American Journal of Epidemiology
Vol. 174, No. 8
DOI: 10.1093/aje/kwr204
Advance Access publication:
September 2, 2011

Practice of Epidemiology

Reliability of Self-rated Health in US Adults

Anna Zajacova* and Jennifer Beam Dowd

* Correspondence to Dr. Anna Zajacova, Department of Sociology, College of Arts and Sciences, University of Wyoming, Department 3293, 1000 East University Avenue, Laramie, WY 82071-2000 (e-mail: zajacova@uwyo.edu).

Initially submitted February 3, 2011; accepted for publication May 25, 2011.

General self-rated health (SRH) is widely used to study trends and inequalities in population health. Recently, there has been an increased interest in understanding the measurement properties of SRH. This study evaluated for the first time the test-retest reliability of SRH among US adults. Analyses were based on a nationally representative sample of 9,235 adults interviewed in the 2005–2008 National Health and Nutrition Examination Survey (NHANES). Respondents reported SRH on 2 occasions (about 1 month apart). Kappa statistics, polyserial correlations, and agreement tabulations were used to assess reliability across population subgroups; regression models tested the association of sociodemographic factors and the stability of the rating. Nearly 40% of respondents changed their health rating between interviews, indicating moderate test-retest reliability of SRH. Reliability differed significantly by sociodemographic characteristics: Racial/ethnic minorities and adults with less education had lower reliability of SRH judgments. Health events between interviews did not influence consistency, but conditional on a rating change, they increased the odds of downgrading one's health. The results suggest that 1) there is a substantial amount of error in individuals' self-assessment of health and 2) reliability is worse for disadvantaged sociodemographic groups, potentially biasing estimates of health inequalities among US adults.

adult; epidemiologic measurements; population; reliability; self report; United States

Abbreviations: NHANES, National Health and Nutrition Examination Survey; SRH, self-rated health.

Self-rated health (SRH), a single-item ordinal measure with 5 levels, is a widely used indicator of general health status in epidemiologic and population health research. One reason for its pervasive use is the belief that SRH has high predictive and concurrent validity, as measured by its association with subsequent mortality and varied measures of morbidity, disability, and utilization of health services (1–5).

A necessary, although not sufficient, condition of validity is reliability (also referred to as ''consistency,'' ''reproducibility,'' or ''stability'') or the degree to which repeated measures remain unchanged under the assumption that the measured construct has not changed (6). From a modeling perspective, low reliability contributes to measurement error, which often biases results of analyses in which the construct is used as an independent or dependent variable (7). Furthermore, reliability may differ across population subgroups, affecting empirical comparisons of these groups. Differential reliability might be of particular concern with SRH, which is frequently used to measure health disparities in the US population.

We are aware of only 2 studies, both using non-US data, that analyzed the consistency of general SRH (8, 9). Both found substantial variation in the ratings across 2 administrations of the health status questionnaire. However, cross-country differences in the interpretation and reporting of subjective health are well known and, thus, these results may not be generalizable to the US population (10). The present study tests for the first time the test-retest reliability of general SRH in a nationally representative sample of US adults, with responses from 2 interviews about 1 month apart. Reliability is also compared across key sociodemographic characteristics including age, gender, race, and level of education.

MATERIALS AND METHODS

Data

The analyses are based on data from the National Health and Nutrition Examination Survey (NHANES), 2005–2008

(11). The continuous survey collects an extensive range of sociodemographic, lifestyle, and health-related information from a nationally representative sample of the noninstitutionalized civilian US population, by using a complex probability sampling design with an oversample of African Americans and Mexican Americans. Data were collected in 2 parts, first in a household interview, followed about 1 month later by a physical examination and additional questionnaire items administered in a mobile examination center. The total household response rate for the interview was 74.9% combined for both waves; about 96% of interviewed adults also participated in the medical examination. Detailed information about response rates by sex and age is available in the NHANES documentation (12).

