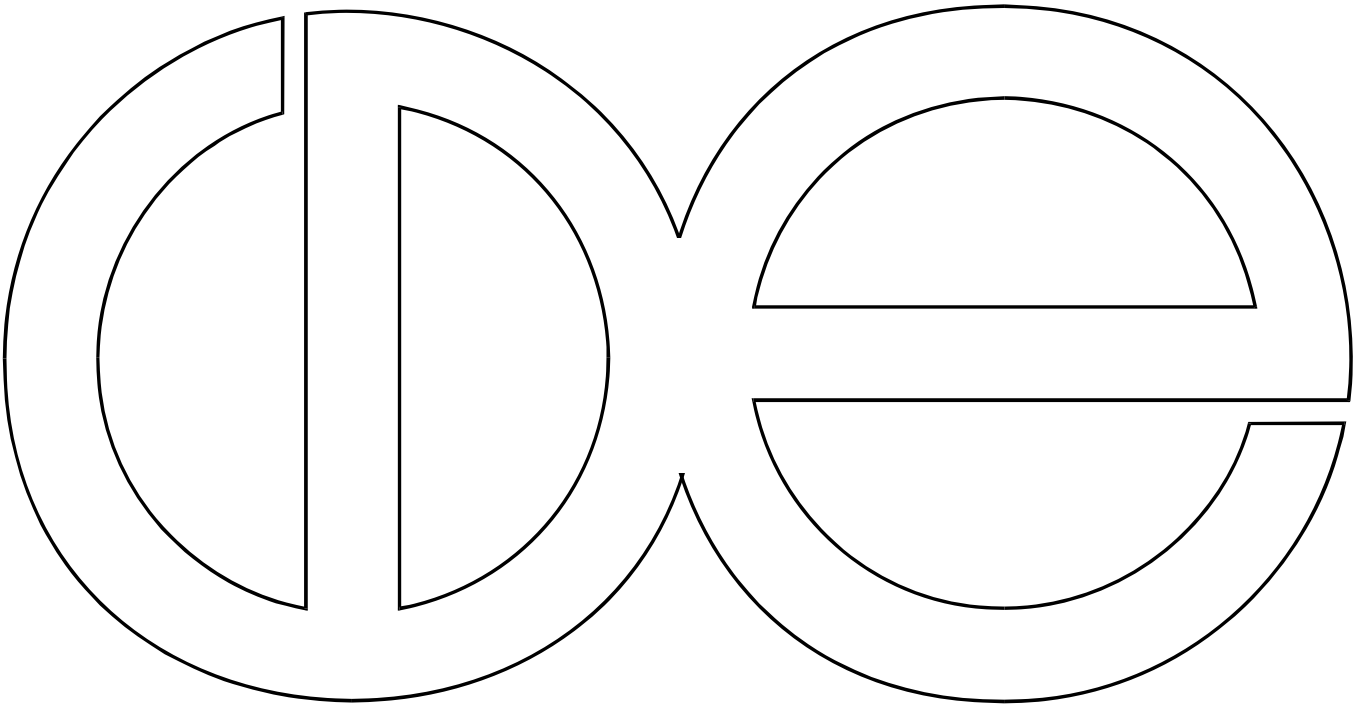
Center for Demography and Ecology

University of Wisconsin-Madison

Anchors – A Way?
Using Anchoring Vignettes to Assess Group Differences in Self-Rated Health

Hanna Grol-Prokopczyk
Jeremy Freese
Robert M. Hauser

CDE Working Paper No. 2010-09



Anchors—A Way?

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

Hanna Grol-Prokopczyk

University of Wisconsin-Madison

Jeremy Freese

Northwestern University

Robert M. Hauser

University of Wisconsin-Madison

Word count: 8,080

Number of tables: 7

Number of figures: 3

Running head: Adjusting for Group Differences in Health-Rating Style

* The development and administration of the general health anchoring vignettes discussed in this paper was enabled by a grant from the Robert Wood Johnson Foundation. Core funding for the Wisconsin Longitudinal Study comes from the National Institute on Aging (R01 AG-09775 and P01 AG-21079). Hanna Grol-Prokopczyk is supported by a training grant in Aging and Population from the National Institute on Aging. We thank Gary King, W. Whipple Neely, Jonathan Wand, our anonymous reviewers, and, especially, John Allen Logan for their assistance. Opinions expressed herein are those of the authors.

Address correspondence to Hanna Grol-Prokopczyk, University of Wisconsin-Madison, Department of Sociology, 8128 Sewell Social Sciences Building, 1180 Observatory Drive, Madison, WI 53706 (email: hgrol@ssc.wisc.edu).

Anchors—A Way?

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

Abstract

This paper identifies a potentially serious problem with the widely used self-rated health (SRH) survey item: that different groups (e.g., different nationalities) have systematically different ways of using the item's response categories. Analyses based on unadjusted SRH may thus produce misleading or entirely erroneous results. We explore 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, we show how certain 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 our sample, models that fail to account for women's greater health-optimism show—incorrectly—that women have better SRH than men. Vignette-based adjustment may reduce risk of such errors. Implications for future research using anchoring vignettes are discussed.

Anchors—A Way?

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

GROUP DIFFERENCES IN HEALTH-RATING STYLE

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' subsequent mortality, even after known health risk factors have been accounted for (e.g., DeSalvo et al. 2006; Idler and Benyamini 1997); and to predict subsequent functional ability among respondents who survive, net of baseline health and socioeconomic variables (Idler and Kasl 1995).

However, a large body of evidence suggests a potentially serious problem with the SRH item, namely, that not all respondents use the response categories ("excellent", "very good", etc.) in the same way. For example, Banks et al. (2007) compared American and English men's health and found a puzzling contradiction: based either on self-reports of specific diseases or on biological measures of disease, American men had objectively worse health than Englishmen, but on the SRH question, they reported having *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' health status, Americans are much more likely to say their health is good than are the English" (28). To borrow a phrase used by Ferraro in a

similar context (1980:381), American men would seem to be more "health-optimistic" than English men. Had Banks et al. relied exclusively on the SRH item, they might have concluded, incorrectly, that American men are healthier than English men. Similar evidence of differential use of response categories when answering the SRH item has been found across Asian countries (Zimmer et al. 2000), across European countries (e.g., Jürges 2007; Jylhä et al. 1998), across racial/ethnic groups (e.g., Cockerham, Sharp, and Wilcox 1983; Menec, Shooshtari, and Lambert 2007; Shetterly et al. 1996), across socioeconomic strata (e.g., Dowd and Zajacova 2007; Ferraro 1980; Idler 1993); and across age groups (e.g., Ferraro 1980; Idler 1993; Jylhä et al. 1998).

