AMERICAN SOCIOLOGICAL ASSOCIATION

Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health

Journal of Health and Social Behavior
52(2) 246–261

DOI: 10.1177/0022146510396713
http://jhsb.sagepub.com
SAGE

Hanna Grol-Prokopczyk[1], Jeremy Freese[2], and Robert M. Hauser[1]

Abstract

This article addresses a potentially serious problem with the widely used self-rated health (SRH) survey item: that different groups have systematically different ways of using the item's response categories. Analyses based on unadjusted SRH may thus yield misleading results. The authors evaluate anchoring vignettes as a possible solution to this problem. Using vignettes specifically designed to calibrate the SRH item and data from the Wisconsin Longitudinal Study (WLS; *N* = 2,625), the authors show how demographic and health-related factors, including sex and education, predict differences in rating styles. Such differences, when not adjusted for statistically, may be sufficiently large to lead to mistakes in rank orderings of groups. In the present sample, unadjusted models show that women have better SRH than men, but this difference disappears in models adjusting for women's greater health-optimism. Anchoring vignettes appear a promising tool for improving intergroup comparability of SRH.

Keywords

differential item functioning (DIF), gender and health, health disparities, reporting heterogeneity, survey methods

The general self-rated health (SRH) question—"In general, would you say your health is excellent, very good, good, fair, or poor?" or some minor variant thereof—is an extremely common survey item, both in the United States and internationally. The item has been shown to provide a good summary of overall physical health (e.g., Frankenberg and Jones 2004; Jylhä, Volpato, and Guralnik 2006); to predict respondents' mortality, even after controlling for known risk factors (e.g., DeSalvo et al. 2006; Idler and Benyamini 1997); and to predict functional ability among survivors, net of baseline health and socioeconomic variables (Idler and Kasl 1995).

However, accumulating evidence suggests a potentially serious problem with SRH, namely, that different groups use its response categories ("excellent," "very good," etc.) in different ways. This article assesses a recently developed survey method, anchoring vignettes, as a means of correcting for this problem. Our results indicate that anchoring vignettes are a promising tool for improving intergroup comparability of SRH.

GROUP DIFFERENCES IN HEALTH-RATING STYLE

Banks et al. (2007) compare American and English men's health and find a puzzling contradiction: Based on self-reports of disease or biological measures, American men have objectively worse health than Englishmen, but on the SRH question, they report *better* health. After ruling out other explanations, the authors conclude that this "contradiction most likely stems from different thresholds used by Americans and English. . . . For the same 'objective'

[1]University of Wisconsin-Madison, Madison, WI, USA
[2]Northwestern University, Evanston, IL, USA

Corresponding Author:
Hanna Grol-Prokopczyk, University of Wisconsin-Madison, Department of Sociology, 8128 Sewell Social Sciences Building, 1180 Observatory Drive, Madison, WI 53706, USA
Email: hgrol@ssc.wisc.edu

health status, Americans are much more likely to say their health is good" (p. 28). That is, American men appear more "health-optimistic" (Ferraro 1980:381) than Englishmen. Similar evidence of differential use of SRH's response categories is found across Asian countries (Zimmer et al. 2000), European countries (e.g., Jürges 2007; Jylhä et al. 1998; Murray et al. 2002), racial/ethnic groups (e.g., Menec, Shooshtari, and Lambert 2007; Shetterly et al. 1996), socioeconomic strata (e.g., Dowd and Zajacova 2007), and age groups (e.g., Ferraro 1980; Groot 2000; Idler 1993).

Men and women, too, may vary in health-optimism. It has been amply demonstrated that despite lower mortality rates at most ages, women report "more intense, more numerous, and more frequent" physical health problems than men across the life course (e.g., Barsky, Peekna, and Borus 2001:266); some studies find that "most physical symptoms are typically reported at least 50 percent more often by women" (Kroenke and Spitzer 1998:150). While at young and middle ages SRH scores are consistent with women's greater number of health problems, in later life (roughly age 60), this pattern disappears or reverses (Case and Paxson 2005). That is, among older adults, women's SRH appears statistically equivalent to men's (Benyamini, Leventhal, and Leventhal 2000:357; Fillenbaum 1979:47; Frankenberg and Jones 2004:444), or more positive than men's (Ferraro 1980:380–81), despite women's greater experience of somatic symptoms. This is the case in the 2005 Wisconsin Longitudinal Study (WLS), in which women give slightly higher health self-ratings than men,[1] even while reporting significantly more health problems (Hauser and Roan 2006:74–75). Such data suggest that in older populations, women may be more health-optimistic than men.

Despite such discrepancies between objective health conditions and subjective health ratings, some researchers argue against "systematic sex differences in [health-]reporting behavior," even claiming that such differences have "tak[en] on the character of an urban folk tale" (Macintyre, Ford, and Hunt 1999:91). Accurately evaluating such claims, however, requires theoretical clarity about the concept of "health-reporting behavior." Three meanings of the term—based on differences in conceptualization of health, respondent thoroughness, and use of response categories, respectively—are often conflated in current use. First, groups may have different health-reporting styles because they differ in their meaning of "health"; for example, in whether mental health is considered part of overall health. Though evidence is mixed, studies often find "no significant differences in the frame of reference used by males and females to answer the global health status question" (Krause and Jay 1994:937), nor sex differences in considering "'trivial' or mental health conditions" (Macintyre et al. 1999:89). (Some scholars, however, suggest that men's health ratings are more sensitive than are women's to life-threatening diseases such as heart disease, as opposed to non-life-threatening conditions such as arthritis; e.g., Benyamini et al. 2000; Deeg and Kriegsman 2003:383.) Second, some groups may give less accurate self-reports of health due to lack of self-knowledge or disinterest in survey participation; for example, men might give higher self-ratings than warranted because they do not know, remember, or care to reflect upon their medical problems. Empirical evidence, however, argues against this (Macintyre et al. 1999; Verbrugge 1989). Third, as described earlier, groups may differ in their use of response categories, that is, in where along the health spectrum they locate thresholds between "poor" and "fair," "fair" and "good," and so on (Figure 1, left). This phenomenon—termed "response category differential item functioning," or DIF (King et al. 2004)—is the focus of this article, and subsequent mentions of "health-rating style" will refer to this. Macintyre et al.'s (1999) dismissal of sex differences in rating style as an "urban folk-tale," we note, was based on evidence relating to the first two aforementioned categories; DIF was not addressed.

Response category DIF is generally deduced by process of elimination, namely, by identifying discrepancies in SRH that persist when relatively objective health measures are controlled for. Most commonly, SRH scores are regressed on large numbers of health-related, demographic, and/or behavioral variables in an attempt to make sex (or other group) differences "disappear." Failure to achieve this goal is considered indicative of DIF.