The analytical sample was defined as respondents aged 20–80 years who answered both self-rated health questions without the use of a proxy. Of the 10,566 adults aged 20–80 years in the NHANES sample, all but 7 (0.07%) had a valid SRH value from the household interview, while 1,227 individuals (11.6%) were missing the second SRH rating because they either did not participate in the mobile examination center examination or they did not answer the part of the questionnaire that included the second SRH item. Finally, an additional 97 adults (1.04%) who answered one or both interviews with the help of a proxy were dropped from the analysis, leaving a final analytical sample of 9,235 respondents.

We conducted bivariate and multivariate analyses to understand the correlates of nonresponse. Adults excluded from analyses ($n = 1{,}331$) were significantly more likely to be either younger than 40 years or older than 60 years, nonwhite, less educated, and in ''poor'' health (results available on request). Because these characteristics also tended to be associated with lower reliability, our results provide a conservative estimate of the overall variability in SRH ratings.

Measures

Self-rated health. SRH was assessed identically in 2 face-to-face interviews conducted by a trained NHANES interviewer: ''In general, would you say your health is excellent, very good, good, fair, or poor?'' Because the number of response categories influences reliability (13), we analyzed the measure using both of these original 5 categories, as well as the frequently used dichotomized version of excellent, very good, or good versus fair or poor (2, 14, 15). The interviews differed in their setting: The first interview was conducted in the respondents' homes, and the second one was in a mobile examination center as a part of a medical examination.

Respondents' characteristics. Age was categorized as 20–39, 40–59 (reference), and 60–80. Race was included as non-Hispanic white (reference), non-Hispanic black, Mexican American, and ''other.'' The ''other'' category included all non-Mexican Hispanics and also all non-Hispanics from racial groups other than white or black. Education was coded in 5 categories: 0–8 years of schooling, 9–11 years, 12 years, some college or associate's degree, and bachelor's degree or higher. Descriptive analyses showed education to be linearly related to reliability; for parsimony, we entered this variable continuously in multivariate models.

Time between interviews. The data included age in months at the household and mobile examination center interviews. Using these 2 pieces of information, we calculated the approximate time difference between the interviews. We included it in analyses as 0–1 month (reference), 2 months, and ''missing.'' This last category comprised 3.7% of respondents who did not have valid information on age in months so we could not calculate the time between interviews.

Intervening health. During the second interview, respondents were asked several questions pertaining to their health during the previous 30 days. For most respondents, this time period corresponded to the time between interviews, so it potentially influenced changes in health ratings. The questions included getting a head or chest cold; stomach or intestinal illness; flu, pneumonia, or ear infection; and additional questions about the number of days with poor physical or mental health. We created 2 summary variables: a dichotomous variable capturing any or no illness (reference) and the total number of days in poor physical and/or mental health (continuous, range of 0–60).

Analysis

We calculated 5 measures of agreement and association between the pairs of health ratings. First, polychoric correlation coefficients, a measure of association for ordinal data such as SRH (16), can be interpreted similarly to a Pearson correlation. Second, kappa coefficients measured agreement between the 2 ratings beyond what would be expected by chance. Guidelines for interpreting kappa suggest that kappa below 0.4 indicates low agreement, from 0.4 to below 0.6 moderate agreement, and from 0.6 to below 0.8 substantial agreement (6). Third, weighted kappa with linear weights (17) incorporated information about the distance between the 2 ratings, so that ratings 1 category apart counted as ''less disagreement'' than a pair of ratings 2 categories apart. This index depends on the choice of weights used for the estimation and is best suited to comparing population groups. Details about the calculation and interpretation of kappa and weighted kappa coefficients are available elsewhere (17, 18). Fourth, we tabulated the proportion of respondents who changed their rating between interviews. Fifth, we calculated the proportion who would be classified differently across interviews if SRH were dichotomized as excellent/very good/good versus poor/fair.

Next, we compiled a table of all pairs of responses or a ''transition'' matrix. This table shows the distribution of SRH levels at the second interview for each SRH level reported at the first interview. Finally, we estimated 2 models of change to examine reliability by sociodemographic factors, baseline health, and recent illness: a logistic model of any change in the SRH rating versus no change and a multinomial model that, conditional on a rating change, distinguished whether the second rating was better or worse than the first one.