Men and women, too, may vary in their health-optimism. It has been amply demonstrated in health survey literature that, despite lower mortality rates at most ages, women report "more intense, more numerous, and more frequent" physical health problems than men across the lifecourse (Barsky, Peekna, and Borus 2001:266; cf. Austad 2006; Fillenbaum 1979); indeed, some studies find that "[m]ost physical symptoms are typically reported at least 50% more often by women" (Kroenke and Spitzer 1998:150). While at young and middle ages, self-rated health 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 even more positive than men's (Ferraro 1980:380,381), despite their greater number of individual somatic symptoms. This is also the case in the 2005 Wisconsin Longitudinal Study interviews of sibling respondents and their spouses, in which women on average

give slightly higher self-ratings of health than men (3.73 out of 5 versus 3.58 for men; $p<.01$; n=2,625—see Figure 1), even while reporting significantly more health problems (Hauser and Roan 2006:74-75).[1] Such data suggest that, in older populations, women may be more health-optimistic than men.

[FIGURE 1 ABOUT HERE ("Self-Rated Health in the Wisconsin Longitudinal…")]

Despite such discrepancies between objective health conditions and subjective ratings of overall health, 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). To accurately evaluate such claims, however, it is important to establish theoretical clarity about the concept of "health-reporting behavior". Specifically, 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. 1) First, people may have different health-reporting styles simply because they differ in what they mean by "health"; for example, in whether mental health is considered a component of overall health. While evidence is mixed in the context of sex differences, 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), or in their likelihood of considering "'trivial' or mental health conditions" (Macintyre et al. 1999:89) (but cf. Benyamini et al. 2000; Deeg and Kriegsman 2003). 2) Second, some groups may give less accurate self-reports of health due either to lack of self-knowledge or to disinterest in survey participation. For example, men might give higher health self-ratings than warranted because they do not know, remember, or care to reflect upon their individual

medical problems. Again, however, empirical evidence argues against gender differences of this type (Macintyre et al. 1999; Verbrugge 1989). 3) Third, as described at the start of this paper, groups may differ in how they use survey response categories to rate their health, that is, in where along the health spectrum they locate thresholds between "poor" and "fair", "fair" and "good", etc. This phenomenon—which has been termed "response category differential item functioning", or DIF for short (King et al. 2004)—is the focus of this paper (and for convenience, 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 a generalization based on evidence relevant only to the first two categories described here; the issue of DIF was not addressed.

While differences in health conceptualization and respondent thoroughness have been studied directly, response-category DIF is generally deduced by process of elimination, that is, by identifying discrepancies in SRH that persist even when relatively objective health measures are controlled for. Most commonly, SRH scores are regressed on a large number 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.

There are several problems with this indirect approach to assessing DIF, however. First, it is prone to Type I error if insufficient control variables are included (for example, disease severities as well as disease diagnoses), as well as Type II error, due to possible suppression effects if controls are cherry-picked to remove evidence of DIF. Second, in some contexts this approach is impossible to undertake at all, if researchers do not have access to relevant control variables, e.g., when costs make extensive health history

questionnaires or biomarker collection impossible, or when the national or ethnic groups being compared differ in their disease taxonomies or access to disease diagnoses. (We note that if the SRH variable cannot be used unless accompanied by large amounts of objective diagnostic data, its selling point as "a surrogate for more objective health measures" [Mansyur et al. 2008:46] is greatly undermined.) Finally, even if the residual regression approach is both doable and correct in identifying DIF, the approach does not suggest any clear solutions for statistically overcoming the problem of DIF in subsequent analyses involving SRH. Some authors have suggested doing separate analyses by subgroup (e.g., men separately from women [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 cursorily list it as a potential source of error, or helplessly 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—may be subject to response-category DIF in the context of self-rated health, a fact which threatens 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 with confidence, and even more difficult to adjust for DIF statistically. In this paper, we investigate a technique with potential to help overcome such methodological problems, that is, a technique that directly measures and statistically adjusts for group differences in use of response categories: anchoring vignettes.

ANCHORING VIGNETTES

Whenever subjective ordered response categories are used (not only in the context of health), differences in responses can potentially reflect response-category DIF rather than differences in the actual variable of interest.[2] To correct for this, King and colleagues have proposed the use of anchoring vignettes (King et al. 2004; King and Wand 2007): brief texts depicting hypothetical individuals who manifest the trait of interest to a lesser or greater degree. Respondents are asked to rate each character's level of the trait on the same response scale as their own self-rating. Typically respondents are given several vignettes, intended to represent various levels of the trait.

Vignette methodology posits that self-assessments reflect some systematic combination of the real value of interest plus DIF. By giving multiple respondents the same vignette, researchers fix the real value across respondents, so that any differences in ratings must be attributable to DIF (plus random error). Through rescaling of response-category thresholds, self-ratings can then be adjusted for DIF, to yield measures that are directly comparable across groups. The logic of this method is presented visually in Figure 2. While anchoring vignettes do not address *why* there are systematic differences in rating styles among groups, they can be used to demonstrate, quantify, and adjust for such differences.

[FIGURE 2 ABOUT HERE ("Schematic Diagram of Logic…")]

The two measurement assumptions made by the anchoring vignette method are *response consistency* and *vignette equivalence* (King et al. 2004:194). Response consistency refers to the assumption that respondents use response categories the same

way when rating the vignettes as they do when rating themselves (rather than holding themselves to higher or lower standards than the vignette characters). Vignette equivalence refers to the expectation that all respondents perceive a vignette as representing the same underlying concept, with vignettes in the same series all seen as part of a unidimensional scale.

Anchoring vignettes are being used in a growing number of surveys, including those conducted in international contexts, such as the World Health Organization's World Health Surveys (Salomon, Tandon, and Murray 2004; cf. Kapteyn, Smith, and van Soest 2007 and Kristensen and Johanssen 2008). They have been applied to a wide variety of research areas, including political efficacy, specific domains of health (such as vision, mobility, and affect), job satisfaction, school community strength, happiness, and women's autonomy (Hopkins and King 2008:2; see also numerous examples of vignette-based studies on the 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 the clear indications that response-category DIF is an issue in analyses using SRH. Some originators of the vignette method have expressed skepticism that vignettes could be fruitfully used to calibrate SRH, as they consider overall physical health too complex to be represented in a single dimension (King 2005). We believe this claim is worth testing directly, and do so in what follows.