This residual approach to identifying DIF has several shortcomings, however. It is prone to Type I error if sufficient controls are lacking (e.g., disease severities) and to Type II error due to possible suppression effects if controls are cherry-picked to remove evidence of DIF. Furthermore, the approach may be unrealizable when costs make extensive health questionnaires or biomarker collection impossible, or when groups being compared differ in their disease taxonomies or access to disease diagnoses. Finally, even if the residual regression approach is both doable and correct in identifying DIF, it does not suggest any clear

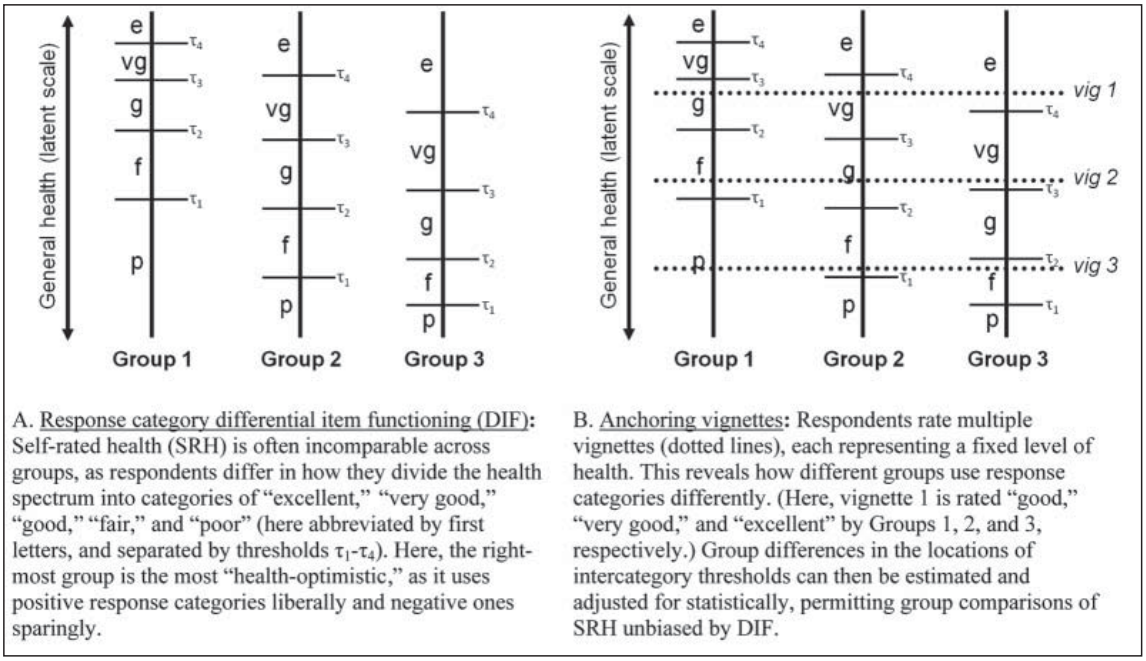

Figure 1. Schematic Diagram of Logic Underlying the Anchoring Vignette Method

method for overcoming DIF in subsequent analyses. Some authors suggest doing separate analyses by subgroup (Ferraro 1980:381), but this approach is limited if response style varies across overlapping subgroups, and of course, group comparison is often the goal of analyses. Thus, most authors finding evidence of DIF can do little but helplessly list it as a potential source of error and warn against direct group comparisons.

To summarize, there is evidence (even if indirect) that the demographic categories of greatest interest to health researchers—nationality, race/ethnicity, socioeconomic status, age, and sex—are subject to response category DIF in the context of SRH, a fact threatening the correctness of research findings relying on SRH. (Multilingual surveys may also be subject to DIF triggered by language differences.) Conceptual and methodological challenges have made it somewhat difficult to identify DIF in SRH with confidence, and even more difficult to adjust for DIF statistically. In what follows, we investigate a technique with potential to help overcome such problems by directly measuring and adjusting for DIF: anchoring vignettes.

ANCHORING VIGNETTES

Whenever surveys use subjective ordered response categories, group differences in responses potentially reflect response category DIF rather than differences in the actual variable of interest. Figure 1 (left half) presents a hypothetical example of groups differing in how they divide the health spectrum into categories of "excellent," "very good," and so on. Group 1, relatively sparing in its use of positive categories such as "excellent," is comparatively "health-pessimistic," while the opposite holds for Group 3. In such a scenario, groups may use the same response category while actually referring to very different underlying levels of health. Generally, researchers have no direct information about intercategory thresholds (τ), and so have no way of knowing whether one group's "good" is higher, lower, broader, or narrower than another's.

While various techniques have been proposed for establishing comparable response scales across groups, recent reviews describe *anchoring vignettes* as "the most promising" of available strategies (e.g., Murray et al. 2002:429). Anchoring vignettes are brief texts depicting hypothetical individuals who manifest the trait of interest (e.g., health) to a lesser or greater degree. Respondents rate each character on the same scale as their own self-rating. Typically respondents rate several vignettes, representing various levels of the trait. These ratings reveal what different groups mean by response categories such as "good." Figure 1, right half, presents this logic visually: The level of health represented by vignette 1 is rated "good" by Group 1, "very good" by Group 2, and "excellent"

by Group 3, revealing the groups' different health-rating styles. Additional vignettes provide comparable information elsewhere along the health spectrum.

Anchoring vignettes, in short, reveal DIF. Phrased more formally, vignettes can be used to estimate where on the latent spectrum groups locate the thresholds between response categories (τ_1 to τ_4 in Figure 1). These threshold differences can then be adjusted for statistically, allowing for valid intergroup comparisons of self-ratings, unbiased by DIF. While anchoring vignettes do not address *why* there are group differences in rating styles, they can demonstrate, quantify, and adjust for such differences. (For additional information, see King and Wand 2007; King et al. 2004.)

The primary measurement assumptions of the anchoring vignette method are *response consistency* and *vignette equivalence* (King et al. 2004:194). Response consistency means that respondents use response categories the same way when rating vignettes as when rating themselves (rather than holding themselves to higher or lower standards than vignette characters). Vignette equivalence means that all respondents perceive a vignette as representing the same underlying concept, with vignettes in a series all seen as part of a unidimensional scale.

Anchoring vignettes appear in a growing number of surveys worldwide (e.g., the 70-country World Health Survey) and have been applied to a wide variety of research areas, including political efficacy, job satisfaction, women's autonomy, and specific domains of health (e.g., mobility and vision) (Hopkins and King 2010; cf. Anchoring Vignettes Web site: http://gking.harvard.edu/vign/). However, thus far, anchoring vignettes have not been applied to the general self-rated health question, despite the widespread use of SRH and clear indications that DIF is an issue in analyses using SRH. Some originators of the vignette method express skepticism that vignettes could be used to calibrate SRH, given the complexity of overall physical health (King 2005). In what follows, we test this directly.

ANALYTIC GOALS

In this article we create and evaluate anchoring vignettes that calibrate the general SRH item. Specifically:

1. We create three series of general health anchoring vignettes and test whether they meet the assumptions of vignette equivalence and response consistency.
2. We assess whether demographic and health-related variables affect vignette ratings, that is, whether they are associated with DIF. (If there is no DIF, there is no need to proceed further, as unadjusted SRH will be unbiased and comparable among groups.) We test whether women are more health-optimistic than men, whether mention of specific diseases affects men's vignette ratings more than women's, and whether personal experience with a disease affects respondents' ratings of vignettes mentioning that disease.
3. We compare a standard analysis of predictors of SRH with an analysis that statistically accounts for DIF to see how DIF affects the strength and/or direction of coefficients. We attend closely to sex differences to see if vignette-based adjustments resolve the aforementioned paradox of women's greater number of physical ailments but higher SRH.

