 1941 and 1991. In this table, "Excellent" has a mean value of 92, "Very good" is 79, "Good" is 68, "Fair" is 51, "Poor" is 23, "Bad" is 20, and "Very bad" is rated 14 on a scale from 0 to 100 (Krosnick 2003).

2. Methodology

2.1. Cognitive lab participants

Since the purpose of the QDRL project was to assess the relative value of eight sets of questions to measure impairments, functioning, and behavior of persons with disability, lab participants were recruited through a newspaper advertisement asking for volunteers who are limited in any way in any activities due to physical, mental, or emotional problems or who need or use special equipment such as wheel chairs, walkers or hearing aids. Sixteen people were interviewed for 90 minutes and completed all items on both Visual Analog Scales and both SAHS questions. A 17th person completed most, but not all, scale ratings. This person's data was included where it was complete. There were ten men and seven women; eight were Non-Hispanic black, eight were Non-Hispanic white, and one gave her race as both black and white. Their ages ranged from 32 to 83. Their education ranged from 8 to 19 years. They had a variety of health conditions such as diabetes, osteoarthritis, rheumatoid arthritis, amputations, fibromyalgia, multiple sclerosis, depression, anxiety, schizophrenia, hearing and vision problems, HIV, cerebral aneurysm, post-polio syndrome, and congestive heart failure.

2.2. Testing protocol

Once in the QDRL people were told that they would be asked eight alternative sets of questions about disabilities and with each new set they were asked to try to wipe their mental slate clean and start over just as though they had not been answering similar questions. This was done to lessen any need they might feel to be consistent. It was also intended to reduce confusion or annoyance at being asked the same or similar questions repeatedly. The lab participants accepted the ground rules.

Data for the experiment were gathered from the QDRL participants by asking them to mark Visual Analog Scales to show where the two different sets of response options should be located. Visual Analog Scales can be very elaborate, but in this case they were simply two pieces of paper, each with a straight horizontal ten-inch line marked with numbers at every inch. Participants drew marks to show where each set of response options would fall along the lines from 1 (bad health) to 10 (good health). The Visual Analog Scales were marked at two different times during the interview.

The presentation order of the eight sets of questions was rotated, as was the presentation of the Visual Analog Scale task, so that some people did the bipolar WHODAS question first and others did the unipolar question first. The instruction for the unipolar question was, "The ruler on this page shows the numbers 1 to 10 where 10 means as good as your health can be and 1 means as bad as your health can be. Please show me where you think excellent, very good, good, fair and poor should be by putting marks on the line and writing the words above the marks." The instruction for the WHODAS question was comparable.

3. Findings

3.1. Numerical equivalence of verbal labels

"Good" and "Very good" received different numerical ratings on the Visual Analog Scale for the two Self-Assessed Health Status questions. The response option "Good" in the bipolar WHODAS scale was given an average rating of 7.3 inches by the seventeen QDRL participants. This was substantially higher than the 5.4-inch average rating that "Good" received in the unipolar scale. Similarly, "Very good" was rated higher (9.4) in the bipolar WHODAS scale, which did not have an "Excellent" option, than in the unipolar scale (7.8), which did. See Table 1.

3.2. Verbal labels

Comparing the lower ends of both scales suggests that the choice of words does matter. As the fourth of five options in the unipolar question, "fair" was given an average rating of 3.8. But it is still better than "bad," the fourth option in the bipolar WHODAS, which was rated 2.8. "Poor" also was rated higher (1.5) than "very bad" (1.2).

3.3. Positivity bias

When it came to rating their own health for the SAHS questions, the QDRL lab participants, all of whom had disabilities, were generally positive. Thirteen out of seventeen (76 percent) answered that their health was "good," "very good," or "excellent" on the unipolar scale and twelve out of seventeen (71 percent) answered "good" or "very good" on the bipolar WHODAS scale (Figures 1a and 1b).

Although some of the participants had serious conditions, it was apparent that they did not regard their own health as "bad" or "very bad." In the bipolar WHODAS scale only one of the seventeen people did so. In contrast, three of the same seventeen people were willing to rate their health as "fair" and one rated her health "poor" in the unipolar scale.