ANALYTIC STRATEGY

The primary goal of this paper is to create and test a series of anchoring vignettes to be used to calibrate the general SRH item. Specifically, we test whether our vignettes meet the assumptions of the anchoring vignette method, and whether they support findings suggested by previous research, namely, that female sex (among older adults), age, and higher levels of education all predict greater health-optimism (e.g., Ferraro 1980; Idler 1993; Jylhä et al. 1998). (Our data do not allow us to assess national, linguistic, or racial-ethnic differences in rating style.) If our anchoring vignettes function as intended, they could enhance the SRH item's validity as both a dependent and an independent variable and allow for more accurate intergroup comparisons. A secondary goal is to compare the merits of general health vignettes that mention specific disease diagnoses (e.g., diabetes) with those that focus on non-disease-specific aspects of health.

There are three stages to our analysis. First, we create and administer three series of general health vignettes, and examine whether the vignettes satisfy assumptions of response consistency and vignette equivalence. Second, we assess whether demographic and health-related factors such as age, sex, education, and disease experience affect vignette ratings, i.e., whether these factors are associated with differential item functioning. (If there is no DIF, then there is no need to proceed further, as unadjusted self-ratings of health will be unbiased and comparable among groups.) Third, we compare a standard, unadjusted analysis of predictors of SRH with an analysis that statistically incorporates information on DIF. This reveals how DIF affects the strength and/or direction of coefficients in predicting SRH. We pay particular attention to sex differences, to see if vignette-based adjustments resolve the previously discussed paradox of women's greater number of physical ailments but higher SRH scores.

DATA AND METHODS

Data

The Wisconsin Longitudinal Study (WLS) began in 1957 as a 1/3 random sample (n=10,317) of graduating Wisconsin high schools seniors, and was expanded in subsequent waves to include a randomly selected sibling of each graduate ("siblings") and the spouses of those siblings ("sibling-spouses"). Our analyses are based on sibling and sibling-spouse data collected in the 2005-2007 surveys. A primary limitation of the data is that, reflecting the demographics of Wisconsin high school graduates at the time the survey began, respondents are almost without exception white (99% identify as exclusively white). WLS documentation and data are available on its website (http://www.ssc.wisc.edu/wlsresearch/).

The vignettes were administered as part of the WLS telephone survey to a random subset of sibling (n=1,221) and sibling-spouse (n=1,404) respondents, giving us an overall working sample size of 2,625. Because sibling respondents, but not spouses, were also administered a mail survey containing health-related information, some of the following analyses are conducted with the sibling sample only. Descriptive statistics pertaining to the analytic sample appear in Table 1.

[TABLE 1 ABOUT HERE ("Descriptive statistics for analytic sample")]

Vignette texts

Though evidence about sex differences in conceptualizations of health is mixed, some scholars suggest that men's health-ratings are more sensitive than women's to life-

threatening diseases (e.g., heart disease), as opposed to non-life-threatening factors affecting daily functioning (e.g., pain from arthritis) (Benyamini et al. 2000; Deeg and Kriegsman 2003:383). To ensure that this potential sex difference and source of multidimensionality would not bias our findings, we created three series of general health vignettes: one describing health-related daily functioning, but referring to no specific diseases (the "No Specific Disease" series); one supplementing the above with references to heart disease or related conditions (the "Heart Disease" series); and one supplementing the above 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, as well as to see whether having first- or second-hand experience with a particular medical condition affects how respondents rate characters with that condition.

Each series consisted of 4 vignettes of varying severity. The exact symptoms mentioned in the vignettes were chosen to capture typical variations in health among WLS participants, based on their responses to the general health question and to questions about specific health conditions and diagnoses. The Heart Disease and Diabetes vignettes were formed by adding a single disease-specific sentence to the corresponding No Specific Disease vignette.[3] Table 2 shows the vignette texts, as well as the text preceding vignette administration, which encourages respondents to rate the vignette characters just as they would rate themselves, and to consider them as age-peers. To further encourage response consistency, vignette characters were of the same sex as the respondents; the first names used in the vignettes (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 wording of the SRH question ("In general, would you say [vignette character]'s health is: excellent, very good, good, fair, or poor?"). The vignettes were administered shortly after the SRH question in the survey.

[TABLE 2 ABOUT HERE ("Text of general health vignettes.")]

Each respondent was given 3 vignettes—one from each of our 3 series—representing 3 different severity levels. Both the order of the series and assignment of severity levels to each series were randomly determined.

Variables and models

For ease of interpretation, both self-rated health and vignette ratings were reverse-coded so that higher values indicate better health (1 = "poor", 5 = "excellent").

To identify factors predicting vignette ratings, we estimated two ordered probit models: one including basic demographic variables (sex, age, education, and income), and one adding personal and familial health variables.

To assess how accounting for response category DIF affects apparent predictors of self-rated health, we implemented the parametric strategy used by King et al. (2004) and elaborated by Rabe-Hesketh and Skrondal (2002): namely, to compare a) a standard ordered probit regression of SRH on various independent variables, to b) a joint "chopit" regression for SRH and vignette ratings on the same independent variables.[4,5] Chopit, short for "compound hierarchical ordinal probit", re-scales the threshold parameters in a standard ordered probit model based on respondents' ratings of the anchoring vignettes, and thus reveals how self-assessments (in our case, of health) truly differ among groups,

after differences in rating styles have been controlled for (see Appendix I for formal specifications). In other words, chopit models estimate group differences after adjusting for DIF. Comparing the standard ordered probit with the chopit model thus shows how the associations between basic demographic covariates and SRH are affected by DIF.

RESULTS

Assessment of the general health vignettes

Mean vignette ratings showed the expected ordinality, both within and across disease series, as shown in Table 3. We note the smaller standard deviations for Severity 4 vignettes, suggesting a floor effect of the response categories. Among individual respondents, fewer than 9% gave ratings that violated the intended rank-ordering of vignettes by severity. These results, showing little evidence of multidimensionality, are consistent with the assumption of vignette equivalence.