DATA AND METHOD

Data

The Wisconsin Longitudinal Study began in 1957 as a one-third random sample (N = 10,317) of graduating Wisconsin high schools seniors, and expanded in subsequent waves to include a randomly selected sibling of each graduate ("siblings") and the sibling's spouse ("sibling-spouses"). Our analyses are based on a random subset of siblings (N = 1,221) and sibling-spouses (N = 1,404) surveyed by telephone in 2005-2007, yielding a sample size of 2,625. Because siblings, but not spouses, were also administered a mail survey containing health-related information, some analyses are conducted with siblings only. A primary limitation of the data is that reflecting the demographics of Wisconsin high schools in 1957, 99 percent of respondents identify as exclusively white. (See www.ssc.wisc.edu/wlsresearch/ for WLS documentation and data.)

Table 1 presents descriptive statistics for the analytic sample as well as descriptions of our independent variables.

Table 1. Descriptive Statistics for Analytic Sample

	Proportion or Mean	Standard Deviation	*N*
Female	.55		2,625
Self-rated health (SRH); 1 = *poor* to 5 = *excellent*	3.67	.99	2,625
Age at time of interview, in years	63.79	7.73	2,624
Education			
Less than high school	.05		139
High school degree	.41		1,056
Some college	.19		497
Four-year college degree	.18		463
Postcollege education	.16		410
Household income, 2005 (in dollars)	74,979	121,265	2,609
Respondent ever diagnosed with diabetes/high blood sugar?	.16		2,620
Respondent ever diagnosed with heart problems?	.15		2,622
Respondent ever diagnosed with hypertension?	.48		2,622
Health Utilities Index (HUI-3) score: 0 = health-state equivalent to death, 1 = best health	.81	.22	2,625
Health Symptoms Scale (HSS) score:[a] Count of physical health symptoms (out of 25) experienced in past six months	8.88	5.11	999
Respondent's parent(s)/sib(s)/spouse had diabetes?[a]	.40		1,012
Respondent's parent(s)/sib(s)/spouse had heart attack?[a]	.47		1,012

[a]These items were administered on the Wisconsin Longitudinal Study (WLS) 2005 mail survey and are available only for sibling respondents, not for sibling-spouses.

Vignette Texts

We wrote three series of vignettes (Table 2): one describing health as daily functioning/disability and referring to no specific diseases (the "No Specific Disease" series), one supplementing the aforementioned with references to heart disease or related conditions (the "Heart Disease" series), and one supplementing the aforementioned with reference to diabetes or related conditions (the "Diabetes" series). These variations allowed us to test whether response consistency and/or substantive findings (especially about sex differences) are affected by inclusion of medical diagnoses in vignettes, to see whether personal experience with a medical condition affects ratings of characters with that condition, and to heed the call of contemporary scholars to treat health as involving daily, lived well-being, rather than being strictly synonymous with mortality risk (e.g., Murray and Chen 1992).

Each series consisted of four vignettes of varying severity. Symptoms described in vignettes represent typical health variations among WLS participants at different levels of SRH. Heart Disease and Diabetes vignettes were formed by adding a disease-specific sentence to the corresponding No Specific Disease vignette. Table 2 shows both vignette texts and instructions, which encouraged respondents to rate vignette characters just as they would rate themselves and to consider them age peers. To further encourage response consistency, vignette characters' sex was matched to respondents' sex; first names used (Nancy, Joan, and Karen for women; David, Tom, and William for men) were drawn from the 10 most common names among respondents; and the question following each vignette exactly replicated the SRH question's wording ("In general, would you say [character]'s health is: excellent, very good, good, fair, or poor?").

For ease of interpretation, SRH and vignette ratings were reverse-coded so higher values indicate better health (1 = *poor*, 5 = *excellent*). Each respondent received three vignettes—one from each series—representing three different severity levels. The order of the series and assignment of severity levels to each series were randomly determined.

Table 2. Text of General Health Vignettes

Introductory text	Earlier we asked you to rate your own health overall. We are interested in how you would use these same categories to rate the health of other people your age. Now I am going to describe the health of some people your age; then I am going to ask you to rate their health using the same categories you used to rate your own health.
No Disease series	*These also serve as base texts for the Heart Disease and Diabetes series.*
Severity 1	[Name/she/he] is energetic, and has little trouble with bending, lifting, and climbing stairs. [She/he] rarely experiences pain, except for minor headaches. In the past year [Name/she/he] spent one day in bed due to illness.
Severity 2	[Name/she/he] is usually energetic, but occasionally feels fatigued. [He/she] has some trouble bending, lifting, and climbing stairs. [His/her] occasional pain does not affect [his/her] daily activities. In the past year, [Name/she/he] spent a few days in bed due to illness.
Severity 3	About once a week, [Name/she/he] has no energy. [He/she] has some trouble bending, lifting, and climbing stairs, and each week experiences pain that limits some of [his/her] daily activities. In the past year, [Name/she/he] spent a week in bed due to illness.
Severity 4	[Name/she/he] feels exhausted several days a week. [He/she] has trouble bending, lifting, and climbing stairs, and every day experiences pain that limits many of [his/her] daily activities. In the past year, [Name/she/he] spent a few nights in a hospital, and over a week in bed due to illness.
Heart Disease series	*The following sentences are added to the base text from the No Disease series.*
Severity 1	[Name]'s doctor says [Name] has good blood pressure, and that [his/her] heart is in good health.
Severity 2	[Name]'s doctor says [Name] has borderline high blood pressure and high cholesterol, but does not need medication for them.
Severity 3	[Name] has high blood pressure and high cholesterol. [He/she] once underwent angioplasty to unblock an artery, and takes medication for these problems.
Severity 4	[Name] has very high blood pressure and cholesterol. [He/she] once had a heart attack, and subsequently had successful bypass surgery.
Diabetes series	*The following sentences are added to the base text from the No Disease series.*
Severity 1	[Name]'s doctor says [Name] has healthy blood sugar levels.
Severity 2	[Name]'s doctor says [Name] must lower [his/her] blood sugar levels to avoid getting diabetes.
Severity 3	[Name] has diabetes, and controls it by managing [his/her] diet.
Severity 4	[Name] has diabetes that requires [him/her] to take daily insulin injections, and is experiencing some diabetes-related complications.
Question following each vignette	In general, would you say [Name]'s health is: excellent, very good, good, fair, or poor?

Analytic Models

Vignette equivalence predicts that rankings of vignettes in a series will be consistent across respondents. To test this assumption, we measured violations of intended rank orderings of vignettes (King et al. 2004). To test response consistency, we regressed SRH on vignette ratings while controlling for (relatively) objective measures of overall health, to confirm that more optimistic self-raters are also more optimistic vignette raters.