3.4. Consistency of response

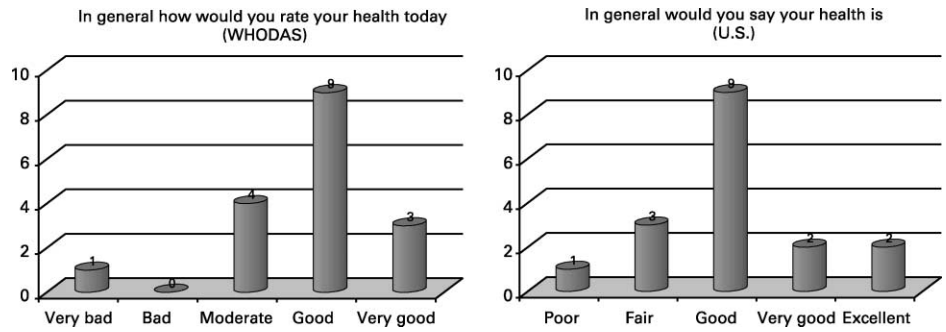
Some people were consistent in their choice of verbal label when asked to rate their own health in the two SAHS questions. Four of the seven women and three of the ten men answered "good" to both self-assessed health questions even though the numerical value they assigned to the words differed for the two theoretical Visual Analog Scales (Table 2). Other participants seemed inclined to be consistent about the rank order of their choice. Five of the ten men and one of the seven women chose the response options with the same rank order even though the label was different. Thus, for some of the QDRL participants there was an apparent disconnect between how they rated their own health and the numerical value they applied to the words in the two scales.

3.5. The importance of label selection

Table 2 shows how the participants rated their own health for each question and the measurement mark they assigned to the label on the Visual Analog Scale. The average

Table 1. Measurement in inches on Visual Analog Scale of response options for two versions of the Self-Assessed Health Status question by 17 cognitive lab participants

	Unipolar Question: "In general would you say your health is Excellent, Very good, Good, Fair or Poor?"					Bipolar WHODAS question: "In general, how would you rate your health today: Very good, Good, Moderate, Bad, Very Bad?"				
	Excellent	Very good	Good	Fair	Poor	Very good	Good	Moderate	Bad	Very Bad
Women										
1	10.0	9.0	6.5	4.0	1.0	10.0	7.0	5.0	3.0	1.0
2	10.0	8.0	5.0	3.0	1.0	10.0	8.0	5.0	3.0	1.0
3	9.0	7.5	5.0	3.0	2.0	9.5	7.0	4.0	2.0	0.5
4	10.0	8.0	6.0	4.0	2.0	10.0	7.0	5.0	3.0	1.0
5	10.5	8.0	6.0	4.0	0.5	8.0	5.0	3.0	2.0	0.5
6	10.0	8.0	6.0	3.5	1.3	9.5	8.5	5.5	3.5	1.8
7	8.0	–	–	4.0	2.0	8.0	7.0	6.0	2.0	1.0
Men										
8	10.0	8.0	6.0	4.0	3.0	9.0	7.0	6.0	4.0	3.0
9	10.0	8.0	5.0	3.0	1.5	10.0	8.0	5.0	3.0	1.0
10	9.0	7.5	5.0	4.5	1.0	10.0	8.0	5.0	3.0	1.0
11	9.5	8.0	5.5	3.0	1.0	9.5	8.0	5.5	3.0	1.5
12	10.0	4.0	2.0	6.0	1.0	8.0	5.0	3.0	2.0	1.0
13	10.0	8.0	7.0	6.0	5.0	10.0	9.0	8.0	2.0	1.0
14	10.0	8.0	5.0	3.0	1.0	9.0	7.5	5.5	3.5	2.0
15	10.0	9.0	6.0	3.0	1.0	9.0	6.0	5.0	3.0	1.0
16	10.0	8.5	6.0	3.5	1.0	10.0	8.5	6.5	2.5	1.0
17	10.0	7.5	5.0	2.5	1.0	10.0	8.0	5.0	2.5	1.0
Average	9.8	7.8	5.4	3.8	1.5	9.4	7.3	5.2	2.8	1.2



Figs. 1a and 1b. Number of participants who chose each response option for two versions of Self-Assessed Health Status questions

numerical value that people assigned to the words they chose for their own self-assessment was higher in the bipolar WHODAS scale (6.9) than in the unipolar scale (5.5).