[TABLE 3 ABOUT HERE ("Mean ratings of general health vignettes")]

To test for response consistency, we performed ordered probit regressions of SRH scores 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]). If two respondents have the same objective level of health, but nonetheless give different self-ratings of health, the difference in self-ratings should (if response consistency holds) be positively correlated with the difference in vignette ratings. In other words, the more health-optimistic self-rater should also be the more health-optimistic vignette-rater.[6]

[TABLE 4 ABOUT HERE ("Ordered probit regression of SRH on vignette…")]

Results of these regressions are displayed in Table 4.[7] Unsurprisingly, the association between physical health scores and SRH is both substantively large and statistically significant in all vignette series. More importantly for our purposes, vignette ratings are positively and significantly ($p < .001$) associated with self-ratings in all three vignette series. Thus, greater health-optimism in vignette ratings is indeed associated with greater health-optimism in self-ratings, providing evidence of response consistency.

Differences in vignette rating styles

[TABLE 5 ABOUT HERE ("Ordered probit reg'n of vig rating on demographic…")]

Table 5 presents estimates from ordered probit regressions of vignette ratings on sociodemographic variables. For all three vignette series, women give higher ratings than men, a difference which is both statistically significant and not trivial in size. This is evidence that women are indeed more health-optimistic than men. The magnitude of this sex difference may be conveyed by some simple comparisons: 34% of men considered the Heart Disease Severity 1 vignette character to have "excellent" health, while 48% of women did. For the Diabetes Severity 3 vignette, 33% of men selected "poor" and 13% selected "good"; the comparable percentages for women were 17% and 24%, respectively. Fifty-eight percent of men, but only 40% of women, gave a rating of "poor" to the No Disease Severity 4 vignette.[8] These examples of lower ratings by men are typical. The only two vignettes in which statistically significant sex differences were not observed were Heart Disease Severity 4 and Diabetes Severity 4. Unfortunately, it is unclear whether these exceptions indicate that men and women's ratings converge when

severe, specific diseases are mentioned, or whether these exceptions are artifacts of category floor effects. Further experimentation would be needed to clarify this issue.

Relatedly, models including interactions between sex and series (not shown) find no evidence that men's ratings of health are differently affected than women's by mention of specific health ailments or diseases, except in the case of the Heart Disease Severity 4 vignette—which women rated *more negatively* than men. (This was true whether or not health-related variables, as found in Table 6, were in the models.) Again, response truncation must be considered, but given that this lone interaction effect was opposite the direction predicted by the aforementioned theory of sex differences, we conclude that our current data do not support the theory. (Further comparisons with differently-worded vignettes may still be warranted, however, to test for other possible sources of multidimensionality).

Age shows a significant association with vignette ratings in the No Disease series, in which higher age predicts more negative ratings, though the effect size is small. This finding is at odds with previous literature (e.g., Idler 1993), a fact discussed further below. Consistent with previous literature, higher levels of education are positively correlated with vignette ratings. Indeed, there appears to be a roughly linear, positive effect of years of schooling on vignette ratings. Income is unrelated to ratings net of other variables; this is confirmed by a Wald test of the joint significance of the income dummy variables.

Having established that certain basic demographic variables predict vignette ratings, we next tested a model that also included several measures of first- and second-hand experience with specific health conditions, as well as the HSS and HUI-3. We

reasoned that people with personal or familial experience of heart disease, diabetes, or related conditions might respond differently to disease-specific vignettes than those without such experience, even when controlling for overall health.[9]

[TABLE 6 ABOUT HERE ("…vignette rating on demographic and health-related…")]

This hypothesis is borne out by our results (Table 6). Respondents with hypertension ranked vignettes in the Heart Disease series significantly more positively than did those without. So, too, did respondents whose parents, siblings, or spouses had suffered heart attacks (p=.06). These findings suggest that direct experience with heart-related conditions leads respondents to consider them less severe (perhaps understandably so: our respondents with hypertension have clearly lived to tell about it). It is somewhat surprising that respondents' heart problems are not similarly associated with Heart Disease rankings, but this could be a result of 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 specific Heart Disease vignettes (rather than the series of four vignettes taken as a whole), personal experience with heart problems significantly predicts more positive ratings when angioplasty (β=.282; p=.019; n=672) or heart attack (β=.327; p=.017; n=680) are mentioned. We found no parallel evidence that experience with diabetes affects ratings of the Diabetes series.

In addition to the models shown in Tables 5 and 6 above, we tested others which included measures of personality, depression, religiosity, social participation, and psychological well-being, but none of these showed systematic association with vignette

ratings. However, in *all* models tested, sex is strongly and significantly related to vignette ratings, in all three series. The sex effect is thus the most robust finding from our analyses thus far, and it is consistent with our suspicions, expressed in our introduction, that in this age group women are more health-optimistic than men.

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

Group differences in self-rated health (SRH)

Assuming response consistency, the presence of DIF in vignette ratings implies the presence of DIF in SRH scores. How does taking this DIF into account affect apparent predictors of SRH? As explained earlier, we answer this question by running a standard ordered probit regression of SRH on demographic variables, and comparing this with a "chopit" regression that adjusts for DIF by re-scaling groups' response category thresholds based on their vignette ratings. Due to space restrictions, we show only the findings based on the No Disease series of vignettes. The findings from the other series were extremely similar, with one exception, mentioned below.

[TABLE 7 ABOUT HERE ("Ordered probit and chopit regressions…")]

Table 7 presents our comparison of ordered probit and chopit models regressing self-rated health on demographic variables. The columns pertaining to the ordered probit results show that nearly all our independent variables significantly predict SRH. As mentioned in our introduction, women in this sample report better health than men. Consistent with expectations, younger respondents report better health than older ones,

and education is positively associated with better SRH. The association of income with SRH is as expected except for an inversion in the bottom two income quartiles, which supplementary analyses indicate is accounted for in a model that adds measures of wealth (not shown); this finding 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 standard ordered probit to the chopit regression. Perhaps most strikingly, the coefficient for female, which had been positive, now becomes negative (though not statistically significant). In other words, the apparent better health of women disappears when vignette rating style is accounted for. The puzzle of women's surprisingly high self-reported ratings of health appears, then, due at least in part to sex differences in use of response categories.

Age continues to be negatively associated with SRH in the chopit model, though this effect ceases to be statistically significant. (It remains significant when using the Heart Disease series, however.) The lack of a statistically significant age effect in two of the three chopit models is surprising, even implausible, though it is consistent with—indeed, caused by—the earlier finding that older respondents show greater health-pessimism in vignette ratings (and thus have their self-ratings adjusted upwards by chopit). Given previous research finding that increasing age is associated with health-*optimism* (Idler 1993; Jylhä et al. 1998), we suspect our vignettes may not be correctly measuring age differences in rating style. In particular, respondents may not attend to the instructions to treat vignette characters as age peers (a possibility suggested by survey audio recordings in which respondents ask the age of the vignette characters). This

situation could lead to violation of response consistency. Mentioning exact ages directly in vignette texts is a potential solution, which we plan to test experimentally.