To identify factors predicting differences in vignette ratings, we estimated two ordered probit models: one including basic demographic variables, and one adding personal and familial health variables. Finally, to assess how accounting for DIF affects apparent predictors of SRH, we compared (a) a standard ordered probit regression of SRH on various independent variables to (b) a joint "hopit" regression for SRH and vignette ratings on the same independent variables.[2,3] Hopit, short for "hierarchical ordinal probit," uses respondents' ratings of vignettes to rescale the thresholds of the standard ordered probit model, revealing how self-assessments differ among

Table 3. Mean Ratings of General Health Vignettes

Series	Least Severe	2	3	Most Severe
No Specific Disease	4.04	2.78	2.06	1.59
(*n* = 2,623)	(.91)	(.78)	(.77)	(.62)
Heart Disease	4.19	2.86	1.63	1.32
(*n* = 2,621)	(.82)	(.81)	(.68)	(.51)
Diabetes	4.03	2.50	1.98	1.41
(*n* = 2,620)	(.92)	(.77)	(.70)	(.59)

Note: Means calculated by assigning scores to responses of 1 = *poor*, 2 = *fair*, 3 = *good*, 4 = *very good*, 5 = *excellent*. Standard deviations in parentheses. Fewer than .3 percent of respondents answered "don't know" or "refused"; these are excluded from analyses.

groups after differences in rating styles are accounted for (Rabe-Hesketh and Skrondal 2002; cf. King et al. 2004). See Appendix A (online supplement available at http://jhsb.sagepub.com/supplemental) for formal specifications. We examined how coefficients changed in sign and statistical significance between the ordered probit and hopit models.

RESULTS

Adherence to Measurement Assumptions

Table 3 shows that within and across each disease series, mean vignette ratings display the expected ordinality when moving from the least to most severe vignette. The smaller standard deviations for Severity 4 vignettes (.51 to .62 vs. .68 to .92 for other severities) suggest a floor effect of response categories. Among individual respondents, fewer than 9 percent gave ratings that violated the intended rank ordering of vignettes by severity (data not shown). These results, showing little evidence of multidimensionality, are consistent with the first assumption of the anchoring vignette method, vignette equivalence.

The model in Table 4 tests adherence to the second key assumption of the method, response consistency, which asserts that respondents use the same standards to rate themselves as to rate vignettes. Response consistency predicts that if two respondents have the same objective level of health but nonetheless give different self-ratings, the difference in self-ratings should be positively correlated with the difference in respondents' vignette ratings. That is, the more optimistic self-rater should also be the more optimistic vignette rater.[4] To test this, we performed ordered probit regressions of SRH on vignette ratings with two more objective self-report measures of general health as controls: the Health Utilities Index Mark 3 (HUI-3) score and a count of physical symptoms (the Health Symptoms Scale; HSS).

Results (Table 4) show a strong association between physical health measures (both HSS and HUI) and SRH ($p < .001$ in all three series). More importantly for our purposes, vignette ratings are positively and significantly associated with self-ratings in all series (β between .137 and .186; $p < .001$).[5] That is, greater health-optimism in vignette ratings indeed predicts greater health-optimism in self-ratings, providing evidence of response consistency. Our vignettes thus show no major violations of the key assumptions of the anchoring vignette method, and so may serve to answer substantive questions about group differences in health-rating style.

Differences in Health-Rating Styles

Table 5 presents estimates from ordered probit regressions of vignette ratings on sociodemographic variables, and shows that certain basic demographic variables are indeed associated with DIF.[6] In all three series, women give higher ratings than men, a difference both statistically significant and not trivial in size (β ranging from .224 to .371; $p < .001$). This is evidence that women are more health-optimistic than are men. The magnitude of this difference may be conveyed by some simple comparisons: 48 percent of women, but only 34 percent of men, rated the Heart Disease Severity 1 character's health as "excellent." For Diabetes Severity 3, 17 percent of women selected "poor" and 24 percent selected "good"; comparable percentages for men were 33 percent and 13 percent, respectively. Only 40 percent of women, but 58 percent of men, rated the No Disease Severity 4 vignette as "poor." These examples of women's higher ratings are typical. The only vignettes not showing significant sex differences were Heart Disease Severity 4 and Diabetes Severity 4. It is unclear whether these exceptions

Table 4. Ordered Probit Regressions of Self-Reported Health on Vignette Ratings and Other Measures of Health Status

	No Specific Disease Series (n = 2,623)	Heart Disease Series (n = 2,621)	Diabetes Series (n = 2,620)
Vignette rating	.186***	.137***	.153***
	(.027)	(.030)	(.028)
Health Symptoms Scale score (÷ 10)	−.636***	−.627***	−.626***
	(.074)	(.074)	(.074)
Heath Utilities Index	2.251***	2.239***	2.244***
	(.125)	(.124)	(.125)

Note: Standard deviations in parentheses. Models also include controls for vignette severity. Where missing, Health Symptoms Score is imputed based on Health Utilities Index, to maintain sample size (this model only).
***p < .001, two-tailed.

Table 5. Ordered Probit Regressions of Vignette Rating on Demographic Variables

	No Specific Disease Series (n = 2,546)	Heart Disease Series (n = 2,546)	Diabetes Series (n = 2,543)
Female	.371***	.224***	.370***
	(.046)	(.047)	(.046)
Age (÷ 10)	−.069*	.008	−.057†
	(.031)	(.032)	(.031)
Less than high school	−.148	−.207†	−.189†
	(.104)	(.108)	(.104)
Some college	.172**	.097	.125*
	(.0610)	(.064)	(.062)
Four-year college degree or more	.242***	.181***	.265***
	(.053)	(.055)	(.054)
Household income, second quartile	.100	.112	.066
	(.067)	(.070)	(.068)
Household income, third quartile	−.033	.020	.025
	(.065)	(.068)	(.066)
Household income, fourth (top) quartile	.070	.062	−.004
	(.066)	(.069)	(.067)

Note: Standard errors in parentheses. Omitted reference categories: "high school degree" (for education) and "household income, bottom quartile" (for income). Models also include controls for vignette severity.
†p < .10. *p < .05. **p < .01. ***p < .001, two-tailed.

indicate that men and women's ratings converge when severe, specific diseases are mentioned or whether they are artifacts of category floor effects.

Relatedly, models interacting sex and series (not shown) find no evidence that men's ratings of health are affected more than women's by mention of specific health conditions. Indeed, women rated the Heart Disease Severity 4 vignette *more* negatively than men. Again, response truncation must be considered, but since this lone interaction effect was opposite the direction predicted by the aforementioned theory of sex differences, we conclude that our data do not support the theory. Further comparisons with differently worded vignettes may still be warranted, however, to test for other sources of multidimensionality.

Table 5 also shows a negative association between age and vignette ratings in the No Disease ($\beta = -.069$; $p < .05$) and Diabetes series ($\beta = -.057$; $p < .10$). The effect size is very small, but is at odds with previous literature (e.g., Groot 2000; Idler 1993), and suggests that respondents are not attending to instructions to treat vignette characters as age peers. This is discussed further in our treatment of Table 7.