We conducted extensive sensitivity analyses to ascertain the robustness of the findings. Among these, models that estimated the magnitude of change (ordered logistic models of no change vs. differences of 1 SRH level, 2, or more levels) yielded results very similar to those shown here. Because the educational attainment of adults in their early twenties is

often not completed, we also reestimated all models for adults aged 25–80 years. This restriction also did not change our conclusions; all results are available on request.

Analyses were conducted by using STATA, version 11.0, software (19). With the exception of polychoric correlation and kappa coefficients (for which the adjustment procedures were not available), the analyses were adjusted for the complex NHANES sampling design. Readers should interpret the correlation and kappa measures accordingly.

RESULTS

At the first interview, 17% of respondents reported their health as excellent, while 32% and 34% rated their health as very good or good, respectively. Fewer than 17% judged their health to be fair or poor. Each of the measures of agreement/association shown in the remaining 5 columns in Table 1 provides a different perspective on the reliability of the SRH. For the sample, the correlation between the 2 sets of ratings was 0.75, indicating moderate to somewhat strong association. The kappa coefficient of 0.43 was at the low end of the ''moderate'' range. Overall, nearly 40% of respondents changed their rating between the 2 interviews. If SRH were dichotomized, 10.5% would be classified in a different category.

Reliability differed across population characteristics. Generally, the comparisons were similar regardless of which 1 of the 5 measures was used; we therefore comment on the overall patterns. Age had a nonlinear relation with reliability: Middle-aged respondents were more consistent than either younger or older adults. Women's ratings were slightly more stable than men's, but the differences were small. Reporting an acute illness and the number of sick days in the last 30 days did not systematically affect reliability. In contrast, considerable differences in reliability emerged by race/ethnicity. Mexican-American adults changed their health ratings more than any other groups, although black adults also had lower reliability compared with white or ''other'' adults. An even greater disparity appeared by education, whereby more schooling was associated with higher consistency. Respondents with a bachelor's degree had the highest rating stability of any group (only 34% changed their rating). The difference for the proportion who would cross a threshold using dichotomized SRH was particularly striking: Nearly 22% of those with the least education would switch categories, compared with fewer than 4% among those with the most education. These differences are substantial, even taking into account the fact that less educated adults had lower average health ratings and were thus closer to the threshold.

Table 2 shows the distribution of ratings across the 2 sets of interviews. Among respondents who rated their health as excellent during the first interview, more than half (51.8%) changed their rating, making it the least stable of the 5 SRH levels. ''Very good'' and ''good'' health categories were more stable: nearly two thirds of observations remained unchanged, followed by the lowest 2 health categories, for which about 50% of the observations did not change. Most respondents who changed ratings did so by only 1 category; however, almost 14% of the respondents who changed their rating did so by 2 or 3 levels.

Table 3 shows results from 2 models of rating change in a multivariate framework. The first is a logistic model of change versus no change. Several variables were associated with the reliability of SRH, including age, education, race/ethnicity, and the level of health reported at the first interview. Respondents older than 60 years, black and Mexican-American adults, and respondents with less schooling provided significantly less reliable health ratings than did younger respondents, white adults, and those with more schooling, respectively. The magnitude of the differences was substantial: The odds that a health rating would change increased by 11% for each education category and were over 25% higher for black and Mexican Americans compared with white adults. Respondents with a more ''extreme'' first health rating were significantly more likely to change their rating than were those who rated their health as the middle (''good'') category. This pattern may reflect a ''regression to the mean'' or a ''floor/ceiling'' effect, reflecting a limit to the range of values available to respondents at the extreme health categories (20, 21).

The second model in Table 3 distinguishes upgrading or downgrading one's rating relative to no change. Age, race, and education were significant correlates of at least one direction of change. Younger adults were more likely than the middle aged to upgrade their rating and less likely to downgrade it. Nonwhite respondents and those with less schooling were significantly more likely to rate their health as worse, relative to their white and more educated counterparts. For both directions of change, the level of the initial SRH judgment was clearly important: The better the health rating during the first interview, the more likely it was to be downgraded; the worse the initial rating, the more likely it was to be upgraded. Finally, respondents who experienced poor health between the interviews, captured particularly with the number of ''sick'' days, were significantly more likely to rate their health as worse and significantly less likely to rate their health as better, compared with those without intervening health events.

DISCUSSION

We investigated the reliability of self-rated health in a large sample representative of US adults, using data from 2 interviews about 1 month apart. A substantial proportion of individuals (40%) changed their ratings across the 2 interviews. Kappa statistics and other measures of agreement indicated only a moderate level of consistency. Most individuals who changed their ratings did so by only one level, but 14% differed by 2 or more SRH levels.

There was a strong floor/ceiling effect of the initial rating, whereby the highest health levels during the first interview were more likely to become lower, and the lowest ratings were more likely to improve. An initial rating of ''good'' health, the middle SRH level, was associated with the highest consistency. Additionally, acute health events or periods of poor health between the 2 interviews had an interesting impact on reliability, as they did not affect stability or the likelihood that a respondent would change his/her rating. Conditional on a change in the rating, however, a respondent who experienced any period of poor health was significantly more likely to ''downgrade'' his/her rating and was

Table 1. Sample Characteristics and Measures of Agreement Between Ratings, by Population Subgroups of US Adults Aged 20–80 years, NHANES, 2005–2008 ($n = 9,235$)

	Proportion, %	Polychoric Correlation	Kappa	Weighted Kappa	% Changed Rating	% Changed if Dichotomized
Total	100	0.75	0.43	0.56	39.6	10.5
Age, years						
20–39	38.2	0.70	0.41	0.52	39.6	9.5
40–59	39.9	0.80	0.46	0.60	37.7	10.2
60–80	21.9	0.74	0.41	0.55	42.8	12.9
Sex						
Male	48.6	0.74	0.42	0.55	39.8	10.7
Female	51.4	0.77	0.44	0.57	39.4	10.4
Race						
White	71.3	0.79	0.46	0.59	38.2	8.5
Black	11.1	0.74	0.40	0.53	43.8	13.9
Mexican American	12.3	0.68	0.38	0.49	45.0	19.1
Other	5.3	0.79	0.44	0.59	37.4	10.2
Education						
0–8 years	5.9	0.65	0.37	0.46	43.8	21.9
9–11 years	12.1	0.64	0.34	0.46	45.6	18.3
12 years	25.0	0.74	0.42	0.55	41.3	12.7
Some college	30.6	0.76	0.42	0.55	39.5	9.2
Bachelor's degree or higher	26.3	0.80	0.48	0.60	34.3	3.8
Any recent illness						
No	74.9	0.75	0.43	0.56	39.3	9.6
Yes	25.1	0.74	0.43	0.56	40.7	13.4
No. of days sick						
0	44.7	0.70	0.41	0.52	39.4	8.4
1–7	29.6	0.73	0.44	0.55	38.6	9.0
8–20	12.8	0.73	0.40	0.53	42.8	14.0
21–60	12.8	0.77	0.43	0.57	39.8	18.0
Time between interviews						
1 month or less	89.6	0.75	0.43	0.56	39.3	10.3
2 months[a]	10.4	0.77	0.43	0.57	40.2	10.3
SRH at household interview[b]						
Excellent	17.2				51.8	1.8
Very good	31.8				36.8	2.7
Good	34.1				34.2	11.1
Fair	13.7				42.6	37.8
Poor	3.2				45.6	12.4
MEC rating compared with household interview						
No change	60.4					
Improved/better	16.7					
Worsened	22.9					

Abbreviations: MEC, mobile examination center; NHANES, National Health and Nutrition Examination Survey; SRH, self-rated health.

[a] A delay of 3 months between interviews was experienced by 0.2% of respondents; they are included here.

[b] Measures of agreement and association are not defined for individual categories of self-rated health because they require variance in the distribution of both sets of ratings.

Table 2. Health Ratings[a] at Each Interview, for Total Sample and by Household-Interview Ratings of US Adults Aged 20–80 Years, NHANES, 2005–2008 ($n = 9,235$)[b]

Health Ratings	Household Interview, %	MEC Interview, %	Distribution of SRH Responses in the MEC Interview at Each Level of Household-Interview SRH				
			Excellent, %	Very Good, %	Good, %	Fair, %	Poor, %
Excellent	17.2	11.4	48.2	6.9	2.4	0.8	0.5
Very good	31.8	34.3	37.9	63.2	20.7	4.2	1.3
Good	34.1	38.0	12.1	27.2	65.8	32.8	10.7
Fair	13.7	13.7	1.7	2.7	10.6	57.4	33.2
Poor	3.2	2.6	0.0	0.0	0.5	4.8	54.4
	100.0	100.0	99.9	100.0	100.0	100.0	100.1

Abbreviations: MEC, mobile examination center; NHANES, National Health and Nutrition Examination Survey; SRH, self-rated health.

[a] Adjusted for sampling design.

[b] Values may not total "100" because of rounding.

also significantly less likely to "upgrade" his/her rating, compared with those who experienced no salient health events.

We also found substantial sociodemographic differences in reliability. Minority adults and those with less education had significantly lower test-retest reliability than whites and those with more schooling. Given the pervasiveness of SRH in literature on health inequalities, these group differences are important. They may, for instance, help to explain why SRH is a weaker predictor of mortality for those with lower socioeconomic status in the United States (4) and for blacks

Table 3. Correlates of Change[a] in Health Ratings Across 2 Interviews of US Adults Aged 20–80 Years, NHANES, 2005–2008 ($n = 9,235$)[b]

	Change		Multinomial Model of Better/Worse Rating			
			Better		Worse	
	Exponentiated Coefficients	95% CI	Exponentiated Coefficients	95% CI	Exponentiated Coefficients	95% CI
Age, years						
20–39	1.01	0.90, 1.13	1.21*	1.04, 1.41	0.83*	0.71, 0.98
60–80	1.18*	1.01, 1.39	1.11	0.92, 1.34	1.22	0.99, 1.50
Female sex	1.00	0.91, 1.10	1.00	0.89, 1.11	1.02	0.90, 1.15
Education	0.89**	0.84, 0.94	1.07	0.97, 1.18	0.76**	0.72, 0.82
Race						
Black	1.26**	1.12, 1.41	0.89	0.72, 1.10	1.68**	1.40, 2.02
Mexican	1.25**	1.11, 1.41	0.89	0.74, 1.06	1.75**	1.43, 2.16
Other	1.03	0.83, 1.27	0.74	0.49, 1.10	1.33*	1.03, 1.72
Any illness	1.07	0.94, 1.22	0.93	0.78, 1.10	1.20	0.99, 1.46
Days sick	1.00	1.00, 1.00	0.97**	0.97, 0.98	1.03**	1.02, 1.04
Time between interviews						
1–2 months	1.04	0.90, 1.20	1.31*	1.06, 1.62	0.83	0.67, 1.03
Missing	1.31	0.97, 1.78	1.15	0.74, 1.78	1.39	0.96, 2.02
Health rating at first interview						
Excellent	2.28**	1.86, 2.81	0.00**	0.00, 0.00	10.58**	8.64, 12.95
Very good	1.21*	1.01, 1.46	0.28**	0.23, 0.33	4.01**	3.30, 4.88
Fair	1.28**	1.12, 1.47	2.39**	2.04, 2.79	0.26**	0.19, 0.37
Poor	1.38*	1.06, 1.79	4.56**	3.42, 6.07	0.00**	0.00, 0.00

Abbreviations: CI, confidence interval; NHANES, National Health and Nutrition Examination Survey.

* $P < 0.05$; ** $P < 0.001$.

[a] Adjusted for sampling design.

[b] Reference levels for categorical variables are age 40–59 years, male, non-Hispanic white, no illness during the 30 days prior to second interview, less than 1 month between interviews, and "good" health at the baseline.

compared with whites (22, 23): With more measurement error, the estimates of the effect of SRH may be more biased toward zero for these disadvantaged groups.

Systematic differences in rating reproducibility across subgroups could arise from greater error or uncertainty about actual health status for some groups compared with others. The unequal reliability could also be caused by differential underlying distribution of health (8) if population groups vary in their actual health status and reliability is lower at some levels of health than others. In general, such reliability differentials could be problematic for the measurement and explanation of health inequalities among American adults, where researchers often use SRH as a measure of general health.

To the best of our knowledge, this is the first study to test the reliability of general SRH in the US population. Despite the importance of understanding the reliability of the SRH measure, only 2 studies have examined this issue in detail, both using non-US samples (8, 9). Crossley and Kennedy (8) found that, among Australian respondents, 28% changed their health rating within the same interview (before and after a detailed battery of health questions). Their data thus indicate a greater stability of SRH compared with our sample, in which nearly 40% of respondents changed their ratings (albeit over the course of about 1 month). Lundberg and Manderbacka (9) also reported considerably higher consistency in a Swedish sample, with fewer than 15% of respondents changing their ratings over 1 month's timeframe. They, however, used a 3-category SRH measure, and it is known that the number of levels in a variable influences its reliability (13). Taking this into account, we found that their results are not far from our results for *dichotomized* SRH, for which fewer than 11% of respondents would have changed categories between interviews. Also consistent with our findings for the United States, Crossley and Kennedy (8) found that the propensity to change ratings varied by sociodemographic characteristics, with older and lower-income respondents more likely to revise their ratings.

In contrast to the paucity of studies analyzing the short-term consistency of SRH, several studies have examined SRH stability over longer periods of time. In a 1973 study of the US elderly, only 34% changed their health during 15 years of the study — although the sample was likely to suffer from selection bias (24). Among Norwegian adolescents, 41% changed their health evaluation over a 4-year period (25), closely in line with findings for young Australian women, 37% of whom changed their health rating over a period of the same duration (26). It is somewhat surprising that the respondents in the current study changed health categories in 1 month, with a frequency comparable to what occurred over several years in these studies. Perhaps this is because the long-term studies tapped into the stability of health ratings due to both relatively unchanging underlying health and fixed reporting styles. In contrast, short-term test-retest reliability might capture random error associated with any individual health rating, such that the second response reflects a new draw from the measurement error distribution that can lead to a different categorization.

It is also possible that the different interview settings for the 2 interviews contributed to the variability in the respondents' health judgments in our study. Although the 2 SRH ratings were elicited with identical wording and mode of administration via face-to-face interviews, the setting was changed. The first rating was given during a household interview, but the second one occurred as a part of a medical examination in a mobile examination center. The ''medical'' environment of the second examination interview may have framed the rating (8), perhaps by making existing health problems more salient to the respondents, prompting them to downgrade their health status. This possibility is consistent with the finding that, among respondents who changed their rating, more downgraded (23%) than upgraded (17%).

The medical framing hypothesis also has some support from the fact that minority and less educated respondents, who tend to have more health conditions and problems, were more likely to downgrade their health between the 2 waves, perhaps reflecting the increased attention brought to their health during the medical examination. Our data do not allow us to examine the framing hypothesis in more detail, but future studies could randomize the SRH question placement within surveys, which would help to establish the strength and nature of such effects.

Whatever the source, our findings imply substantial measurement error in self-rated health among US adults, particularly for those with less education and nonwhite race/ethnicity. This error could have important implications for a wide range of epidemiologic and social science research that utilizes SRH as a summary measure of overall health, whether as a predictor or an outcome. When a covariate in multivariate analysis is measured with error, its effect is biased and inconsistent; moreover, the effects of other predictors may become biased as well (7, 27). Dichotomization of the 5-point scale may be a useful strategy for increasing the reliability of SRH in the general population, although it may not decrease the group differentials in consistency. Future research using self-rated health should consider using one of several available procedures (27) aimed at correcting for measurement error in linear and nonlinear regression models.

ACKNOWLEDGMENTS

Author affiliations: Department of Sociology, College of Arts and Sciences, University of Wyoming, Laramie, Wyoming (Anna Zajacova); Department of Epidemiology and Biostatistics, Hunter College, School of Public Health, City University of New York (CUNY), New York, New York (Jennifer Beam Dowd); and CUNY Institute for Demographic Research, New York, New York (Jennifer Beam Dowd).

This project was partly supported by award P20 RR16474-10 from the National Center for Research Resources.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. *Am J Public Health*. 1982;72(8):800–808.
2. Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. *J Health Soc Behav.* 1997;38(1):21–37.
3. Idler EL, Russell LB, Davis D. Survival, functional limitations, and self-rated health in the NHANES I Epidemiologic Follow-up Study, 1992. First National Health and Nutrition Examination Survey. *Am J Epidemiol*. 2000;152(9):874–883.
4. Dowd JB, Zajacova A. Does the predictive power of self-rated health for subsequent mortality risk vary by socioeconomic status in the US? *Int J Epidemiol*. 2007;36(6):1214–1221.
5. Finch BK, Hummer RA, Reindl M, et al. Validity of self-rated health among Latino(a)s. *Am J Epidemiol*. 2002;155(8):755–759.
6. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires*. New York, NY: Oxford University Press; 2006.
7. Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ*. 1996;313(7048):41–42.
8. Crossley TF, Kennedy S. The reliability of self-assessed health status. *J Health Econ*. 2002;21(4):643–658.
9. Lundberg O, Manderbacka K. Assessing reliability of a measure of self-rated health. *Scand J Soc Med*. 1996;24(3):218–224.
10. Salomon JA, Tandon A, Murray CJ. Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes [electronic article]. *BMJ*. 2004;328(7434):258. (doi:10.1136/bmj.37963.691632.44).
11. *National Health and Nutrition Examination Survey Data 2007–2008*. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention, US Department of Health and Human Services; 2010.
12. *Unweighted response rates for NHANES 2005–2006 and 2007–2008 by age and gender.* Hyattville, MD: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2010. (http://www.cdc.gov/nchs/nhanes/response_rates_cps.htm). (Accessed April 7, 2011).
13. Cicchetti DV, Shoinralter D, Tyrer PJ. The effect of number of rating scale categories on levels of interrater reliability: a Monte Carlo investigation. *Appl Psych Meas*. 1985;9(1):31–36.
14. Browning CR, Cagney KA. Moving beyond poverty: neighborhood structure, social processes, and health. *J Health Soc Behav.* 2003;44(4):552–571.
15. Lynch SM. Cohort and life-course patterns in the relationship between education and health: a hierarchical approach. *Demography.* 2003;40(2):309–331.
16. Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*. 1979;44(4):443–460.
17. Vanbelle S, Albert A. A note on the linearly weighted kappa coefficient for ordinal scales. *Stat Methodol*. 2009;6(2):157–163.
18. Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. *J Clin Epidemiol*. 1993;46(9):1055–1062.
19. STATA statistical software, release 11.0. College Station, TX: StataCorp LP, 2009.
20. Beaton DE, Bombardier C, Hogg-Johnson SA. Measuring health in injured workers: a cross-sectional comparison of five generic health status instruments in workers with musculoskeletal injuries. *Am J Ind Med*. 1996;29(6):618–631.
21. Seymour J, McNamee P, Scott A, et al. Shedding new light onto the ceiling and floor? A quantile regression approach to compare EQ-5D and SF-6D responses. *Health Econ*. 2010;19(6):683–696.
22. Lee SJ, Moody-Ayers SY, Landefeld CS, et al. The relationship between self-rated health and mortality in older black and white Americans. *J Am Geriatr Soc.* 2007;55(10):1624–1629.
23. Ferraro KF, Kelley-Moore JA. Self-rated health and mortality among black and white adults: examining the dynamic evaluation thesis. *J Gerontol B Psychol Sci Soc Sci.* 2001;56(4):S195–S205.
24. Maddox GL, Douglass EB. Self-assessment of health: a longitudinal study of elderly subjects. *J Health Soc Behav.* 1973;14(1):87–93.
25. Breidablik HJ, Meland E, Lydersen S. Self-rated health during adolescence: stability and predictors of change (Young-HUNT Study, Norway). *Eur J Public Health*. 2009;19(1):73–78.
26. Shadbolt B. Some correlates of self-rated health for Australian women. *Am J Public Health*. 1997;87(6):951–956.
27. Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2006.