Education continues to be positively associated with health in the chopit model, though the effect is weakened, with only the college degree variable remaining statistically significant. Income effects on health change little between the probit and chopit models. Again, respondents in the bottom income quartile show better health than those in the second quartile, for the reason explained above.

Chopit's information about predictors of threshold variation helps explain why findings differ between the standard probit and the chopit models. For example, the coefficient for female sex under Threshold 1 (-0.469) means that women have a lower threshold for the distinction between "poor" and "fair", i.e., women are more likely than men to choose "fair" over "poor" to describe a given vignette character. Since only the coefficients for Threshold 1 are open to such straightforward interpretation (because higher-order thresholds depend on previous ones, and involve exponentiation of coefficients/covariates [Appendix A, equation 1; cf. King et al. 2004:198]), estimated group differences in thresholds are best presented visually. In Figure 3, we show chopit's mean predicted thresholds for our sample by sex and by educational category. While the differences are not overwhelming in size, they are by no means trivial.

[FIGURE 3 ABOUT HERE ("Mean predicted intercategory thresholds…")]

There are fewer statistically significant predictors of thresholds as one moves from Threshold 1 to Threshold 4, reflecting the fact that our vignettes elicited many more ratings at the "poor" or "fair" end of the health spectrum than at the "very good" or "excellent" end, and so provided the chopit method with less information about ratings at

better levels of health. Future vignette series would be improved by including characters in better health. Table 7 also shows that theta (θ) values for the vignettes are monotone decreasing, supporting our claim of vignette equivalence (King et al. 2004:199).

In sum, our ordered probit/chopit analysis demonstrates that DIF can indeed affect apparent predictors of SRH. Some variables affect DIF, but do not lead to errors in rank ordering of groups' unadjusted SRH. For example, higher levels of education are associated with greater health-optimism, but unadjusted ordered probit analyses will still be correct in showing a positive relationship between education and health—they may just overstate its strength. In other cases, however, failure to take DIF into account does lead to outright errors in ranking groups by SRH. Notably, a standard analysis of WLS data would incorrectly show women in our sample to have better self-rated health than men, when in fact 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 (SRH) survey item is possible: our vignettes were comprehensible to respondents, appeared to meet assumptions of vignette equivalence and response consistency, and revealed a number of demographic and health-related variables associated with differences in rating styles (most consistently, sex and education). More importantly, we have shown that failure to account for response-category DIF in SRH can yield incorrect research findings, including ones involving very fundamental demographic categories. In analyses with SRH as a dependent variable, we demonstrated that not accounting for DIF can lead to misestimation of an effect's strength (e.g., that of

education on health), 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 non-trivial.

There were few differences, either in adherence to measurement assumptions or in substantive findings, among our three series of vignettes; we also found no support for the idea that men's health ratings are more greatly affected than women's by mention of specific disease conditions. There was, however, some evidence that first- or second-hand experience with a health problem (such as hypertension) can lead to more health-optimistic ratings of vignette characters with that problem. An argument could thus be made for preferring the No Specific Disease series, to avoid bias due to differential disease prevalence or 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 the residual regression approach described earlier; they can be used not only to identify DIF, but also to 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 in taxonomies of disease. Vignettes may be particularly 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 health theory.

At the same time, it would be premature to recommend that our vignettes, with their precise wording, be used more generally. Not only have our tests been limited to a racially homogenous, American sample with a narrow age range, but even within the sample our vignettes were not optimal. The unexpected negative correlation between age and vignette rating suggests that respondents neglect the instructions to treat vignette characters as age peers; if confirmed, rewording vignettes would be in order. Also, as mentioned earlier, the vignettes elicited more rankings of poor or fair health than of very good or excellent health, while participants' SRH scores were skewed in the opposite direction. King and Wand (2007:61) explain that chopit and related analyses are most effective when vignette- and self-ratings have similar distributions.

Furthermore, our study was limited by the fact that respondents only received one vignette from each series, rather than a complete series. This design forced us to use a parametric statistical analysis (chopit) rather than any of a rich array of newer, non-parametric techniques, most notably those found in Wand's "anchors" package in R (http://wand.stanford.edu/anchors/; cf. King et al. 2004 on parametric and non-parametric approaches). While our design allowed us to compare group differences in response styles and in SRH, non-parametric techniques would have permitted us to adjust *individuals'* SRH scores, and thus to directly compare individuals' health ratings. Individually-adjusted measures of SRH would be useful as both dependent and independent variables (chopit, in contrast, requires that SRH be the dependent variable), and would allow one to test, among other things, whether adjusted SRH better predicts subsequent mortality than raw SRH scores.[10] We thus recommend designing future surveys in such a way that non-parametric analyses are possible, that is, by giving

respondents a full series of anchoring vignettes. (If this is cost-prohibitive, however, a parametric design is still useful for identifying sources of DIF and correcting for DIF in certain contexts.)

Another potential design improvement pertains to placement of the vignettes vis-à-vis the self-rating. In our study, the SRH question was administered several minutes before the vignette questions, according to the prevailing wisdom at the time, which held that priming effects of vignette questions on self-ratings should be avoided. Recently, however, Hopkins and King (2008) have argued *in favor* of placing vignettes immediately before the self-assessment, since this can "clarify the meaning of the self-assessment… and familiarize the respondents with the response scale, further improving measurement" (6). Their experimental findings support such intentional use of priming.

As survey researchers have become increasingly interested in comparative studies and the problems of DIF have become more widely appreciated, anchoring vignettes have been proposed as a means of improving the validity of comparative self-report measures. While some skepticism has been raised about whether vignettes could be used to improve self-ratings of general health, our work so far indicates they are promising. Nevertheless, the method remains fairly new and continued refinement can be expected as investigators explore vignettes further and apply vignettes to more samples.

NOTES

[1] Based on the same WLS respondents who constitute our analytic sample, described in Table 1. The sex difference remains statistically significant when controlling for age and/or other demographic variables.

[2] Because the anchoring vignette method is predicated on differences in use of ordered response categories, its current use is restricted to ordinal (or perhaps binary) variables. We are unaware of comparable techniques for nominal or continuous variables.

[3] The disease-specific sentence was prepended to the base text for half the vignettes, and appended for the other half. The placement was found to have no effect on ratings. We also found no effect of respondents' first names matching those of vignette characters.