Table 3 summarizes the self-ratings and sheds light on the question of whether people relied on the verbal labels for their self-assessment or on the rank order of the presentation of the response options. If the QDRL participants were relying on rank, the numbers in the cells would fall on a diagonal line from highest on the left to lowest on the right. Of the

Table 2. Participant's ratings of their own Self-Assessed Health Status (SAHS) and the measurement mark they assigned to the label on the Visual Analog Scale

	Unipolar Question*		Bipolar WHODAS question**	
		Visual Analog Scale		Visual Analog Scale
Women				
1	Fair	4.0	Moderate	5.0
2	Good	5.0	Good	8.0
3	Good	5.0	Good	7.0
4	Good	6.0	Good	7.0
5	Poor	0.5	Very bad	0.5
6	Good	6.0	Good	8.5
7	Fair	4.0	Moderate	6.0
Men				
8	Good	6.0	Good	7.0
9	Good	5.0	Good	8.0
10	Fair	4.5	Moderate	5.0
11	Good	5.5	Very good	9.5
12	Good	2.0	Moderate	3.0
13	Very good	8.0	Good	9.0
14	Good	5.0	Good	7.5
15	Excellent	10.0	Very good	9.0
16	Excellent	10.0	Very good	10.0
17	Very good	7.5	Good	8.0
Average		5.5		6.9

Note: *The unipolar question is "In general, would you say your health is: excellent, very good, good, fair, or poor?" **The bipolar WHODAS question is "In general, how would you rate your health today: very good, good, moderate, bad, very bad?"

Table 3. Summary of responses to two different Self-Assessed Health Status questions

Bipolar WHODAS response options	Unipolar question response options:				
	Excellent	Very good	Good	Fair	Poor
Very good	2		1		
Good		2	7		
Moderate			1	3	
Bad					
Very bad					1

seventeen disabled participants, a minority of six chose self-assessments that had the same rank in the two scales. The majority seemed to have made their choices based on the verbal labels. In fact seven participants chose "Good" as their self-assessment regardless of whether it was the second item in the bipolar WHODAS scale or third item in the unipolar scale. Table 3 also shows is that if people in the general population behaved the way this small sample did, a substantial minority might choose a different option in response to the two scales.

4. Discussion

The choice of verbal labels for rating scales that measure subjective phenomena has received a great deal of study. In reviewing the body of work, Krosnick and Fabrigar note: "If verbal labels are to be useful, they must have reasonably precise meanings for respondents. It is also important that the labels one chooses reflect relatively equal intervals along a continuum, particularly if an analyst is to capture all variance in the latent construct and plans to treat the results as an interval-level variable in statistical analysis" (1997, p. 150).

When the QDRL participants were marking the Visual Analog Scale, many of them seemed to approach the task by marking the line at 10 inches if the response option was "excellent" and something less than 10 if it was "very good." Then they arrayed the remaining options at intervals of approximately two inches, but moved the mark up or down depending on the verbal label.

The question naturally arises as to how typical the QDRL participants were in their judgments on the Visual Analog Scale. Johnson et al. (1997) published results of a study of response differences among culturally diverse populations where they used a methodology similar to ours, using an 11-point semantic differential ratings scale to assess the degree of good or ill health represented by each of the five precoded responses to a variant of the unipolar SAHS question, "Would you say your health is excellent, very good, good, fair, or poor?" This variant lacks the term "in general." The scale included a zero and was not a Visual Analog Scale but the methodology of associating a verbal label with a number was similar. Also, unlike the QDRL participants, the Johnson study participants did not have disabilities. In comparing cross-cultural groups Johnson et al. found considerable agreement regarding the numerical values assigned to the two most positive health ratings ("excellent" and "very good"). Differences by race were observed, though, for more neutral and less positive responses. Table 4 compares the numerical ratings found by Johnson et al. with those collected in the QDRL study.

Table 4. Average numerical equivalence ratings for response options to the Unipolar Self-Assessed Health Status question by race: Two studies

In general would you say your health is. . .	African American		White Non-Hispanic	
	Johnson, et al. ($N=109$)	QDRL ($N=8$)	Johnson, et al. ($N=108$)	QDRL ($N=8$)
Excellent	9.6	9.6	9.8	9.8
Very good	8.1	7.6	7.9	8.0
Good	6.4	5.4	6.1	5.4
Fair	4.3	4.4	4.0	3.1
Poor	0.9	1.9	1.4	1.3

The differences found between the bipolar WHODAS and the unipolar version of the SAHS may also be partially attributable to "floor and ceiling effects." If a large proportion of the responses fall in the highest category, there is a "ceiling" that limits differentiation. If too few respondents choose the lowest category, the "floor" may be too low. Sturgis et al. caution: "Floor and ceiling effects concern both the sensitivity of instruments to differences in health state between population sub-groups and also their ability to detect longitudinal changes in health state at the population level" (Sturgis et al. 2001, p. 71).

The question also arises as to why people with chronic conditions and disabilities, like the population in general, rated their own health as "good" or "very good," and occasionally "moderate," but clearly shunned "bad" and "very bad." When asked to explain their ratings, QDRL participants said that their health meant the condition of their vital organs as measured by blood pressure, blood tests, cholesterol, etc. Even with mobility, vision and hearing problems, their health could be good. One man who has both diabetes and heart disease said his health was "Good" because his conditions were "under control." The rare QDRL participants who rated their health as "Poor" have had conditions that cannot be controlled or have things that affect cognition or mood.

The positivity bias found in our QDRL participants is also seen in national distributions of the U.K., the U.S., Australia and Canada. The positive skew could be a problem in data analysis because an artificially low ceiling may produce ordinal, rather than interval, data. McCarty and Shrum (2000, p. 271) caution, ". . .because personal values are inherently positive constructs, respondents often exhibit little differentiation among the values and end-pile their ratings toward the positive end of the scale. Such lack of differentiation may potentially affect the statistical properties of the values and the ability to detect relationships with other variables." Because of the positivity bias the scale may function like an ordinal rather than an interval scale.

The QDRL results are clear that the participants did not assign the same numerical value to the same words (Good and Very good) when they were used in two different scales. On the other hand some participants did choose the same term to describe their own health when asked the two questions. This paradox may be explained by work done by Wildt and Mazis (1978). They designed a study to test two hypotheses: one that scale response is a function of scale position and the other that scale response is a function of scale labels. Using the Chi-square statistic they found no consensus in results. "Rather, there is some indication that both scale label and position influence response, and that an interaction or

extraneous factor is also influencing response" (Wildt and Mazis 1978, p. 265). Thus in comparing results from two similar but not quite the same SAHS questions, it should not be assumed that the meanings of scale points are equivalent, or that the meanings of verbal labels are equivalent if they occupy different scale points, because scale points and words combine to create meaning.

As is typical of qualitative studies, our QDRL sample was small and biased, with participants selected for having disabilities. A larger survey done with the general population shows findings similar to those regarding our small sample. The 1992 Wisconsin Longitudinal Survey (WLS) used still another version of the SAHS (Hauser and Freese 2003). The WLS question is close to the WHODAS with a restriction of "at the present time" functioning much like the "today" and using the term "rate your health." The question also has a second part, "compared with people your age and sex" that implies that some difference in reference period or reference group should be inferred for the first part of the question. The WLS is asked in a mail survey and the response options are presented in a table from lowest (shown with a number 1) to highest (shown with a number 5). For both 1992 and 2003 the WLS asks:

1. How would you rate your health. . .
Circle one number for each lettered item.

	Very Poor,	Poor,	Fair,	Good,	Excellent
a. at the present time?	1	2	3	4	5
b. compared with other people your age and sex?	1	2	3	4	5

In 2003, in addition to asking the 1992 SAHS again in the mail survey, there was a preceding telephone interview (with the same individuals) that used the unipolar SAHS question, so there is data from the same people for both versions of the question. There are five differences between the mail WLS and the phone WLS. 1) The mode for the two versions differs (phone and mail). 2) The presentation of the options is reversed. 3) The context may be affected because the second part of the question asks for a comparison with other people of the same age and sex. 4) The mail WLS does not have the response option "Very good" and has a lower "floor" offering the option "Very poor." Thus, the mail version shares three terms with the phone question that are in different rank orders (Good, Fair, and Poor). The mail version also has "Excellent" in common with the phone version and it is in the same rank order. 5) Having respondents circle a number (rather than answering the phone interviewer) reinforces the concept of this being an interval scale. The interviewer-administered telephone version gives the respondent some cognitive wiggle room. "Excellent" may seem further from "Good" than "Poor" is from "Very poor." For example, in Krosnick's table of mean numerical ratings of quality terms, "Excellent" is 92 and "Good" is 68, while "Poor" is 23 and "Very poor" is 11 (Krosnick 2003).

Table 5 shows the results for 536 WLS respondents in 2003.

1. As could be expected, WLS respondents rated their health as above average on both versions of the SAHS. There is a positivity bias in both distributions of answers.
2. A substantial number of people chose different response options in the two scales. Of course, respondents who answered "Very good" in the phone survey did not even

Table 5. Consistency of responses to each option on the unipolar self-assessed health status (SAHS) phone survey question according to subsequent responses to a different version of the SAHS by the same 536 respondents on the 2003 mail version of the Wisconsin Longitudinal Survey (WLS)

Response options for phone WLS	Response options to SAHS on subsequent 2003 WLS mail survey					
	Very poor 1	Poor 2	Fair 3	Good 4	Excellent 5	Total number
Excellent	1	0	1	58	***92***	152
Very good	0	0	9	*168*	31	208
Good	0	0	*33*	**107**	0	140
Fair	0	*6*	**14**	7	0	27
Poor	*1*	**4**	4	0	0	9
Total number	2	10	61	340	123	536

have that option in the mail version, but of the 328 people who did have the option of being consistent because four terms were the same in both scales (poor, fair, good, excellent) only 217 (66.2 percent) of respondents gave the same response. (The number of respondents who answered the same in both versions is shown in bold font in Table 5.)

3. In the phone question, 208 respondents chose "Very Good," an option that was not provided in the mail survey. When they later answered the mail survey, these respondents had to choose a higher or lower response. Only 31 chose the higher response "Excellent," while most (168) chose the lower "Good." The remaining 9 chose an even lower option ("Fair"). Of the 152 people who chose "Excellent" on the phone, only 92 chose that on the mail survey. What may have been operating was the tendency to give more socially desirable answers to an interviewer on the phone than on mail surveys (Hochstim 1967).
4. WLS researchers observe that the data suggest that the scale position of the category matters more in determining respondents' choice of response options than the verbal labels. They point out that there is somewhat more grouping on the diagonal of the cross-tabulation than there is matching on identical verbal categories.

5. Conclusion

The difference in response options in the bipolar WHODAS Self-Assessed Health Status question and the unipolar question will distort direct comparisons between countries using the different response options for these seemingly comparable questions. For many people, "Good" does not mean the same thing in the two questions, and "Bad" and "Very bad" definitely are not equivalent to "Fair" or "Poor." The shift in meaning can be due to the extreme adjective "Very bad" of the bipolar WHODAS question and to the lower "ceiling" on the question which does not offer an "Excellent" option.

The deceptively simple but truly useful self-assessed health questions are widely used and deserve a standard format that would enhance comparability. The cognitive, statistical, and linguistic aspects of the question variations should all be evaluated by the interdisciplinary research community and consensus reached. There may be linguistic reasons for the WHODAS terms. However, we do think that "fair health " is a better

midpoint in English than "moderate health." The terms "Poor health" and "Bad health" could be studied cognitively to find out how they are interpreted. Or, it may be that the term "Bad health" is preferable to "Poor health" for the practical reason that its meaning is more constant when translated across languages. In Krosnick's table "Bad" usually is given a slightly lower numeric equivalent than "Poor." It should be noted, though, that while "Bad" on the face of it seems like a linguistic opposite of "Good," the two terms are not balanced about the scale's midpoint. In Krosnick's table "Bad" is 20, considerably farther from "Fair" (51) than "Good" is at 68.

We also think that the bipolar WHODAS question, lacking "Excellent" as it does, has a ceiling that is too "low" to distribute the positivity bias found in the response distributions. There should be an "Excellent" option.

Another issue is polarity. Work should be done to establish whether the underlying construct is bipolar or unipolar in respondents' minds. If a seven point bipolar scale is acceptable, the response options "Excellent," "Very good," "Good," "Fair," "Slightly bad," "Bad," and "Very bad" would produce points that would be rather evenly distributed as measured by the average values in Krosnick's table (92, 79, 68, 51, 32, 20, 14). The disadvantage is that the low end of the scale gets so few responses already that the data are usually collapsed for analysis. Furthermore, while a seven-point scale would work easily in a mail or Web survey, it would be harder to administer by phone or in person than the current five-point scales.

References

Australian Bureau of Statistics (2002). Website: http://www.abs.gov.au/ausstats

Bierman, A.S., Bubolz, T.A., Fisher, E.S., and Wasson, J.H. (1999). How Well Does a Single Question About Health Predict the Financial Health of Medicare Managed Care Plans. Effective Clinical Practice, 2, 56–62. See Http://www.ahcpr.gov/research/jun99/ra13.htm

Centers for Disease Control (2002). See http://apps.nccd.cdc.gov/brfss

Dalstra, J.A.A., Kunsht, A.E., Geurts, J.J.M., Frenken, F.J.M., and Mackenbach, J.P. (2002). Trends in Socioeconomic Health Differences in the Netherlands, 1981–1999. Journal of Epidemiology and Community Health, 56, 927–934.

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.

Hauser, R.M. and Freese, J. (2003). Personal communication.

Hochstim, J.R. (1967). A Critical Comparison of Three Strategies of Collecting Data from Households. Journal of the American Statistical Association, 62, 976–989.

Idler, E.L. and Benyammi, Y. (1997). Self-rated Health and Mortality: A Review of Twenty-seven Community Studies. Journal of Health and Social Behavior, 38, 21–37.

Idler, E.L. and Kasl, S.V. (1995). Self-ratings of Health: Do They Also Predict Change in Functional Ability? Journal of Gerontology, 50B, S344–S353.

Johnson, T., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., Lacey, L., and Horm, J. (1997). Social Cognition and Responses to Survey Questions Among Culturally Diverse

Populations. In Survey Measurement and Process Quality, L. Lyberg, et al. (eds). New York: Wiley, 87–113.

Klockars, A.J. and Yamagishi, M. (1988). The Influence of Labels and Positions in Rating Scales. Journal of Educational Measurement, 25, 85–96.

Krosnick, J.A. (2003). Advanced Questionnaire Design Class Notes. Essex Summer School in Social Science Data Analysis and Collection. Unpublished manuscript.

Krosnick, J.A. and Fabrigar, L.R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In Survey Measurement and Process Quality, L. Lyberg, et al. (eds). New York: Wiley, 141–164.

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, 32, 657–674.

McCarty, J.A. and Shrum, L.J. (2000). The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures. Public Opinion Quarterly, 64, 271–298.

Sturgis, P., Thomas, R., Purdon, S., Bridgwood, A., and Dodd, T. (2001). Comparative Review and Assessment of Key Health State Measures of the General Population. United Kingdom Department of Health Web Site. http://www.doh.gov.uk/pdfs/healthreport.pdf

Washington State Department of Health, Center for Health Statistics (2002). http://www.doh.wa.gov/EHSPHL/CHS/CHS-Data/brfss_homepage.htm

Wikman, A. and Wärneryd, B. (1990). Measurement Errors in Survey Questions: Explaining Response Variability. Social Indicators Research, 22, 199–212.

Wildt, A.R. and Mazis, M.B. (1978). Determinants of Scale Response: Label Versus Position. Journal of Marketing Research, 15, 261–267.

Received February 2003
Revised December 2003