Consistent with previous literature (e.g., Dowd and Zajacova 2007), higher levels of education pre-

Table 6. Ordered Probit Regressions of Vignette Rating on Demographic and Health-Related Variables

	No Specific Disease Series (n = 942)	Heart Disease Series (n = 942)	Diabetes Series (n = 938)
Female	.412***	.250**	.401***
	(.078)	(.081)	(.079)
Age (÷ 10)	−.073	.019	−.107†
	(.057)	(.060)	(.057)
Less than high school	−.117	−.080	−.301
	(.188)	(.194)	(.192)
Some college	.231*	.129	.068
	(.103)	(.107)	(.104)
Four-year college degree or more	.257**	.211*	.228**
	(.088)	(.091)	(.089)
Household income, second quartile	.130	.154	.041
	(.107)	(.111)	(.109)
Household income, third quartile	.000	.034	.030
	(.110)	(.114)	(.111)
Household income, fourth (top) quartile	.136	.236†	.090
	(.117)	(.121)	(.118)
Respondent's diabetes diagnosis	−.050	−.042	−.074
	(.106)	(.108)	(.106)
Respondent's heart problems diagnosis	.081	−.012	−.012
	(.112)	(.114)	(.114)
Respondent's hypertension diagnosis	.015	.167*	.140†
	(.076)	(.079)	(.078)
Parent/sibling/spouse had diabetes	−.055	−.060	.084
	(.076)	(.079)	(.077)
Parent/sibling/spouse had heart attack	.011	.143†	.039
	(.074)	(.077)	(.076)
Health Symptoms Scale score (÷10)	.144†	.093	.135
	(.080)	(.083)	(.082)
Health Utilities Index	−.259	−.077	.242
	(.188)	(.194)	(.195)

Note: Standard errors in parentheses. Models also include controls for vignette severity. Omitted reference categories: "high school degree" (for education) and "household income, bottom quartile" (for income).
†p < .10. *p < .05. **p < .01. ***p < .001, two-tailed.

dict more health-optimistic ratings, an effect that appears roughly linear. The effect of a college degree (compared to a high school degree) approaches the size of the difference between men and women, as shown by the relatively large parameter estimates (β between .181 and .265; p < .001). Perhaps more highly educated respondents feel greater confidence regarding their capacity to handle a given level of health impairment, and thus rate it more positively. Income, in contrast, is unrelated to ratings net of other variables (confirmed by a Wald test of the joint significance of the income dummies).

Our next model, including measures of first- and secondhand experience with specific health conditions, is shown in Table 6. We hypothesized that people with personal or familial experience of heart disease, diabetes, or related conditions might respond differently to disease-mentioning vignettes than those without such experience, even when controlling for overall health.

Our results bear out this hypothesis. Respondents with hypertension ranked Heart Disease vignettes significantly more positively than did respondents without hypertension (β = .167; p < .05). So too did respondents whose parents, siblings, or spouses had suffered heart attacks (β = .143; p = .06). This suggests that familiarity with heart-related conditions leads respondents to con-

Table 7. Ordered Probit and Hopit Regressions of Self-Rated Health (SRH) on Demographic Variables

	Ordered Probit		Hopit	
	β	*SE*	β	*SE*
Female	.173***	.044	−.050	.061
Age (÷ 10)	−.122***	.030	−.034	.042
Less than high school	−.248**	.097	−.174	.135
Some college	.144**	.059	.054	.082
Four-year college degree or more	.460***	.052	.309***	.073
Household income, second quartile	−.176**	.064	−.265**	.089
Household income, third quartile	.020	.063	.093	.088
Household income, fourth (top) quartile	.199**	.064	.177†	.091
Threshold 1 (poor–fair)				
Sex (female)			−.469***	.055
Age (÷ 10)			.026	.038
Less than high school			.100	.106
Some college			−.201**	.072
Four-year college degree or more			−.285***	.063
Household income, second quartile			−.107	.076
Household income, third quartile			−.110	.075
Household income, top quartile			−.156*	.077
Constant	−2.544***	.216	−2.138***	.355
Threshold 2 (fair–good)				
Sex (female)			.231***	.053
Age (÷ 10)			.009	.037
Less than high school			.056	.103
Some college			.042	.069
Four-year college degree or more			.097	.059
Household income, second quartile			.047	.077
Household income, third quartile			.181*	.074
Household income, top quartile			.136†	.076
Constant	−1.761***	.211	−.225	.265
Threshold 3 (good–very good)				
Sex (female)			−.048	.048
Age (÷ 10)			.045	.033
Less than high school			−.097	.111
Some college			.119†	.062
Four-year college degree or more			.091	.057
Household income, second quartile			−.047	.069
Household income, third quartile			.008	.067
Household income, top quartile			−.062	.070
Constant	−.754***	.210	−.300	.236
Threshold 4 (very good–excellent)				
Sex (female)			.101*	.048
Age (÷ 10)			.049	.032
Less than high school			−.083	.125
Some college			−.070	.063
Four-year college degree or more			−.098†	.054
Household income, second quartile			.007	.072
Household income, third quartile			−.050	.070
Household income, top quartile			.085	.067
Constant	−.350	.209	−.286	.225
Vignettes				
θ_1			.279	.294
θ_2			−1.061***	.295
θ_3			−1.849***	.296
θ_4			−2.477***	.298
ln σ			−.209***	.029

Note: N = 2,548. Hopit uses No Disease vignettes. Reference categories: "high school degree" (education), "household income, bottom quartile" (income).

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$, two-tailed.

sider them less problematic. It is surprising that respondent's own heart problems do not similarly predict higher Heart Disease ratings, but this could result from question wording: All four Heart Disease vignettes mention "blood pressure" (specifically "high blood pressure" in Severities 2 through 4), but only Severity 3 mentions "angioplasty," and only Severity 4 mentions a "heart attack." The Heart Disease series, then, might be more accurately seen as a Hypertension series. In bivariate analyses of individual Heart Disease vignettes, personal experience with heart problems predicts more positive ratings when angioplasty ($\beta = .282$; $p = .019$; $n = 672$) or heart attack ($\beta = .327$; $p = .017$; $n = 680$) are mentioned. We found no parallel evidence that experience with diabetes affects ratings of Diabetes vignettes. Perhaps in this case, awareness of the daily challenges of maintaining healthy blood sugar levels negates the optimism-producing "familiarity effect."

In addition to the models in Tables 5 and 6, we tested others including measures of personality, depression, and psychological well-being, but none of these showed systematic association with vignette ratings. However, in *all* models tested, sex was strongly and significantly related to vignette ratings in all series. The sex effect is thus the most robust finding from our analyses, and it is consistent with our suspicions, expressed in our introduction, that in this age group, women are more health-optimistic than are men.[7]

More generally, we have shown that there are significant differences in how different groups use response categories to rate general health. We next assess how this affects apparent differences in groups' SRH.

Group Differences in Self-Rated Health

The group differences in vignette-rating style, described above, imply the presence of those same group differences in *self*-rating style (assuming response consistency). How does taking such group differences into account affect analyses of SRH? To answer this, we compare two models: one involving no attempt to adjust for DIF (a standard ordered probit regression) and one that adjusts for DIF by rescaling groups' response category thresholds based on vignette ratings (hopit). Due to space restrictions, we show only findings based on the No Disease vignettes. Findings from the other series were extremely similar.

Table 7 presents our comparison of ordered probit and hopit models regressing SRH on demographic variables. In the ordered probit model, nearly all the independent variables significantly predict SRH. As mentioned earlier, women in this sample report better health than do men ($\beta = .173$; $p < .001$). Consistent with expectations, older respondents report worse health than do younger ones ($\beta = -.122$; $p < .001$), and education is positively and roughly linearly associated with better SRH (e.g., $\beta = .460$; $p < .001$ for college vs. high school degrees holders). The association of income with SRH is as expected aside from an inversion in the bottom two quartiles, which supplementary analyses indicate is accounted for in models adding measures of wealth (not shown); this reflects the fact that income is not an ideal measure of economic standing in a population with mixed retirement statuses.

Next, we look at how coefficients change in sign and statistical significance as we move from the ordered probit to the hopit model (Table 7, right). Perhaps most strikingly, the coefficient for female, which had been positive, now becomes negative (though not statistically significant: $\beta = -.050$; $p = .41$). In other words, the apparent better health of women disappears when health-rating style is accounted for. The puzzle of our female respondents' surprisingly high SRH appears, then, due at least in part to sex differences in response category thresholds.

Age remains negatively associated with SRH in the hopit model, though this effect ceases to be statistically significant ($\beta = -.034$; $p = .42$). The lack of a significant effect of age on SRH is surprising, though consistent with—indeed, caused by—the earlier finding that older respondents are more health-pessimistic than younger ones (and so have self-ratings adjusted upwards by hopit). Datta Gupta, Kristensen, and Pozzoli (2010), analyzing disability vignettes, report very similar findings, which they show result from age-related response *in*consistency—the failure of respondents to treat vignette characters as age peers. Our vignettes appear to suffer the same problem (a possibility supported by survey audio recordings in which respondents ask the vignette characters' ages).[8] The problem appears surmountable, however: A recent fielding of our vignettes to a nationally representative sample ($n = 1{,}765$) included more prominent instructions regarding characters' ages, and no negative correlations between age and health ratings were found (while all other major findings of the present study were replicated) (Grol-Prokopczyk 2010). We counsel future users of health vignettes to attend carefully to instrument wording to maximize age-related response consistency.

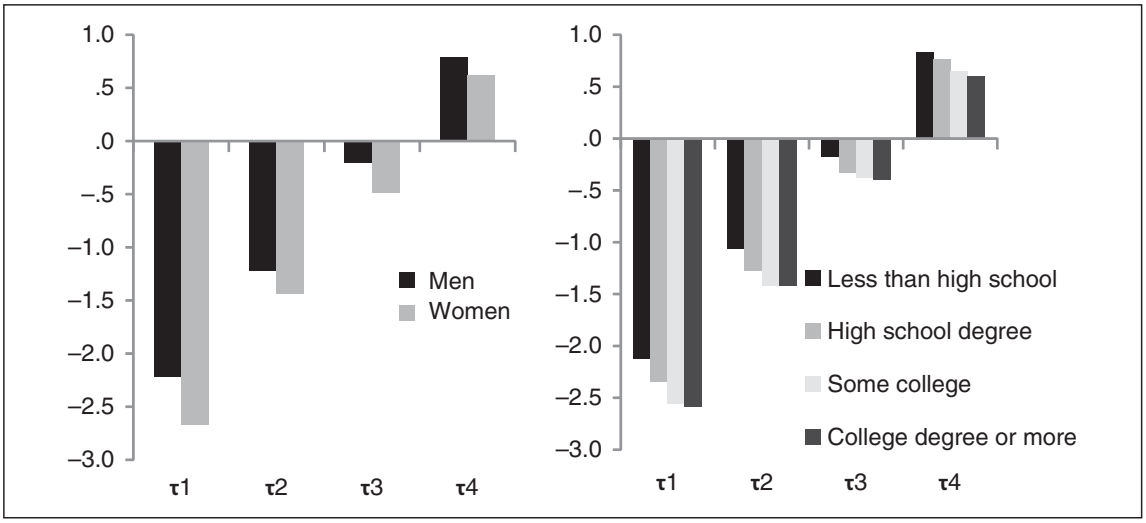

Figure 2. Mean Estimated Intercategory Thresholds, by Sex (Left) and Education (Right).
Note: τ_1 through τ_4 refer to thresholds "poor/fair," "fair/good," "good/very good," and "very good/excellent," respectively, as in Figure 1. Estimates are derived by applying threshold-predicting coefficients from the Hopit model in Table 7 to the analytic sample. Y-axis units are standard deviations of self-rated health (SRH).

Education continues to be positively associated with health in the hopit model, though the effect is weakened, with only the college degree variable remaining statistically significant ($\beta = .309$; $p < .001$). This reflects the hopit model's correction for the greater health-optimism of more highly educated respondents. In contrast, parameter estimates for income change little between the two models, since, as shown in Tables 5 and 6, income has no strong association with rating style.

The hopit model's information about predictors of threshold variation (also in Table 7) explains why findings differ between the probit and hopit models. For example, the hopit coefficient for female sex under Threshold 1 (–.469; $p < .001$) indicates that women have a lower threshold than men for the distinction between "poor" and "fair"; that is, women are more likely to choose "fair" over "poor" to describe a given vignette. Furthermore, since higher order thresholds depend on previous ones in hopit's parameterization (online Appendix A, Equation 1), this substantial sex difference in the lowest cutpoint sets the stage for sex-related difference in higher cutpoints.

Since coefficients for thresholds beyond the first are challenging to interpret directly (they both depend on previous thresholds and involve exponentiation of coefficients), group differences in thresholds are best presented visually. Figure 2 presents hopit's mean estimated thresholds for our sample by sex and by education. As shown, all four intercategory thresholds are noticeably lower for women than for men, reflecting our female respondents' greater health-optimism across the health spectrum. Similarly, cutpoints consistently decrease with rising education (albeit with small or no differences between "some college" and "college degree" categories). Figure 2 underscores that different demographic groups ascribe substantially (though not dramatically) different meanings to health-related response categories.

Our earlier claim of vignette equivalence is supported by the monotone decreasing theta (θ) values calculated by hopit (Table 7) (King et al. 2004:199).

In sum, our ordered probit/hopit analyses demonstrate that DIF indeed affects apparent predictors of SRH. Some variables affect rating style but do not lead to errors in rank ordering of groups' unadjusted SRH. For example, greater education is associated with greater health-optimism, but unadjusted ordered probit analyses still correctly show a positive relationship between education and health—they just overstate its strength. In other cases, failure to adjust for DIF leads to outright errors in group rankings by SRH. Notably, a standard analysis of WLS data would incorrectly show women in our sample to have better SRH than men, whereas, correcting for DIF, their SRH is equal to or worse than men's.

SUMMARY AND DISCUSSION

Our results indicate that creating anchoring vignettes to adjust the general self-rated health item is possible: Our vignettes are comprehensible to respondents,

show minimal violation of the method's measurement assumptions, and reveal several demographic and health-related variables associated with differences in rating style (DIF)—most consistently, sex and education. More importantly, we show that failure to account for DIF in SRH can yield incorrect research findings involving fundamental demographic categories. Treating SRH as a dependent variable, we demonstrated that neglecting DIF can lead to misestimation of an effect's strength (e.g., education) or even to a reversal of an independent variable's correct sign (e.g., when women in our sample appear to have better SRH than men, when in fact their SRH is the same or worse). Using SRH as an independent variable could likewise be problematic when DIF is nontrivial.

There were few differences in adherence to measurement assumptions or in substantive findings among our three vignette series. We also found no support for the idea that mention of specific disease conditions affects men's health ratings more than women's. There was, however, some evidence that familiarity with a health problem (e.g., hypertension) leads to more health-optimistic ratings of vignettes mentioning that problem. Researchers may thus prefer the No Specific Disease vignettes, to minimize bias due to differential disease knowledge among groups.

Anchoring vignettes have a number of advantages over earlier approaches to identifying DIF: They are a more direct and potentially less error-prone method than is the residual regression approach; they can both identify DIF and statistically correct for it; their costs are relatively low; the number of additional survey items required is small; and, by focusing on universal experiences such as pain and fatigue (as in our No Specific Disease series), vignettes might avoid problems of cultural or regional differences in access to medical diagnoses or taxonomies of disease. Vignettes may also be useful in multilingual contexts, serving as a safeguard against translation-triggered DIF. We thus believe that general health anchoring vignettes have potential to serve a valuable role in health research, enabling more accurate empirical work and more rigorous honing of theory.

Nevertheless, it would be premature to recommend that our vignettes, with their precise wording, be used more generally. Current analyses were limited to a racially homogenous, American sample with a narrow age range, and even within this sample our vignettes were not optimal. The unexpected negative correlation between age and vignette ratings suggests that respondents neglected to treat vignette characters as age peers; we thus recommend improved wording (see Grol-Prokopczyk 2010). Also, the vignettes elicited more rankings of poor or fair health than of very good or excellent health, while participants' self-ratings skewed in the opposite direction. Better alignment of the distributions would improve hopit's statistical efficiency (King and Wand 2007:61).

Furthermore, our study was limited by the fact that respondents received one vignette from each series, rather than a complete series. This design forced us to use a parametric approach (hopit) rather than Wand's newer, nonparametric techniques (http://wand.stanford.edu/anchors/). While hopit reveals *group* differences in SRH, nonparametric techniques permit adjustment of *individual* SRH scores, which can serve as dependent or independent variables (hopit, in contrast, requires that SRH be the dependent variable). With individually adjusted scores, one could test, for example, whether adjusted SRH better predicts mortality than raw SRH.[9] We recommend researchers give respondents full vignette series to enable nonparametric analyses. (Parametric designs may still be useful for identifying and correcting for DIF in certain contexts, however.)

Another potential design improvement concerns placement of vignettes vis-à-vis self-ratings. We administered the SRH question several minutes before the vignettes, according to prevailing wisdom at the time, which held that priming effects of vignettes on self-ratings should be avoided. Hopkins and King (2010), however, argue *in favor* of placing vignettes immediately before self-assessments, to "clarify the meaning of the self-assessment question and familiarize the respondents with the response scale, further improving measurement" (p. 208). Their experiments support such intentional use of priming.

As survey researchers have become increasingly interested in comparative studies, and as the problem of DIF has become more widely appreciated, anchoring vignettes have been proposed as a means of improving the comparative validity of self-report measures. Our work indicates that anchoring vignettes are a promising, workable method for improving comparability of self-ratings of general health. The method remains fairly new, however, and continued refinement can be expected as investigators explore vignettes further.

ACKNOWLEDGMENTS

We thank Gary King, Mary McEniry, Jesse Norris, Jonathan Wand, Wisconsin Longitudinal Study (WLS) staff, our anonymous reviewers, and, especially, John Allen Logan for their assistance.

FUNDING

The authors disclosed receipt of the following financial support for the research and/or authorship of this article: This research was supported by a grant from the Robert Wood Johnson Foundation, by core grants to the Center for Demography of Health and Aging (P30 AG017266) and the Center for Demography and Ecology (R24 HD047873) at the University of Wisconsin-Madison, and by the Vilas Estate Trust, University of Wisconsin-Madison. Core funding for the Wisconsin Longitudinal Study comes from the National Institute on Aging (R01 AG-09775; P01 AG-21079). Hanna Grol-Prokopczyk was supported by a training grant in Population, Life Course and Aging from the National Institute on Aging (T32 AG00129).

NOTES

1. Women's mean SRH in our analytic sample (described in Table 1) is 3.73 out of 5, versus 3.58 for men ($p < .01$).
2. Some refer to this model as "chopit" (e.g., Rabe-Hesketh and Skrondal 2002), though more commonly "chopit" refers to models that use multiple ratings of each vignette to calculate individual-level random effects.
3. Statistical analyses were done with Stata SE/10.1, using the gllamm program (www.gllamm.org) for hopit. Appendix B (available at http://jhsb.sagepub.com/supplemental) contains complete code for this article.
4. Because self-rated health (SRH) is not reducible to a health index score or physical symptoms list, and because of other random error, we would not expect perfect correlation between self-ratings and vignette ratings, but negative or absent correlation would be a serious cause for concern.
5. Models including sex and age reveal nearly identical coefficients for vignette ratings.
6. The models in Tables 5 and 6 do not meet the parallel regression assumption ($p < .01$ in an approximate likelihood ratio test), meaning that the effects of independent variables are not constant across all binary pairings of response categories. Results shown are broadly correct, however, in that the direction and significance of covariates are entirely consistent with findings from binary response models. Due to lack of preferable alternatives (Greene and Hensher 2010:188), and since the hopit model (Table 7) *does* show separate coefficients by threshold, we retain these models. However, to not grant the models' coefficients undue significance, we base this section's examples of sex differences on simple cross-tabulations of our data, not on the models' output.
7. A companion experiment shows that women rate our vignettes more highly than men regardless of vignette characters' sex (Grol-Prokopczyk 2010). That is, respondents' sex, not vignette characters' sex, drives our findings.
8. Despite this minor violation, we find strong overall evidence of response consistency (Table 4). We control for age in all DIF-related models, and remain confident in our other findings.
9. Vignette-based adjustment may make SRH *less* predictive of mortality if the DIF being erased reflects respondents' knowledge of their mortality risk. The sex differences identified in this article, however, remained strong in models including measures of perceived mortality risk (e.g., "How certain are you that you will live for another 10 years?").

REFERENCES

Banks, James, Michael Marmot, Zoë Oldfield, and James P. Smith. 2007. "The SES Health Gradient on Both Sides of the Atlantic." IFS Working Paper W07/04. Retrieved January 6, 2009 (http://eprints.ucl.ac.uk/2653/1/2653.pdf).

Barsky, Arthur J., Heli M. Peekna, and Jonathan F. Borus. 2001. "Somatic Symptom Reporting in Women and Men." *Psychosomatic Medicine* 62:354–64.

Benyamini, Yael, Elaine A. Leventhal, and Howard Leventhal. 2000. "Gender Differences in Processing Information for Making Self-Assessments of Health." *Psychosomatic Medicine* 62:354–64.

Case, Anne and Christina Paxson. 2005. "Sex Differences in Morbidity and Mortality." *Demography* 42:189–214.

Datta Gupta, Nabanita, Nicolai Kristensen, and Dario Pozzoli. 2010. "External Validation of the Use of Vignettes in Cross-Country Health Studies." *Economic Modelling* 27:854–65.

Deeg, Dorly J. H. and Didi M. W. Kriegsman. 2003. "Concepts of Self-Rated Health: Specifying the Gender Difference in Mortality Risk." *The Gerontologist* 43:376–86.

DeSalvo, Karen B., Nicole Bloser, Kristi Reynolds, Jiang He, and Paul Muntner. 2006. "Mortality Prediction with a Single General Self-Rated Health Question: A Meta-Analysis." *Journal of General Internal Medicine* 21:267–75.

Dowd, Jennifer Beam and Anna Zajacova. 2007. "Does the Predictive Power of Self-Rated Health for Subsequent Mortality Risk Vary by Socioeconomic Status in the US?" *International Journal of Epidemiology* 36:1214–21.

Ferraro, Kenneth F. 1980. "Self-Ratings of Health among the Old and the Old-Old." *Journal of Health and Social Behavior* 21:377–83.

Fillenbaum, G. G. 1979. "Social Context and Self-Assessments of Health among the Elderly." *Journal of Health and Social Behavior* 20:45–51.

Frankenberg, Elizabeth and Nathan R. Jones. 2004. "Self-Rated Health and Mortality: Does the Relationship Extend to a Low Income Setting?" *Journal of Health and Social Behavior* 45:441–52.

Greene, William H. and David A. Hensher. 2010. *Modeling Ordered Choices: A Primer*. New York: Cambridge University Press.

Grol-Prokopczyk, Hanna. 2010. "Age, Sex, and Race Effects in Anchoring Vignette Studies: Methodological and Empirical Contributions." CDE Working Paper No. 2010–18, University of Wisconsin-Madison.

Groot, Wim. 2000. "Adaptation and Scale of Reference Bias in Self-Assessments of Quality of Life." *Journal of Health Economics* 19:403–20.

Hauser, Robert M. and Carol L. Roan. 2006. "The Class of 1957 in their Mid-60s: A First Look (with Variables)." CDE Working Paper No. 2006–03, University of Wisconsin-Madison.

Hopkins, Daniel J. and Gary King. 2010. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly* 74: 201–22.

Idler, Ellen L. 1993. "Age Differences in Self-Assessments of Health: Age Changes, Cohort Differences, or Survivorship?" *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 48: S289–S300.

Idler, Ellen L. and Yael Benyamini. 1997. "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." *Journal of Health and Social Behavior* 38:21–37.

Idler, Ellen L. and S. V. Kasl. 1995. "Self-Ratings of Health: Do They Also Predict Change in Functional Ability?" *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 50:S344–53.

Jürges, Hendrik. 2007. "True Health vs Response Styles: Exploring Cross-Country Differences in Self-Reported Health." *Health Economics* 16:163–78.

Jylhä, Marja, Jack M. Guralnik, Luigi Ferrucci, Jukka Jokela, and Eino Heikkinen. 1998. "Is Self-Rated Health Comparable Across Cultures and Genders?" *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 53:S144–52.

Jylhä, Marja, Stefano Volpato, and Jack M. Guralnik. 2006. "Self-Rated Health Showed a Graded Association with Frequently Used Biomarkers in a Large Population Sample." *Journal of Clinical Epidemiology* 59:465–71.

King, Gary. 2005. Personal communication made at the June meeting of the Robert Wood Johnson Scholars in Health Policy Research Program, Aspen, CO.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Survey Research." *American Political Science Review* 98:191–207.

King, Gary and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15:46–66.

Krause, Neal M. and Gina M. Jay. 1994. "What Do Global Self-Rated Health Items Measure?" *Medical Care* 32:930–42.

Kroenke, Kurt and Robert L. Spitzer. 1998. "Gender Differences in the Reporting of Physical and Somatoform Symptoms." *Psychosomatic Medicine* 60:150–55.

Macintyre, Sally, Graeme Ford, and Kate Hunt. 1999. "Do Women 'Over-Report' Morbidity? Men's and Women's Responses to Structured Prompting on a Standard Question on Long Standing Illness." *Social Science and Medicine* 48:89–98.

Menec, Verena H., Shahin Shooshtari, and Pascal Lambert. 2007. "Ethnic Differences in Self-Rated Health among Older Adults: A Cross-Sectional and Longitudinal Analysis." *Journal of Aging and Health* 19:62–86.

Murray, Christopher J. L. and Lincoln C. Chen. 1992. "Understanding Morbidity Change." *Population and Development Review* 18:481–503.

Murray, Christopher J. L., Ajay Tandon, Joshua A. Salomon, Colin D. Mathers, and Ritu Sadana. 2002. "New Approaches to Enhance Cross-Population Comparability of Survey Results." Pp. 421–31 in *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, edited by C. J. L. Murray, J. A. Salomon, C. D. Mathers, and A. D. Lopez. Geneva: World Health Organization.

Rabe-Hesketh, Sophia and Anders Skrondal. 2002. "Estimating Chopit Models in gllamm: Political Efficacy Example from King et al." Retrieved January 20, 2009 (http://www.gllamm.org/chopit.pdf).

Shetterly, Susan M., Judith Baxter, Lynn D. Mason, and Richard F. Hamman. 1996. "Self-Rated Health among Hispanic vs. Non-Hispanic White Adults: The San Luis Valley Health and Aging Study." *American Journal of Public Health* 86:1798–801.

Verbrugge, Lois M. 1989. "The Twain Meet: Empirical Explanations of Sex Differences in Health and Mortality." *Journal of Health and Social Behavior* 30:282–304.

Zimmer, Zachary, Josefina Natividad, Hui-Sheng Lin, and Napaporn Chayovan. 2000. "A Cross-National Examination of the Determinants of Self-Assessed Health." *Journal of Health and Social Behavior* 41:465–81.

Bios

Hanna Grol-Prokopczyk is a PhD candidate in sociology at the University of Wisconsin-Madison specializing in the sociology of health and medicine. Her dissertation explores the measurement and social meanings of chronic pain.

Jeremy Freese is Professor and Chair of the Department of Sociology and faculty fellow of the Institute for Policy Research at Northwestern University. He is engaged in a variety of research projects that draw connections across social, psychological, and biological processes, especially in the context of technological and social policy change.

Robert M. Hauser is Vilas Research Professor of Sociology, Emeritus, at the University of Wisconsin-Madison and Executive Director of the Division of Behavioral and Social Sciences and Education at the National Research Council. He has been an investigator on the Wisconsin Longitudinal Study (WLS) since 1969 and has led the study since 1980.