[4] Wand, King, and Lau (forthcoming:18; citing Wand 2007) suggest a new estimator as an improvement on the chopit method, but Wand (2008) confirms that in cases where respondents are given a single vignette from a series (as in the present study), there are no advantages to this alternate method over chopit.

[5] All statistical analyses were done using Stata SE/10.1. Chopit models were estimated using the Stata program gllamm (available at www.gllamm.org).

[6] Because SRH is not reducible to a health index score or a list of physical symptoms, and because of other random error, we would not expect a perfect correlation between self-rating and vignette-rating, but negative or absent correlation would be a cause for concern regarding the functioning of the vignettes.

[7] Models including sex and age as controls revealed nearly identical coefficients for the vignette ratings, and so are not shown.

[8] The ordered probit models shown in Tables 5 and 6 fail to meet the parallel regression assumption ($p<.01$ in an approximate likelihood ratio test), meaning that the effects of the independent variables are not constant across all binary pairings of response categories. The 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 2008:69), and since the chopit model in Table 7 does show separate coefficients for each threshold, we maintain the ordered probit models. However, to not grant the models' coefficients undue significance, we base this section's examples of sex differences in rating style on simple cross-tabulations of our data, not on the output of the models.

[9] Including the Health Symptoms Scale and relatives' health experiences in the model substantially reduced our sample size. Omitting those variables, however, had no effect on the direction or statistical significance of the remaining variables.

[10] We are open to the possibility that vignette-based adjustment will make SRH *less* predictive of mortality, if the DIF that is being erased reflects respondents' holistic knowledge of their overall mortality risk. For example, women may give higher vignette- and self-ratings than men because they sense that women, even with greater numbers of physical health symptoms, have lower risk of mortality than men. "Correcting" for this knowledge through vignette adjustment could then lead to underestimation of women's SRH scores, and overestimation of their mortality risks.

REFERENCES

Austad, Steven N. 2006. "Why Women Live Longer Than Men: Sex Differences in Longevity". *Gender Medicine* 3(2):79-92.

Banks, James, Michael Marmot, Zoë Oldfield, and James P. Smith. 2007. "The SES Health Gradient on Both Sides of the Atlantic". No. WP07/04. The Institute for Fiscal Studies, UCL (University College London). Available at http://eprints.ucl.ac.uk/2653/1/2653.pdf [last accessed 6 January 2009].

Barsky, Arthur J., Heli M. Peekna, and Jonathan F. Borus. 2001. "Somatic Symptom Reporting in Women and Men". *Psychosomatic Medicine* 62:354–364.

Benyamini, Yael, Elaine A. Leventhal, and Howard Leventhal. 2000. "Gender Differences in Processing Information for Making Self-Assessments of Health". *Psychosomatic Medicine* 62:354–364.

Case, Anne, and Christina Paxson. 2005. "Sex Differences in Morbidity and Mortality". *Demography* 42(2):189-214.

Cockerham, William C., Kimberly Sharp, and Julie A. Wilcox. 1983. "Aging and Perceived Health Status". *Journal of Gerontology* 38(3):349-355.

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(3):376-386.

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(3):267–275.

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(6):1214-1221.

Ferraro, Kenneth F. 1980. "Self-Ratings of Health among the Old and the Old-Old". *Journal of Health and Social Behavior* 21(4):377–383.

Fillenbaum, G. G. 1979. "Social Context and Self-Assessments of Health among the Elderly." *Journal of Health and Social Behavior* 20(1):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(4):441–452.

Greene, William H. and David A. Hensher. 2008 (June 15). *Modeling Ordered Choices: A Primer*. Social Science Research Network Working Paper Series. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1213093 [last accessed 10 March 2010]. Also forthcoming as Greene, William H. and David A. Hensher. 2010. *Modeling Ordered Choices: A Primer*. Cambridge University Press.

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, Madison, WI.

Hopkins, Daniel J. and Gary King. 2008 (Dec 28). "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." Working paper. Available at http://gking.harvard.edu/files/implement.pdf [last accessed 7 January 2009].

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(6):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(1):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(6):S344-S353.

Jürges, Hendrik. 2007. "True Health vs Response Styles: Exploring Cross-country Differences in Self-Reported Health". *Health Economics* 16(2):163-178.

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(3):S144-S152.

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(5):465–471.

Kapteyn, Arie, James P. Smith, and Arthur van Soest. 2007 (Mar). "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands". *The American Economic Review* 97(1):461-473.

King, Gary. 2005 (June). Personal communication made at the meetings 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 (Feb). "Enhancing the Validity and Cross-Cultural Comparability of Survey Research". *American Political Science Review* 98(1):191-207.

King, Gary and Jonathan Wand. 2007 (Winter). "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes". *Political Analysis* 15(1):46-66.

Krause, Neal M. and Gina M. Jay. 1994. "What Do Global Self-Rated Health Items Measure?" *Medical Care* 32(9):930-942.

Kristensen, Nicolai and Edvard Johansson. 2008. "New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes". *Labour Economics* 15(1):96-117.

Kroenke, Kurt and Robert L. Spitzer. 1998. "Gender Differences in the Reporting of Physical and Somatoform Symptoms." *Psychosomatic Medicine* 60:150-155.

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.

Mansyur, Carol, Benjamin C. Amick, Ronald B. Harrist, and Luisa Franzini. 2008. "Social Capital, Income Inequality, and Self-Rated Health in 45 Countries." *Journal of Health and Social Behavior* 66:43-56.

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(1):62-86.

Rabe-Hesketh, Sophia and Anders Skrondal. 2002 (Sept). "Estimating chopit models in gllamm: Political efficacy example from King et al.". Online document, available at http://www.gllamm.org/chopit.pdf [last accessed 20 January 2009].

Salomon, Joshua A., Ajay Tandon, and Christopher J. L. Murray. 2004 (Jan). "Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes". *BMJ* 328:258-261.

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(12):1798-1801.

Verbrugge, Lois M. 1989. "The Twain Meet: Empirical Explanations of Sex Differences in Health and Mortality." *Journal of Health and Social Behavior* 30(3):282-304.

Wand, Jonathan. 2007. "Improving the Measurement of Interpersonally Incomparable Data, II: Stochastic and Latent Class Models using Anchoring Vignettes or Other Common Survey Questions." Manuscript, Stanford University.

Wand, Jonathan. 2008 (31 Jan). Personal communication to first author via email.

Wand, Jonathan, Gary King, and Olivia Lau. Forthcoming. "Anchors: Software for Anchoring Vignette Data." *Journal of Statistical Software*.

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(4):465-481.

Table 1. Descriptive statistics for analytic sample.

	Proportion or Mean	SD	N
Female	.546		2,625
Self-rated health (SRH) 1=poor to 5=excellent	3.665	.990	2, 625
Age at time of interview, in years	63.791	7.732	2,624
Education			
Less than high school	.054		139
High school degree	.412		1,056
Some college	.194		497
4-year college degree	.181		463
Post-college education	.160		410
Household income, 2005	$74,979	$121,265	2,609
R ever diagnosed with diabetes/high blood sugar?	.156		2,620
R ever diagnosed with heart problems?	.154		2,622
R ever diagnosed with hypertension?	.477		2,622
Health Utilities Index (HUI-3) score: 0= health-state equivalent to death, 1=best health	.814	.218	2,625
Health Symptoms Scale (HSS) score[a]: Count of physical health symptoms (out of 25) experienced in past 6 months	8.875	5.108	999
R's parent(s)/sib(s)/spouse had diabetes[a]?	.402		1,012
R's parent(s)/sib(s)/spouse had heart attack[a]?	.471		1,012

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

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 Health 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. [S/he] 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. [S/he] 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. [S/he] 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 sentences below 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. [S/he] once underwent angioplasty to unblock an artery, and takes medication for these problems.
Severity 4	[Name] has very high blood pressure and cholesterol. [S/he] once had a heart attack, and subsequently had successful bypass surgery.
Diabetes series	*The sentences below 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 [hid/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?

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
	(.91)	(.78)	(.77)	(.62)
Heart Disease	4.19	2.86	1.63	1.32
	(.82)	(.81)	(.68)	(.51)
Diabetes	4.03	2.50	1.98	1.41
	(.92)	(.77)	(.70)	(.59)

Means calculated by assigning scores to responses of 1 = poor; 2 = fair; 3 = good; 4 = very good; 5 = excellent. Standard deviations in parentheses.

Table 4. Ordered probit regression 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	0.186***	0.137***	0.153***
	(.027)	(.030)	(.028)
Health Symptoms Scale score (÷ 10)	-0.636***	-0.627***	-0.626***
	(.074)	(.074)	(.074)
Heath Utilities Index	2.251***	2.239***	2.244***
	(.125)	(.124)	(.125)

*** $p < .001$, two-tailed. Models also include controls for vignette severity. Where missing, Health Symptoms Score is imputed (based on Health Utilities Index), to maintain sample size.

Table 5. Ordered probit regression of vignette rating on demographic variables.

	No Specific Disease series	Heart Disease series	Diabetes series
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)
4-year college degree or more	.242***	.181***	.265***
	(.053)	(.055)	(.054)
Household income, 2nd quartile	.100	.112	.066
	(.067)	(.070)	(.068)
Household income, 3rd quartile	-.033	.020	.025
	(.065)	(.068)	(.066)
Household income, 4th (top) quartile	.070	.062	-.004
	(.066)	(.069)	(.067)
N	2,546	2,546	2,543

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

Table 6. Ordered probit regression of vignette rating on demographic and health-related variables.

	No Specific Disease series	Heart Disease series	Diabetes series
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)
4-year college degree or more	.257**	.211*	.228**
	(.088)	(.091)	(.089)
Household income, 2nd quartile	.130	.154	.041
	(.107)	(.111)	(.109)
Household income, 3rd quartile	.000	.034	.030
	(.110)	(.114)	(.111)
Household income, 4th (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)
N	942	942	938

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

Table 7. Ordered probit and chopit regressions of self-rated health (SRH) on demographic variables.

	Ordered probit		Chopit	
	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
4-year college degree or more	.460***	.052	.309***	.073
Household income, 2nd quartile	-.176**	.064	-.265**	.089
Household income, 3rd quartile	.020	.063	.093	.088
Household income, 4th (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
4-year college degree or more			-.285***	.063
Household income, 2nd quartile			-.107	.076
Household income, 3rd 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
4-year college degree or more			.097	.059
Household income, 2nd quartile			.047	.077
Household income, 3rd 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
4-year college degree or more			.091	.057
Household income, 2nd quartile			-.047	.069
Household income, 3rd 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
4-year college degree or more			-.098†	.054
Household income, 2nd quartile			.007	.072
Household income, 3rd 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

N=2,548. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$, two-tailed. Chopit uses No Disease vignettes. Reference categories: "High school degree" (education); "Household income, bottom quartile" (inc.).

Figure 1: Self-Rated Health in the Wisconsin Longitudinal Study, By Sex

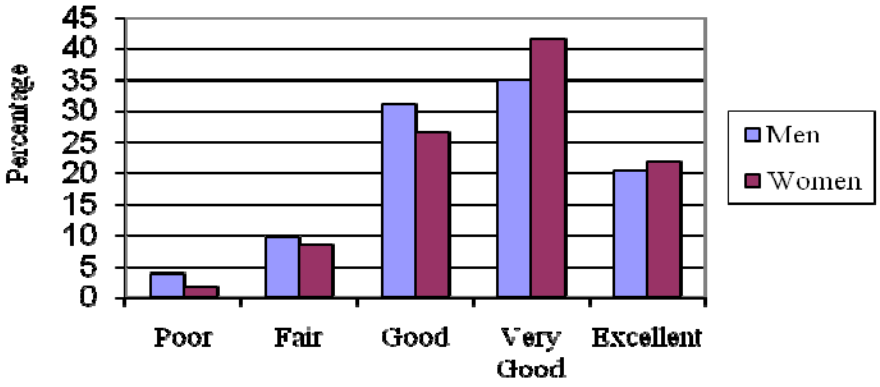
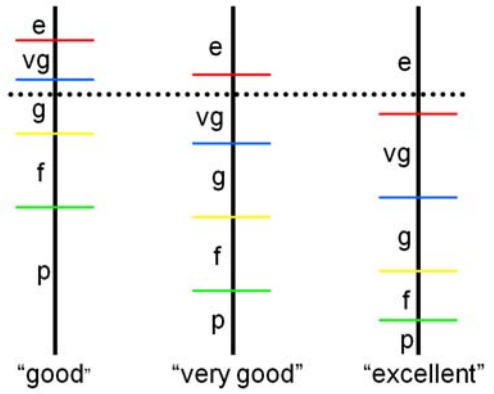
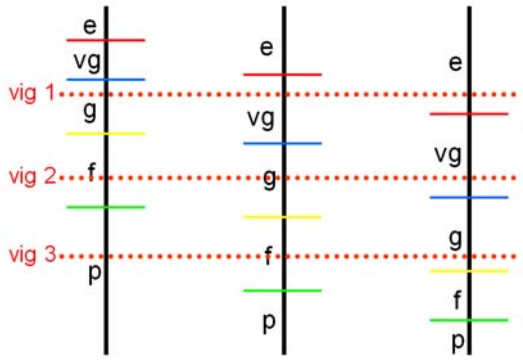


Figure 2: Schematic diagram of logic underlying the anchoring vignette method.

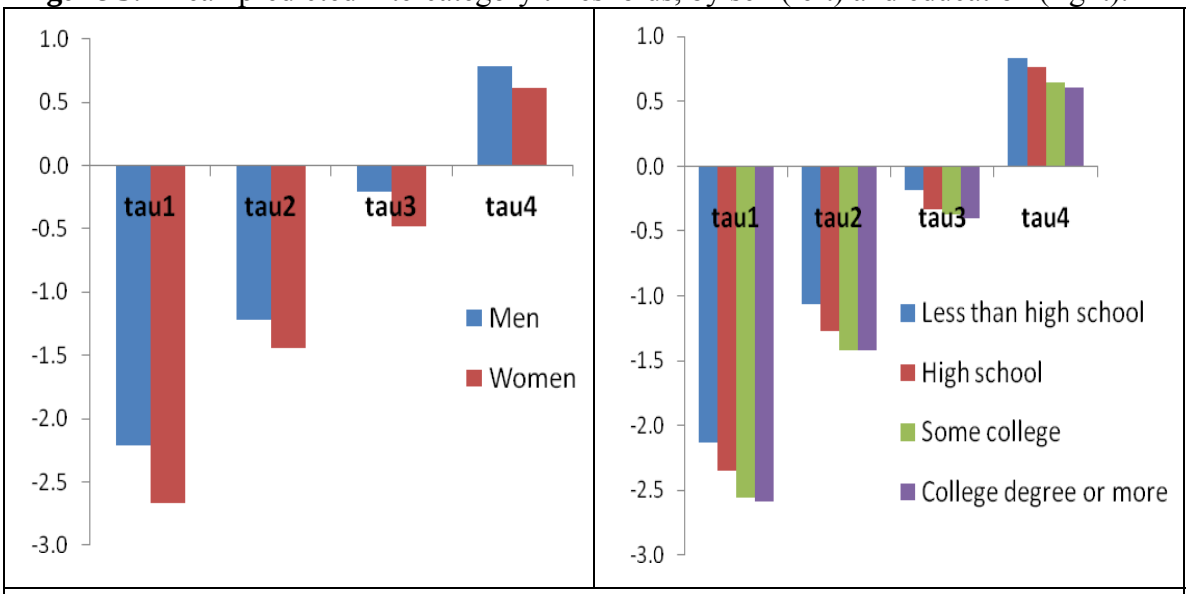


Left panel: Respondents differ in how they divide the health spectrum (vertical black line) into the categories of "excellent", "very good", "good", "fair", and "poor" (i.e., they show response-category DIF). Thus, multiple respondents could have the same level of health (dotted horizontal line), but use different terms to refer to this level of health. In this example, the right-most individual is the most "health-optimistic".



Right panel: Giving respondents multiple vignettes to rate (red dotted lines) reveals what different respondents mean by different response categories, and permits calculation of group differences in the location of intercategory thresholds. These threshold differences can subsequently be adjusted for statistically, allowing for group health comparisons unbiased by response-category DIF.

Figure 3: Mean predicted intercategory thresholds, by sex (left) and education (right).



Notes: tau1 refers to the threshold between "poor" and "fair"; tau2 to the threshold between "fair" and "good", etc. Graphs are based on threshold-predicting coefficients from chopit model (using No-Disease vignettes), applied to analytic sample.

APPENDIX A: Statistical models

I.A. Model for self-rated health.

The latent (unobserved) perceived level of general health of individual i, H_i^*, is modeled as an ordered probit model

$$H_i^* = X_i'\beta + \varepsilon_i$$

where X_i are covariates, β are parameters, and ε_i is an individual residual error term, assumed to be standard normal distributed, $\varepsilon_i \sim N(0, 1)$.

Respondent i reports his or her continuous perceived level of general health as category h_i, where h_i is determined as follows:

$$h_i = k \text{ if } \tau_i^{k-1} \leq H_i^* < \tau_i^k ;$$

$$-\infty = \tau_i^0 < \tau_i^1 < \ldots \tau_i^K = \infty;$$

and K = 5 (since there are 5 response categories).

While the latent H_i^* is comparable across respondents, the observed response h_i reflects different use of thresholds (τs) by different respondents, and so is *not* comparable across respondents. The thresholds vary among respondents as a function of covariates Z_i (which may be identical to X_i):

$$\tau_i^1 = \gamma^1 'Z_i$$

$$\tau_i^k = \gamma^{k-1} + e^{\gamma k' Zi}, \qquad k = 2, \ldots, 5; \qquad (1)$$

where γ^k are parameters.

I.B. Model for vignettes.

The anchoring vignette methodology assumes vignette equivalence, that is, that each vignette character has an objective level of general health θ_j (j = 1, 2, 3, or 4, since

we are using 4 severity levels in each vignette series), with perceptions of this level differing among survey respondents only due to (normal random) error:

$$V_{ij}^{*} = \theta_j + u_{ij},$$

$$u_{ij} \sim N(0, \sigma^2).$$

Anchoring vignette methodology also assumes response consistency, i.e., that respondents use the same thresholds to generate observed vignettes ratings as they use to generate their observed self-ratings:

$$v_{ij} = k \text{ if } \tau_i^{k-1} \leq V_{ij}^{*} < \tau_i^{k}$$

where the thresholds are determined by the same γ coefficients as in equations (1) above.

The likelihood function for the chopit model is composed, additively, of the likelihood for the self-assessment [$L_s(\beta, \gamma \mid h)$] and the likelihood for the vignette assessment [$L_v(\theta, \gamma \mid v)$]:

$$L(\beta, \theta, \gamma \mid h, v) = L_s(\beta, \gamma \mid h) \times L_v(\theta, \gamma \mid v).$$

For additional information, see Rabe-Hesketh and Skrondal 2002, and King et al. 2004.

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: hgrol@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu