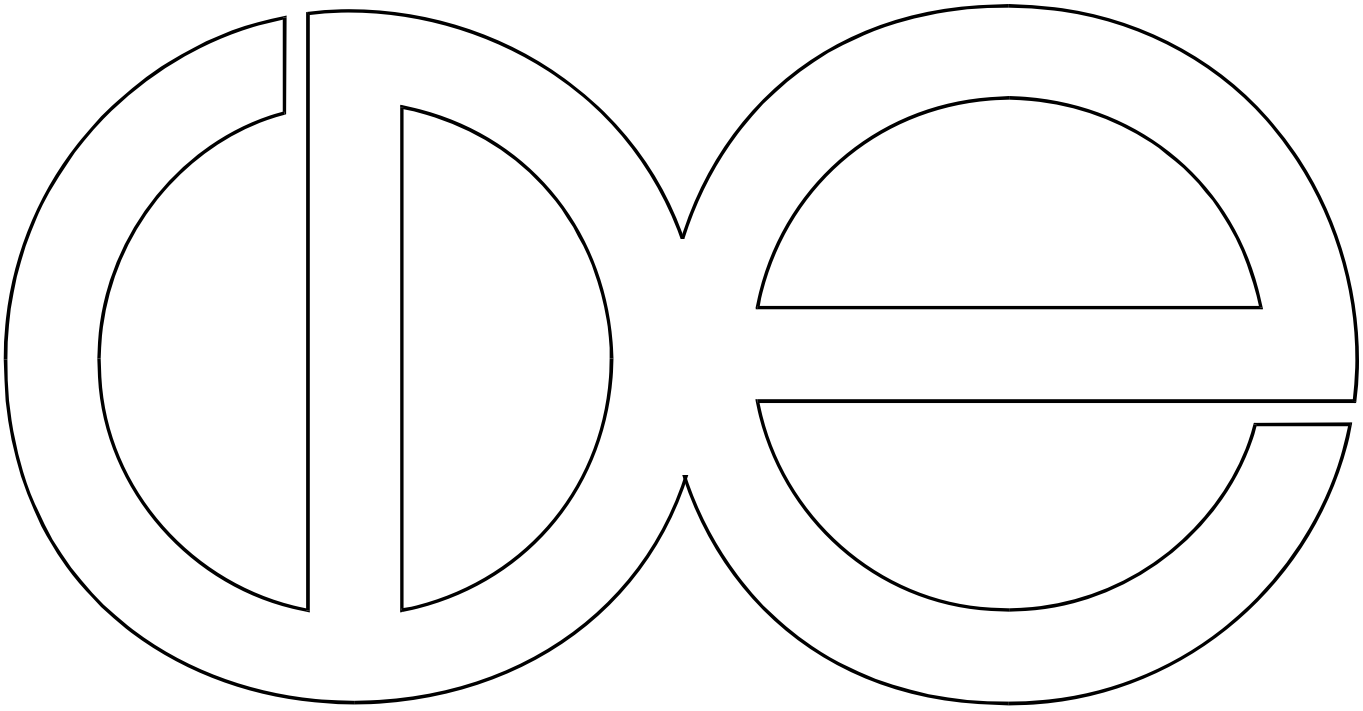
Center for Demography and Ecology

University of Wisconsin-Madison

Age, Sex, and Race Effects in Anchoring Vignette Studies: Methodological and Empirical Contributions

Hanna Grol-Prokopczyk

CDE Working Paper No. 2010-18



Age, Sex, and Race Effects in Anchoring Vignette Studies: Methodological and Empirical Contributions*

Hanna Grol-Prokopczyk

University of Wisconsin-Madison

* This research uses data collected by Time-sharing Experiments for the Social Sciences [NSF Grant 0818839, Jeremy Freese and Penny Visser, Principal Investigators], and is supported 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. The survey instrument was approved by the University of Wisconsin–Madison Social and Behavioral Sciences Institutional Review Board [protocol SE-2009-0278]. I am grateful to Jeremy Freese and Robert M. Hauser for helpful comments about this project. 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).

Age, Sex, and Race Effects in Anchoring Vignette Studies:

Methodological and Empirical Contributions

Abstract: In the past decade, anchoring vignettes have become an increasingly popular tool for identifying and correcting for group differences in use of subjective ordered response categories. However, existing techniques to maximize response consistency (the use of the same standards for self-ratings as for vignette-ratings), which center on matching vignette characters' demographic characteristics to respondents' own characteristics, appear at times to be ineffective or to pose interpretive difficulties. For example, respondents often appear to neglect instructions to treat vignette characters as age peers. Furthermore, when vignette characters are depicted as have the same sex as the respondent, interpretation of observed sex differences in rating style is rendered problematic. This paper applies two experimental manipulations to a national sample (n=1,765) to clarify best practices for enhancing response consistency. First, a comparison of ratings of same- and opposite-sex vignette characters suggests that, with occasional and avoidable exceptions, the sex of the respondent rather than the sex of the vignette character drives observed sex differences in rating style. Second, an analysis of two methods of highlighting vignette characters' age shows that both yield better response consistency than previous, less prominent means. Implications for interpretation and design of anchoring vignette studies are discussed. In addition to methodological contributions, this paper represents the first fielding of general health vignettes to a national sample. Findings show significant differences in health-rating style across racial/ethnic groups, educational categories, and sex. Significant racial/ethnic differences in styles of rating political efficacy are also observed. These findings underscore the incomparability of unadjusted subjective self-ratings across demographic groups.

The past decade has seen a burgeoning of interest in anchoring vignettes as a tool for improving intergroup comparability of survey items. This paper presents experimental findings addressing two methodological questions of interest to users of anchoring vignettes: First, whether to match vignette characters' sex to respondents' sex (and how to interpret subsequent findings about sex differences in rating style), and second, how to optimize vignette wording to encourage respondents to treat vignette characters as age peers. In addition, this paper's empirical findings reveal substantial differences across key demographic groups in how respondents use response categories when rating general health and political efficacy.

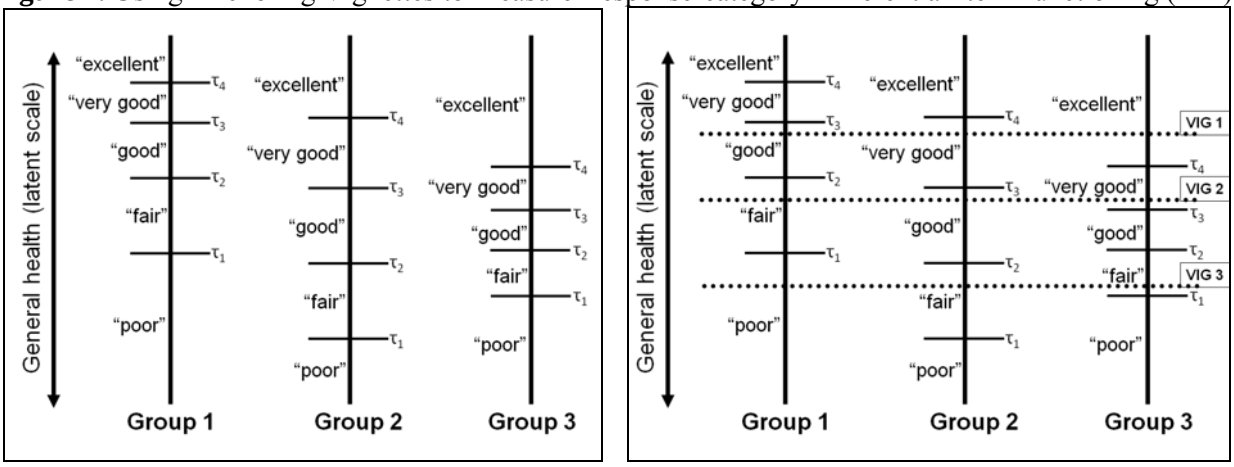
ANCHORING VIGNETTES

Whenever subjective ordered response categories are used in surveys—e.g., "excellent, very good, good, fair, or poor" for self-ratings of health—there is potential that different groups will attribute substantially different meanings to these categories. One group's "good," for example, may represent the same level of health as another group's "fair"; or some groups may be more sparing in use of a given category than are others (see Figure 1, left side). This phenomenon, referred to as "response-category differential item functioning" (DIF) (King et al. 2004) or "reporting heterogeneity" (e.g., Bago d'Uva et al. 2009), is not merely a theoretical possibility: in the context of health ratings alone, evidence supports differences in rating styles across sexes (e.g., Grol-Prokopczyk, Freese, and Hauser 2011), socioeconomic categories (e.g., Dowd and Zajacova 2007), races/ethnicities (e.g., Menec, Shooshtari, and Lambert 2007; Shetterly et al. 1996; Smith 2003), and nationalities (e.g., Iburg et al. 2002; Jürges 2007; Jylhä et

al. 1998; Murray et al. 2002; Sadana et al. 2002; Zimmer et al. 2000). Such reporting heterogeneity can lead to incorrect (and sometimes, highly implausible) research findings (see, e.g., Sadana et al. 2002). In the early 2000s, World Health Organization (WHO) researchers reviewing potential solutions to this problem concluded that anchoring vignettes were "the most promising" of available strategies (Murray et al. 2002:429), and since the formal debut of the method (King et al. 2004), interest in and use of anchoring vignettes has grown dramatically.

Anchoring vignettes are brief texts describing a third-person character who exemplifies a certain level of the trait of interest (e.g., general health). Respondents are asked to rate the character's level of the trait using the same subjective ordered categories that they use for their own self-rating. When the same vignette is given to multiple respondents, the objective level of the trait is being held constant, so differences in ratings can be interpreted as indicative of differences in use of response categories. Typically several vignettes, representing different levels of the trait, are given, and are used to estimate the locations of intercategory thresholds for different groups. By accounting for these different threshold locations, self-ratings can be adjusted to be comparable across individuals or groups, via any of several possible parametric or non-parametric strategies (King et al. 2004; King and Wand 2007; Wand, King, and Lau forthcoming). A schematic diagram of the logic behind the anchoring vignette method is presented in Figure 1; sample vignette texts are shown in Appendix A.

Figure 1. Using Anchoring Vignettes to Measure Response-category Differential Item Functioning (DIF).



Left: Populations or demographic subgroups may differ in how they use subjective ordered response categories. Such "response-category differential item functioning" (DIF) (King et al. 2004), leads to incomparability of responses across groups. Here, members of Group 1 use systematically higher intercategory cutpoints (τ_1 through τ_4) when rating their general health than do members of Group 2. Respondents in Group 3 show a compression of cutpoints relative to the other groups. A level of health rated "good" in Group 1 might thus be considered "very good" in Group 2 and "excellent" in Group 3.

Right: Anchoring vignettes are used to measure and statistically adjust for DIF. Here, respondents from each group receive three vignettes (dotted horizontal lines), each representing a different absolute level of health. Group differences in vignette ratings reveal how each group uses response categories. More formally, vignettes enable estimation of intercategory thresholds (τ's) for each group, which are then adjusted for statistically to permit intergroup comparisons unbiased by DIF.

Since the early 2000s, anchoring vignettes have appeared in numerous studies and surveys, both large and small, national and cross-national (e.g., the 70-country World Health Survey [WHS; http://www.who.int/healthinfo/survey/en/], the Study on Global AGEing and Adult Health [SAGE; http://www.who.int/healthinfo/systems/sage/en/]; the Survey of Health, Ageing and Retirement in Europe [SHARE; http://www.share-project.org/], the Health and Retirement Study [HRS; http://hrsonline.isr.umich.edu/]; and the Wisconsin Longitudinal Study [WLS; http://www.ssc.wisc.edu/wlsresearch/]), and have been applied to domains as diverse as political efficacy, state effectiveness, job satisfaction, women's autonomy, community strength, binge drinking, work disability, health system responsiveness, and specific domains of health such as vision and mobility

(Hopkins and King 2010:202-3; see also examples of vignette-based studies on the Anchoring Vignettes web site: http://gking.harvard.edu/vign/).

Despite its growing popularity, the anchoring vignette method is still relatively new, and advancements continue to be made regarding how to test the method's measurement assumptions (Bago d'Uva et al. 2009; Datta Gupta, Kristensen, and Pozzoli 2010; Rice, Robone, and Smith 2009; van Soest et al. 2007), how to improve upon or adjudicate among strategies for vignette-based adjustment (King and Wand 2007; Wand 2008), and how to optimize vignette wording and implementation (Grol-Prokopczyk et al. 2011; Hopkins and King 2010). The present paper contributes to this latter category of work.

AGE AND SEX OF VIGNETTE CHARACTERS

Two key measurement assumptions are required for the correct functioning of anchoring vignettes: response consistency and vignette equivalence (King et al. 2004:194). Response consistency means that respondents use categories the same way when rating vignette characters as when rating themselves, i.e., they use the same intercategory cutpoints in both situations (rather than holding themselves to higher or lower standards than vignette characters). Vignette equivalence means that all respondents perceive a given vignette as representing the same absolute level of the trait in question (even if differing in the response category they use to describe that level), with vignettes in a series seen as representing points along a unidimensional scale. In the context of the schematic diagram in Figure 1, response consistency means that τ_1 through τ_4 are in the same position for a respondent's vignette ratings as for his or her self-ratings,

and vignette equivalence means that the vignettes can accurately be depicted as flat horizontal lines across all groups of respondents.

To enhance response consistency, respondents are typically encouraged to think of vignette characters as being like themselves in terms of sex, age, and "background." Specifically, vignette characters' sex is often (though not always) matched to respondents' own sex, as recommended by King et al. (2004:194), and instructions introducing vignettes to respondents generally describe the characters as being "of your age and background." (The WHS, SHARE, HRS, and WLS surveys, among others, use this or very similar wording.) However, these aspects of implementation suggest several interpretational or methodological problems.

How to interpret sex differences in vignette ratings?

A number of questions remain unanswered regarding how best to assign and interpret vignette characters' sex. While many surveys consistently sex-match vignettes[1] (e.g., Grol-Prokopczyk et al. 2011), some, for ease of administration, field the same set of

[1] In principle, by matching vignette characters' sex to respondents' sex, the key measurement assumptions of the anchoring vignettes method are put in conflict: response consistency is presumably enhanced, but vignette equivalence may be jeopardized, since respondents are no longer all getting identical vignettes. This is not seen as a problem in existing vignette studies, however, as they assume axiomatically (and tacitly) that vignettes differing in the sex of the character nonetheless represent identical levels of a trait. There appear, then, to be *two* kinds of vignette equivalence assumptions. 1) The first—what is called "vignette equivalence" in existing literature—postulates that all respondents perceive the same absolute value of a trait when looking at a given vignette. 2) The second, introduced here, postulates that a given respondent will perceive the same absolute level of a trait when looking at two vignettes that differ only in the sex (and/or age or background) of the vignette character. That is, if a respondent R rates a male's health differently than an otherwise identical female's health, it is because R uses different cutpoints for the two sexes, not because R sees them as having different underlying levels of health.

We could term these assumptions "cross-respondent vignette equivalence" and "cross-character vignette equivalence," respectively. Though this second assumption is not discussed explicitly in existing vignette literature, it is also crucial, since without it vignette-adjusted self-ratings could not be compared across male and female respondents, or across respondents of different ages.

mixed-sex vignettes to all respondents, male and female (e.g., WHS and SAGE[2]), while others randomly assign each vignette character's sex (e.g., Kapteyn, Smith, and van Soest 2007; van Soest et al. 2007). Is one of these techniques preferable to the others? Can the findings across such studies be compared? Answers to these questions hinge on whether respondents' own sex or vignette characters' sex (or both) drive differences in use of intercategory thresholds.

There are many domains in which the sex of a vignette character could plausibly affect vignette ratings. For example, a male character experiencing pain or fatigue may be rated as having worse health than a woman with identical symptoms, since pain and fatigue are often considered unmanly (Courtenay 2000). Because women traditionally have less political power than men, a woman who meets with an elected official may be rated as having greater political efficacy than a man who does the same. If male characters elicit different ratings than female ones in this manner, then using opposite-sex vignettes could undermine response consistency, as vignettes reveal how, e.g., women rate men, not how they rate themselves.[3]

Even when characters' sex *is* matched to respondents' sex, however, it would be desirable to understand whether observed sex differences in rating style should be interpreted as true differences in how men and women use response categories, or whether the differences are partially or entirely artifacts of vignette characters' sex. Neither case would invalidate vignette-based adjustments, since, when sexes are

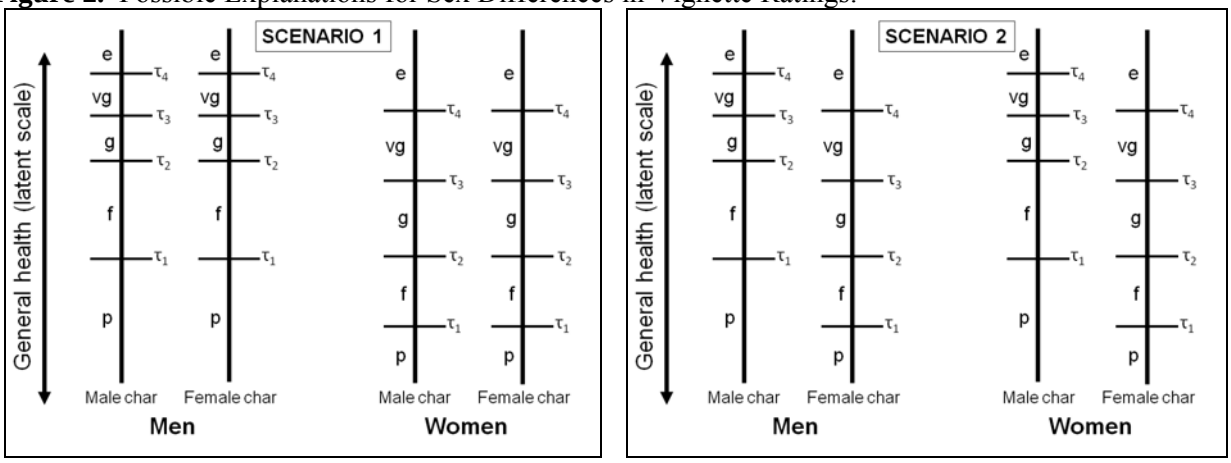
[2] While some documentation suggests that WHS and SAGE sex-match vignette characters, this appears to be in error, as confirmed by WHO researchers responsible for questionnaire design and fielding (Verdes 2011), and as shown in studies using these vignettes (e.g., Rice et al. 2009).

[3] Kapteyn et al.'s (2007) and van Soest et al.'s (2007) suggestion to use a dummy variable indicating vignette characters' sex may help identify and mitigate such threats to response consistency, but is useful only when sex is randomly assigned—in the other cases, characters' sex will be completely collinear with respondents' sex or with vignette severity.

matched, response consistency should not be threatened, but clarifying the interpretation of such ambiguous scenarios would be of interest to scholars studying gender differences in many domains, and would have practical applications even in unrelated survey settings. For example, knowing the relative effect of raters' versus ratees' sex on ratings could help researchers assess and improve the validity of proxy reports about opposite-sex spouses or family members.

To clarify such issues, the current study presents results of an online experiment in which respondents were randomly assigned to receive same-sex or opposite-sex vignette characters. The experimental data are used to compare two idealized scenarios, depicted visually in Figure 2: 1) Scenario 1: Respondents' sex, but not vignette character's sex, drives observed sex differences in rating style. In this case, sex differences in vignette ratings are truly a reflection of women's and men's different styles of evaluation; proxy ratings of opposite-sex family members will be biased due to these different evaluation styles; matching vignette characters' sex to respondents' sex is optional, since it has no bearing on response consistency; and results from sex-matching and non-sex-matching designs can be unproblematically compared. 2) Scenario 2: Vignette character's sex alone drives observed sex differences in rating style. In this case, men and women do not truly differ in their evaluation styles; proxy ratings by opposite-sex family members will not be biased; and matching vignette characters' sex to respondents' sex is *crucial* for response consistency. (The possibility that both respondents' and vignette characters' sex affect vignette ratings, perhaps interactively, is also considered.)

Figure 2. Possible Explanations for Sex Differences in Vignette Ratings.



Scenario 1: Respondents' sex (but not vignette characters' sex) affects ratings of health vignettes.
Scenario 2: Vignette characters' sex (but not respondents' sex) affects ratings of health vignettes.
Note: Response categories "excellent, very good, good, fair, poor" are here abbreviated by first letters. Implications of each scenario are described in the main text.

How to address age-related response inconsistency?

Another methodological question pertaining to the presentation of vignette characters is how to better encourage age-related response consistency, in light of recent findings that respondents appear to neglect instructions to treat vignette characters as age peers. Grol-Prokopczyk et al. (2011), for example, find that older adults in the WLS rate general health vignettes more "health-pessimistically" (i.e., using more negative response categories) than younger adults. Not only is this finding inconsistent with the predictions of previous literature (e.g., Groot 2000; Idler 1993; van Doorslaer and Gerdtham 2003), but it leads to the implausible result that, when self-rated health is adjusted based on vignette ratings, health appears to *not* deteriorate with age. Datta Gupta et al. (2010) present similar findings based on SHARE's work disability vignettes, and take the extra step of formally testing whether the findings represent a violation of response consistency. They conclude that, indeed, in a model relaxing the response consistency assumption, age dummies show the expected sign (p. 859). It appears, then, that existing

instructions regarding vignette characters' age may not be sufficiently prominent, so that older adults rate vignette characters as though they were younger than themselves, i.e., using higher standards for health.

To attempt to address this problem, this study analyzes two different forms of item wording: one describing vignette characters in prominent and succinct opening instructions as "people your age", and one explicitly presenting *each* characters' age (e.g., "John, age 65, …."), using the multiple of 5 nearest to the respondents' own age. Do either or both of these approaches improve age-related response consistency relative to previous studies?

Racial/ethnic differences in health-rating style

The general health anchoring vignettes created by Grol-Prokopczyk et al. (2011), intended to calibrate the widely-used general self-rated health (SRH) item, have previously been fielded only in a racially homogeneous (White) and geographically limited sample. However, a number of studies provide evidence, albeit indirectly, that some racial/ethnic groups are more "health-pessimistic" in subjective health reports than others—e.g., Hispanics compared to Whites (Shetterly et al. 1996; cf. Menec et al. 2007; Turner and Avison 2003). Other evidence suggests non-trivial racial/ethnic differences in tendency to use extreme response categories, across substantive domains (specifically, African-Americans and Hispanics may use extreme categories more often, and Asians less often, than Whites [see Smith 2003:82]).

The current study, by fielding the general health vignettes to a nationally-representative sample, provides the first opportunity to use anchoring vignettes to directly

identify racial/ethnic differences in use of response categories when rating health. Given the ubiquity of the SRH item, its strong correlation with objective measures of health (e.g., Jylhä, Volpato, and Guralnik 2006), and its well-documented power to predict mortality (see, e.g., DeSalvo et al. 2006; Idler and Benyamini 1997), racial/ethnic differences in use of the item's response categories would likely be of interest to many researchers. This paper thus explores differences in rating style (of general health, and also of a second domain, political efficacy) by race/ethnicity, as well as by other key demographic categories, including sex, age, and education.

ANALYTIC GOALS

To summarize, the three primary analytic goals of this paper are:

1. To experimentally test whether sex differences in rating style are driven by true differences in how men and women use response categories or simply by differences in how male and female vignette characters are evaluated.
2. To experimentally compare the effects of two different forms of wording regarding vignette character's age on rating style (and specifically to assess whether one or both forms appear to overcome problems with age-related response inconsistency reported in previous literature).
3. To extend previous empirical work using anchoring vignettes by identifying differences in rating style across racial/ethnic groups, as well as across other key sociodemographic categories, in the domains of health and political efficacy.

As a whole, this paper contributes to the anchoring vignette literature by identifying and clarifying practices to maximize vignette validity, and by broadening the scope of empirical findings on sociodemographic predictors of reporting heterogeneity.

DATA AND METHODS

Survey sample

Data collection was sponsored by a Time-sharing Experiments for the Social Sciences (TESS) award (http://www.tessexperiments.org/), and fielded by Knowledge Networks (http://www.knowledgenetworks.com/). Knowledge Networks recruits respondents to its nationally-representative (American) "KnowledgePanel" using a dual sampling strategy of random-digit dial (RDD) and address-based sampling, to ensure adequate coverage of respondents likely to be undercovered by RDD alone, e.g., cell phone-only households. After recruitment, respondents are provided with Internet access and necessary hardware, if needed, to allow all respondents to participate in online surveys. (Respondents who already have Internet access receive incentive points, redeemable for cash, to encourage survey completion.) Respondents are asked to complete 4-6 online surveys per month, and receive notice of new surveys by email, allowing them to participate from home and at the time of their choosing. Participants may skip up to seven consecutive surveys without risk of removal from the KnowledgePanel.

The current Web-based survey was fielded in June 2010 to 2,750 Knowledge-Panel respondents, of whom 1,771, or 64.4%, completed the survey. Of these, six respondents who did not answer any vignette questions were dropped from the sample,

leaving a working sample size of 1,765. Non-response rates for individual vignette questions ranged from .45%-1.25%. Descriptive characteristics of the analytic sample are shown in Table 1.

Table 1. Descriptive statistics for analytic sample (n=1,765).

	Proportion or Mean	Standard Deviation	N
Female	.51		1,765
Age in years	48.76	16.69	1,765
Education			
Less than high school	.11		194
High school degree	.28		499
Some college	.30		521
Bachelor's degree or higher	.31		551
Household income			
Less than $24,999	.20		359
$25,000 to $49,999	.26		458
$50,000 to $84,999	.28		490
$85,000 or higher	.26		458
Marital status			
Currently married	.52		921
Separated/Divorced/Widowed	.18		326
Never married	.21		376
Cohabiting	.08		142
Race/ethnicity			
White, non-Hispanic	.77		1,353
Black, non-Hispanic	.09		151
Hispanic	.08		139
Other, including two or more races	.07		121

Vignette texts and experimental manipulations

Each respondent was presented with the four general health vignettes and three political efficacy vignettes shown in Appendix A. These vignette series were designed to calibrate key measures in health research (Grol-Prokopczyk et al. 2011) and political science (Hopkins and King 2010:208), respectively. They were included in the present

study because they represent two very different substantive domains, and because they have been previously validated for adherence to measurement assumptions (Grol-Prokopczyk et al. 2011; Hopkins and King 2010; King and Wand 2007). Vignette ratings were reverse-coded to permit intuitive interpretation, i.e., so that higher ratings indicate better health or greater political efficacy. The order of the two series of vignettes, as well as the order of items within each set, was randomly determined for each respondent.

Assignment to experimental conditions was also random: half of respondents received vignettes with male names, and half vignette with female names (shown in Appendix A). To encourage response consistency, names in the vignettes were selected from the top-ten most common names reported on the 1990 U.S. Census (U.S. Census Bureau 2008). Furthermore, half of respondents received vignettes giving each character's exact age (the "explicit age" condition), where this age was set to be the multiple of five nearest to the respondent's own age; half received vignettes where characters' age was suggested only implicitly in the opening instructions (e.g., "What follows are descriptions of the health of some people your age"). Appendix B compares the opening instructions for the "explicit age" and "no explicit age" conditions; for comparison, it also shows the wording used for the general health vignettes in the Wisconsin Longitudinal Study.

Analytic strategy

Analyses consisted of ordered probit regressions of vignette ratings on key demographic variables (sex, age, education, income, marital status, and race/ethnicity) and on flags of experimental conditions, to identify which factors predict differences in

ratings of vignettes. To explore whether men and women are differently affected by the sex of the vignette character, models including interactions between respondents' sex and vignette character's sex were also examined. Vignette were analyzed both individually and pooled within a series; in the latter case, controls for vignette severity were included among the independent variables.

In addition, "hopit" (hierarchical ordered probit) models were used to identify factors predicting differences in intercategory threshold locations (as described in Rabe-Hesketh and Skrondal 2002; cf. King et al. 2004:198).[4] Unlike standard ordered probit models, which assume identical response-category thresholds for all respondents, hopit models allow cutpoints to vary across groups, based on the groups' ratings of anchoring vignettes. Formally—and using general health as an example—respondent *i* reports his or her perceived level of vignette character *j*'s health (V_{ij}*) as category v_{ij}, where v_{ij} is determined as follows:

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

$$-\infty = \tau_i^{0} < \tau_i^{1} < \ldots \tau_i^{K} = \infty.$$

The thresholds (τ_i^{1} though τ_i^{K}) vary among respondents as a function of Z_i, where Z_i is a vector of covariates (in our case, comprising measures of sex, age, education, income, marital status, and race/ethnicity) and γ^k represents the corresponding parameters:

$$\tau_i^{1} = \gamma^{1\prime} Z_i$$

$$\tau_i^{k} = \tau^{k-1} + e^{\gamma k' Z i}, \qquad k = 2, \ldots, K. \qquad \text{(Equation 1)}$$

[4] Some authors refer to this model as "chopit", with the "c" standing for "compound", though more often this usage is reserved for cases where multiple ratings of each vignette allow for calculation of individual-level random effects (which is not the case here).

All statistical analyses were done in Stata SE/11.1. Hopit was implemented using the `gllamm` program (www.gllamm.org), as described by Rabe-Hesketh and Skrondal (2002). Stata code for all analyses is available from the author upon request.

RESULTS

Table 2 shows mean ratings of the general health and political efficacy vignettes. Ratings of both series decrease/increase monotonically in the expected direction. The standard deviation for health vignette 4—the vignette describing the least healthy vignette character—is noticeably smaller than for other vignettes in the series (0.66 versus 0.82-0.88), suggesting a possible floor effect of response categories.

Table 2. Mean ratings of anchoring vignettes.

	Vignette 1	Vignette 2	Vignette 3	Vignette 4
General Health	4.17	3.10	1.98	1.48
	(.85)	(.88)	(.82)	(.66)
Political Efficacy	2.16	2.32	2.95	n/a
	(.82)	(.78)	(.78)	

Note: Means calculated by assigning scores to general health ratings of 1 = poor; 2 = fair; 3 = good; 4 = very good; 5 = excellent, and to political efficacy ratings of 1= no say at all, 2 = little say, 3 = some say, 4 = a lot of say. Standard deviations in parentheses.

Table 3 presents results of ordered probit regressions of vignette ratings (pooled within a series) on experimental conditions and key demographic variables, revealing which factors predict higher or lower vignette ratings.[5] Regressions of individual (rather than pooled) vignette ratings on the same variables yielded similar results, except where

[5] The parallel regression assumption is not met in these models, meaning that independent variables' effects are not constant across all binary pairings of response categories. Nonetheless, these models constitute a largely accurate summary of findings, as they provide parameter estimates consistent in terms of direction and statistical significance with those obtained from binary response models (not shown due to space constraints). Furthermore, the model shown in Appendix C *does* show the effects of independent variables separately for each cutpoint, providing a more fine-grained picture of how demographic covariates predict differences in rating style across the health and political efficacy spectrums.

Table 3. Ordered probit regression of vignette ratings on demographic variables.

	General Health series	Political Efficacy series
Male vignette character	-.060*	.049
	(.027)	(.030)
Explicit mention of character's age	.030	.024
	(.027)	(.030)
Female respondent	.143***	.062*
	(.027)	(.030)
Age 30-44	.004	-.003
	(.045)	(.051)
Age 45-59	-.021	.104*
	(.044)	(.050)
Age 60 and above	-.070	.061
	(.048)	(.054)
Less than high school degree	-.164**	.099
	(.049)	(.055)
Some college	.062	.033
	(.035)	(.040)
Bachelor's degree or higher	.143***	.157***
	(.037)	(.042)
HH Income: $25,000 to $49,999	-.109**	-.090*
	(.040)	(.046)
HH Income: $50,000 to $84,999	-.052	-.094
	(.042)	(.048)
HH Income: $85,000 or higher	-.085	-.085
	(.045)	(.051)
Separated/Divorced/Widowed	.002	-.032
	(.039)	(.044)
Never married	-.116**	-.014
	(.040)	(.045)
Cohabiting	-.055	.024
	(.053)	(.060)
Black, non-Hispanic	-.423***	.397***
	(.050)	(.056)
Hispanic	-.279***	.349***
	(.052)	(.059)
Other, including two or more races	-.080	.120*
	(.053)	(.061)
N	1,757	1,749

Note: * $p < .05$, ** $p < .01$, *** $p < .001$, two-tailed. Standard errors in parentheses. Higher vignette ratings indicate better perceived health or greater perceived political efficacy. Omitted reference categories: "Male respondent," "Age 18 to 29," "High school degree," "Less than $24,999," "Currently married," and "White, non-Hispanic." Models also include controls for vignette severity, not shown.

noted below. Versions of each model including an interaction between respondent's sex and character's sex were also analyzed, but because the interaction term was never

statistically significant, it was excluded from the presented analyses. Referring to Table 3, we now present results from the two experimental manipulations, followed by findings regarding demographic predictors of differences in rating style.

Results: Does vignette character's sex affect vignette ratings?

Table 3 indicates that, in the general health vignettes, male vignette characters receive lower health ratings than do female ones with identical symptoms (β = -.060; p = .024). However, this effect is driven entirely by the lowest severity health vignette, Severity 1. In analyses of individual vignettes, only this one shows a significant effect of character's sex on ratings (β = -.107; p = .045), and in a pooled analysis excluding this vignette, the relationship is no longer statistically significant (β = -.042; p = .175). Character's sex may be relevant in this vignette but not others because of its mention of "headaches," which afflict women much more than men (e.g., Fillingim et al. 2009; Kroenke and Spitzer 1998:152). A man who *does* have a headache may therefore be rated as having worse health than a women with the same ailment. In contrast, other vignettes in the health series do not mention specific health complaints, and thus appear less likely to elicit such gendered associations. This finding is consistent with Angelini, Cavapozzi, and Paccagnella (2010), who find in a series of work disability vignettes—which include mention of back pain and depression, both of which are substantially more common among women than men (Fillingim et al. 2009; Wetzel 1994)— that "the same [vignettes] are considered less severe for a woman than for a man."

As mentioned above, there was no significant interaction between respondent's and character's sex, for any individual or pooled vignettes. Men and women appear

equally inclined to rate the General Health Severity 1 vignette more negatively when the character is male.

Table 3 also shows that there was no significant effect of vignette characters' sex on ratings of political efficacy vignettes (as was also the case when the vignettes were analyzed individually). In sum, an effect of character's sex was found in only one of the seven vignettes included in this study, and it seems plausible that even this lone effect could have been avoided with different vignette wording, i.e., by not mentioning a gendered health condition (headaches).

Results of age experiment

As shown in Table 3, no significant differences were found between vignettes mentioning each character's exact age and vignettes describing characters in opening instructions as "people your age." This was true in both series of vignettes, and whether analyzed individually or pooled. Furthermore, the problem of age-related response inconsistency reported in Grol-Prokopczyk et al. (2011)—in which older adults gave more negative ratings of general health vignettes—was not replicated, even when the analysis was restricted to White, non-Hispanic respondents aged 60 or more to better resemble the original WLS sample (not shown). This may reflect the fact that the WLS instructions were wordier than the current "no explicit age" instructions (65 versus 28 words; see Appendix B), or that they were understood less well because they were given orally over the telephone, rather than appearing written on a screen. Another possibility is that respondent fatigue was a greater issue in the WLS than in the current study, since

in the WLS the vignettes appeared at the end of the survey's sixth module, rather than as a stand-alone instrument.

At face value, then, the present findings suggest that respondents are as likely to treat vignette characters as age peers when the characters are described once as "people your age" as when each character's numeric age is given explicitly. However, given that findings may differ in oral survey situations, or when respondent fatigue is high, explicit mentioning of characters' age may be the preferred option, since it does not rely on careful attention to opening instructions to avoid age-related response inconsistency.

Results: Group differences in use of response categories

The remaining parameter estimates in Table 3 show how demographic factors predict differences in styles of rating health and political efficacy. For the general health vignettes, women gave systematically higher ratings than did men ($\beta = .143$; $p < .001$). (This significant sex difference was found for all individual health vignettes except Severity 4 [$\beta = .029$; $p = .612$]. It is unclear whether this indicates that men and women's ratings converge when health states are very poor, or whether this is an artifact of category floor effects.) A similar though weaker effect of respondent's sex was found for the political efficacy vignettes. These findings, paired with the largely null findings regarding effects of *character's* sex, suggest that sex differences in rating style are, at least in these domains, driven primarily by respondents' rather than vignette characters' sex. This corresponds to Scenario 1 in Figure 2. Previous findings that women are more "health-optimistic" than men thus appear correct, and not mere artifacts of sex-matching

practices (though for the General Health Severity 1 vignette, sex differences may be exaggerated by sex matching).

As mentioned, respondent's age did not appear to affect ratings of health vignettes (a finding confirmed by a Wald test of the joint significance of the relevant dummies), though for political efficacy, respondents aged 45 to 59 did give significantly higher ratings than those under 30 ($\beta = .104$; $p = .037$). In both vignette series, respondent's education showed a positive (and, for health, roughly linear) effect on vignette ratings, with, e.g., college graduates giving substantially higher ratings than high school graduates ($\beta = .143$, $p < .001$ for health; $\beta = .157$, $p < .001$ for political efficacy). Also in both series, higher levels of income predicted slightly lower vignette ratings, though this association was only marginally significant for those with incomes of $50,000 and up. Never-married respondents ranked health vignettes more health-pessimistically than currently married respondents ($\beta = -.116$, $p = .005$), while marital status was unrelated to ratings of political efficacy.

Finally, racial/ethnic differences in rating styles were observed in both vignette series, and in both were substantively large, with the "non-Hispanic Black" and "Hispanic" dummy variables yielding the largest coefficients in each model. The parameter estimates for "Black, non-Hispanic," for example ($\beta = -.423$, $p < .001$ for health, and $\beta = .397$, $p < .001$ for political efficacy), were at least twice the size of any others in the respective models, including respondent's sex and college degree. (Such associations were observed consistently across all individual political efficacy vignettes, and across all health vignettes except Severity 4, which, as above, may reflect category floor effects.) However, while non-White status predicted more negative ("pessimistic")

ratings of health, it predicted more *positive* ("optimistic") ratings of political efficacy. The effects of race/ethnicity on rating style, then, appear to not take the form of general optimism/pessimism, but rather to be context-dependent.

Overall, the results shown here for general health vignettes are strikingly similar to those reported in Grol-Prokopczyk et al. (2011) (including in terms of the significant effects of respondent's sex and education on health ratings, the lack of a sex difference for the General Health Severity 4 vignette, etc.), but extend those findings by including covariates related to marital status and race/ethnicity. The current data thus provide a more complete picture of how key demographic groups differ in their manner of rating health.

Results: Differences in intercategory cutpoint locations across groups

Table 3 summarized how demographic groups differ in their ratings of vignettes. In turn, Appendix C presents results of a hopit model that uses such differences in vignette ratings to estimate intercategory cutpoint locations by group. For example, in the General Health series, the first cutpoint (τ_1; that between "poor" and "fair") is estimated to be significantly lower for female respondents than for males (β = -.133; $p < .001$). This means that women are less likely to choose a health rating of "poor" rather than "fair", i.e., they are more health-optimistic than men (consistent with what was described above). Coefficients for this first cutpoint also demonstrate the aforementioned greater health-optimism of more highly educated respondents, and the greater health-pessimism of non-Whites. Appendix C also shows that Cutpoint 1's

location differs across educational and racial/ethnic categories in the context of political efficacy.

Cutpoints beyond the first defy such straightforward interpretation, since they are based additively on previous cutpoints and involve exponentiation of coefficients (see Equation 1 above). Estimated cutpoint locations may thus best be presented visually. Figures 3a through 3c apply cutpoint coefficients from the hopit model to the analytic sample, to generate estimated cutpoint locations by sex, level of education, and race/ethnicity for health, and by race-ethnicity for political efficacy. (Political efficacy cutpoints differ only trivially by sex and education, and thus are not pictured.)

The y-axis units in these graphs are standard deviations (SDs) of the relevant self-rating (health or political efficacy). Thus, Figure 3a shows us that, when rating general health, women in our sample use intercategory cutpoints that are approximately .15 SD units lower than men's ($p < .001$ for all cutpoints). The difference across educational categories (Figure 3b) is larger, with college-degree holders using cutpoints approximately .35 SD units lower than respondents who did not complete high school.

Figure 3c shows that, for both of the tested domains, differences in cutpoint locations across racial/ethnic groups are larger still, averaging around a .4 unit difference between Whites and Blacks, and for some cutpoints reaching nearly .6 units. This figure also shows clearly that while non-Whites generally have higher cutpoints than Whites for health, the pattern is reversed in the context of political efficacy. While none of the group differences in cutpoint locations presented here are extremely large, they do represent non-trivial sources of measurement bias, which, depending on the application, could potentially lead to incorrect or misleading research findings.

Figure 3a. Estimated Cutpoints, General Health, By Sex.

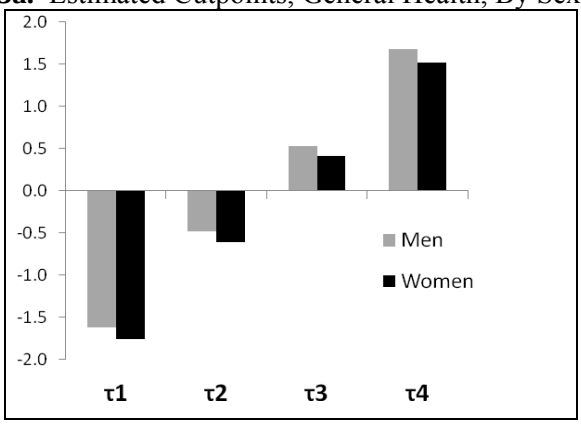


Figure 3b. Estimated Cutpoints, General Health, By Education.

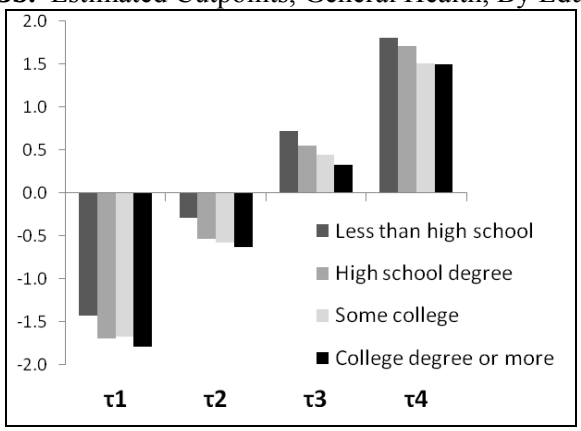
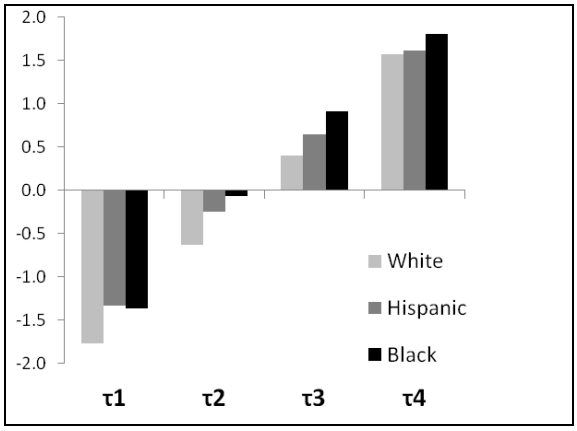
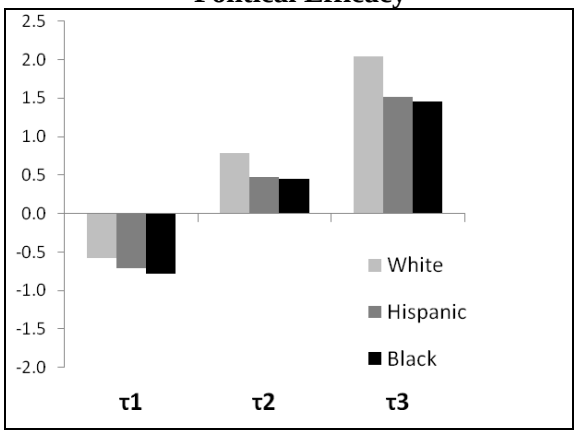


Figure 3c. Estimated Cutpoints for General Health (Left) and Political Efficacy (Right), By Race/Ethnicity.

General Health



Political Efficacy



Note for Figures 3a-3c: Cutpoints are estimated by applying coefficients from the hopit model in Appendix C to the analytic sample. Y-axis units are standard deviations of respondents' self-ratings of general health (SRH), or, in Figure 3c right, of political efficacy.

DISCUSSION

In the past decade, anchoring vignettes have grown in popularity as a tool to correct for individual and group differences in use of subjective response scales. Their ability to serve this purpose, however, depends crucially on adherence to key measurement assumptions, including response consistency (RC). Such adherence cannot be taken for granted: while a number of studies find evidence generally supportive of adherence to RC (e.g., Grol-Prokopczyk et al. 2011, van Soest et al. 2007), others find the opposite (e.g., Bago d'Uva et al. 2009, Datta Gupta et al. 2010), or find mixed evidence (e.g., Grol-Prokopczyk, McEniry, and Verdes 2011). In this light, careful attention to details of anchoring vignette implementation is warranted, to maximize the chances that vignettes will function as intended.

This project used two experimental manipulations to clarify how to maximize sex- and age-related response consistency when using anchoring vignettes (and also to clarify interpretation and comparison of research findings based on the design choices made). Results show that, first, most sex differences in ratings of vignettes—at least in the tested domains of general health and political efficacy—are driven by true differences in how men and women use response categories, rather than by the sex of the depicted vignette characters. The sole observed exception to the pattern occurred when a highly gendered health symptom (headaches) was mentioned in a health vignette. To avoid such situations in the future, researchers may strive to avoid mention of health (or other) conditions with clearly gendered distributions or connotations.

These findings suggest that matching vignette characters' sex to respondents' sex is optional, as it is unlikely to have substantial effects on response consistency (as long as gendered conditions are avoided in vignette texts). It also appears that findings from studies differing in how they assign vignette characters' sex can be fairly compared, and that use of mixed-sex vignettes within a single survey does not introduce significant bias. At the same time, the findings suggest that proxy ratings given by opposite-sex family members or other opposite-sex respondents *are* likely to be biased due to men and women's different evaluation styles, and thus should be interpreted with caution (or adjusted statistically, potentially using vignettes).

Results of the second experimental manipulation suggest that *both* tested techniques of conveying vignette characters' age, i.e., using clear opening instructions and explicitly mentioning age in vignette texts themselves, can effectively improve age-related response consistency. However, given previously reported challenges with this form of response consistency, it may be preferable to use the explicit mention of characters' age when possible, in case contextual factors (such as respondent fatigue) lead to poor attention to vignette instructions.

Future researchers may wish to verify that the current experimental findings hold in other substantive domains—though the similarity of results across domains as different as health and political efficacy suggests at least a certain generalizability across substantive areas.

In addition to presenting experimental results, this study confirms and extends previous empirical findings of non-trivial differences in use of response categories across key demographic groups. In particular, women appear to give systematically higher

(more "optimistic") vignette ratings than men, especially in the context of general health. A similar association is found for education, with more highly educated respondents giving higher ratings (i.e., using lower intercategory thresholds) than less educated ones. Never-married respondents appear more health-pessimistic than their married peers, while marital status appears unrelated to style of rating political efficacy. Finally, and most strikingly, differences in rating style across racial/ethnic groups appear substantively large in both tested domains. When rating health, (non-Hispanic) Blacks used substantially higher category thresholds than (non-Hispanic) Whites, with Hispanics generally appearing in an intermediate position. That is, non-Whites are more "health-pessimistic" than Whites. When rating political efficacy, however, Black respondents used substantially *lower* thresholds than Whites, with Hispanics tracking quite closely to Blacks. Non-Whites may thus be relatively optimistic in the context of political efficacy. Such differences in rating styles may have important implications for health researchers studying racial/ethnic health disparities (who may find exaggerated racial/ethnic health differences if they use unadjusted self-ratings), and for political scientists studying group differences in political behavior and belief.

This study is the first to use anchoring vignettes to test for and demonstrate racial/ethnic differences in styles of rating general health (i.e., by using vignettes specifically designed to calibrate the general self-rated health [SRH] question. Previous fieldings of health vignettes to national samples have represented only specific domains of health, such as mobility). Given the relatively small number of non-Whites in the present sample, and the possibility that non-response bias leads them to differ in some significant ways from the non-White population as a whole, these findings invite attempts

at replication (and at validation. In particular, verification of response consistency among non-Whites would bolster confidence that anchoring vignettes could correctly adjust for racial/ethnic differences in rating style). The current findings are, however, consistent with previous studies suggesting greater "health-pessimism" among non-Whites (e.g., Shetterly et al. 1996). Overall, the present study underscores the incomparability of unadjusted subjective self-ratings across demographic groups, and supports the need for survey tools such as anchoring vignettes to adjust for reporting heterogeneity.

REFERENCES

Angelini, Viola, Danilo Cavapozzi, and Omar Paccagnella. 2010. "Dynamics of work disability reporting in Europe". Paper presented at the Royal Statistical Society conference on Anchoring Vignettes in Social Science Research, London, U.K., 17 November. Retrieved 27 January 2011 from http://membership.rss.org.uk/main.asp?group=&page=1321&event=1194&month=11&year=2010&date=17%2F11%2F2010.

Bago d'Uva, Teresa, Maarten Lindeboom, Owen O'Donnell, and Eddy van Doorslaer. 2009. "Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity." HEDG Working Paper 09/30. Retrieved 20 January 2011 from http://www.york.ac.uk/res/herc/documents/wp/09_30.pdf.

Courtenay, Will H. 2000. "Constructions of Masculinity and Their Influence on Men's Well-Being: A Theory of Gender and Health". *Social Science & Medicine* 50:1385-1401.

Datta Gupta, Nabanita, Nicolai Kristensen, and Dario Pozzoli. 2010. "External Validation of the Use of Vignettes in Cross-Country Health Studies." *Economic Modelling* 27:854-865.

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.

Fillingim, Roger B., Christopher D. King, Margarete C. Ribeiro-Dasilva, Bridgett Rahim-Williams, and Joseph L. Riley. 2009. "Sex, Gender, and Pain: A Review of Recent Clinical and Experimental Findings." *The Journal of Pain* 10(5): 447–485.

Grol-Prokopczyk, Hanna, Jeremy Freese, and Robert M. Hauser. 2011. "Using Anchoring Vignettes to Assess Group Differences in Self-Rated Health." *Journal of Health & Social Behavior* 52(2):246-261.

Grol-Prokopczyk, Hanna, Mary McEniry, and Emese Verdes. 2011. "Categorical Borders Across Borders: Can Anchoring Vignettes Identify Cross-National Differences in Health-Rating Style?" Paper presented at the Population Association of America annual conference in Washington, D.C., 2 April. Manuscript available from first author.

Groot, Wim. 2000. "Adaptation and Scale of Reference Bias in Self-Assessments of Quality of Life." *Journal of Health Economics* 19:403-420.

Hopkins, Daniel J. and Gary King. 2010. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly* 74(2): 201–222.

Iburg, Kim Moesgaard, Joshua A. Salomon, Ajay Tandon, and Christopher J. L. Murray. 2002. "Cross-Population Comparability of Physician-Assessed and Self-Reported Measures of Health". Ch. 8.4 (pp. 433-448) in Murray, Christopher J.L., Joshua A. Salomon, Colin D. Mathers, and Alan D. Lopez (eds.), *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. 2002. Geneva: World Health Organization.

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.

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. "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." *The American Economic Review* 97(1):461-473.

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(1):191-207.

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

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

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.

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". Ch. 8.3 (pp. 421-431) in Murray, Christopher J.L., Joshua A. Salomon, Colin D. Mathers, and Alan D. Lopez (eds.), *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. 2002. Geneva: World Health Organization.

Rabe-Hesketh, Sophia and Anders Skrondal. 2002. "Estimating chopit models in gllamm: Political efficacy example from King et al." Retrieved 20 January 2009 from http://www.gllamm.org/chopit.pdf.

Rice, Nigel, Silvana Robone, and Peter Smith. 2009. "Analysis of the Validity of the Vignette Approach to Correct for Heterogeneity in Reporting Health System Responsiveness." HEDG Working Paper 09/28. Retrieved 20 January 2011 from http://www.york.ac.uk/res/herc/documents/wp/09_28.pdf.

Sadana, Ritu, Colin D. Mathers, Alan D. Lopez, Christopher J. L. Murray and Kim Moesgaard Iburg. 2002. "Comparative Analyses of More than 50 Household Surveys on Health Status". Ch. 8.1 (pp. 369-386) in Murray, Christopher J.L., Joshua A. Salomon, Colin D. Mathers, and Alan D. Lopez (eds.), *Summary*

Measures of Population Health: Concepts, Ethics, Measurement and Applications. Geneva: World Health Organization.

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.

Smith, Tom W. 2003. "Developing Comparable Questions in Cross-National Surveys." Ch. 5 (pp. 69-91) in Janet A. Harkness, Fons J. R. van der Vijver, and Peter Ph. Mohler (eds.), *Cross-Cultural Survey Methods*. Hoboken: John Wiley & Sons.

Turner, R. Jay, and William R. Avison. 2003. "Status Variations in Stress Exposure: Implications for the Interpretation of Research on Race, Socioeconomic Status, and Gender." *Journal of Health and Social Behavior* 44(4):488-505.

U.S. Census Bureau, Population Division. 2008. "Frequently Occurring First Names and Surnames From the 1990 Census". Available at http://www.census.gov/genealogy/names/ [last accessed 10 March 2009].

Van Doorslaer, Eddy, and Ulf-G. Gerdtham. 2003. "Does Inequality in Self-Assessed Health Predict Inequality in Survival by Income? Evidence from Swedish Data." *Social Science and Medicine* 57:1621–1629.

van Soest, Arthur, Liam Delaney, Colm Harmon, Arie Kapteyn, and James P. Smith. 2007. "Validating the Use of Vignettes for Subjective Threshold Scales." IZA Discussion Paper No. 2860. Retrieved 20 January 2011 from ftp://repec.iza.org/RePEc/Discussionpaper/dp2860.pdf.

Verdes, Emese. 2011 (10 March). Personal communication via email.

Wand, Jonathan. 2008 (28 April). "Evaluating the Credibility of Interpersonal Comparisons using Survey Experiments and Benchmarks." Unpublished manuscript.

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

Wetzel, Janice Wood. 1994. "Depression: Women-at-Risk." Pp. 85-108 in Olson, Miriam Meltzer (ed.), *Women's Health and Social Work: Feminist Perspectives*. Binghamton, NY: The Haworth Press.

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.

APPENDIX A: Texts of vignettes and self-assessments.

General Health, Severity 1	[Barbara/David][, age XX,] is energetic, and has no trouble with bending, lifting, and climbing stairs. [She/he] rarely experiences pain, except for minor headaches. In the past year [Barbara/David] spent one day in bed due to illness. In general, would you say [Barbara/David]'s health is: excellent, very good, good, fair, or poor?
General Health, Severity 2	[Jennifer/John][, age XX,] is usually energetic, but once in a while feels fatigued. [S/he] has very slight trouble bending, lifting, and climbing stairs. [His/her] occasional pain does not affect [his/her] daily activities. In the past year, [Jennifer/John] spent two days in bed due to illness. In general, would you say [Jennifer/John]'s health is: excellent, very good, good, fair, or poor?
General Health, Severity 3	About once a week, [Mary/Michael][, age XX,] 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, [Mary/Michael] spent a week in bed due to illness. In general, would you say [Mary/Michael]'s health is: excellent, very good, good, fair, or poor?
General Health, Severity 4	[Susan/Richard][, age XX,] 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, [Susan/Richard] spent a few nights in a hospital, and over a week in bed due to illness. In general, would you say [Susan/Richard]'s health is: excellent, very good, good, fair, or poor?
General Health Self-Assessment	In general, would you say your own health is excellent, very good, good, fair, or poor?
Political Efficacy, Level 1	[Elizabeth/James][, age XX,] is concerned about cars speeding by [his/her] house, and [he/she] would like to see the speed limit on [his/her] street reduced. However, [he/she] knows that [his/her] local elected official is from another part of town, and so is very unlikely to help him/her. How much say do you think [Elizabeth/James] has in getting [his/her] local government to consider issues that interest him/her? A lot of say, some say, little say, or no say at all?
Political Efficacy, Level 2	[Linda/Robert][, age XX,] is concerned about cars speeding by [his/her] house, and [he/she] would like to see the speed limit on [his/her] street reduced. [He/she] writes a letter to [his/her] local elected official and receives a form letter in reply. How much say do you think [Linda/Robert] has in getting [his/her] local government to consider issues that interest him/her? A lot of say, some say, little say, or no say at all?
Political Efficacy, Level 3	[Patricia/William][, age XX,] is concerned about cars speeding by [his/her] house, and [his/her] would like to see the speed limit on [his/her] street reduced. [He/she] brings the issue up at a public town meeting. The issue is thoroughly debated by [his/her] local elected officials. How much say do you think [Patricia/William] has in getting [his/her] local government to consider issues that interest him/her? A lot of say, some say, little say, or no say at all?
Political Efficacy Self-Assessment	How much say do you have in getting your local government to consider issues that interest you? Do you have a lot of say, some say, little say, or no say at all?

Note: Half of respondents received female names, and half received male names. Half received vignettes containing the phrase ", age XX, " where XX is the multiple of five nearest to the respondent's own age.

APPENDIX B: Opening instructions for vignettes.

General Health, "explicit age" condition	Please rate the health of the following people using the same categories you would use to rate your own health. [Followed by mention of specific ages in vignettes themselves.]
General Health, "no explicit age" condition	What follows are descriptions of the health of some people your age. Please rate their health using the same categories you would use to rate your own health.
Political Efficacy, "explicit age" condition	Please rate the say in government of the following people using the same categories you would use to rate yourself. [Followed by mention of specific ages in vignettes themselves.]
Political Efficacy, "no explicit age" condition	What follows are descriptions of some people your age concerned about speeding cars in their neighborhood. Please rate their say in government using the same categories you would use to rate yourself
General Health, instructions in Wisconsin Longitudinal Study	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.

APPENDIX C: Predictors of intercategory cutpoint locations, based on vignette ratings (hopit model).

	General Health series (n=1,757)		Political Efficacy series (n=1,749)	
	β	SE	*β*	SE
Cutpoint 1 (Poor-Fair / No say-Little say)				
Female respondent	-.133***	.037	-.043	.043
Age 30-44	-.075	.062	-.027	.069
Age 45-59	-.085	.061	-.143*	.069
Age 60 and above	-.161*	.066	-.144	.074
Less than high school degree	.199**	.065	-.072	.077
Some college	.000	.049	-.094	.055
Bachelor's degree or higher	-.114*	.051	-.201**	.059
HH Income: $25,000-$49,999	.082	.055	.087	.064
HH Income: $50,000-$84,999	.061	.059	-.010	.068
HH Income: $85,000 or higher	.137*	.062	.014	.073
Separated/Divorced/Widowed	.109*	.054	-.029	.064
Never married	.130*	.055	.026	.062
Cohabiting	.129	.072	.097	.082
Black, non-Hispanic	.349***	.068	-.232**	.083
Hispanic	.336***	.070	-.193*	.086
Other, including two+ races	.194**	.073	-.100	.086
Constant	-.718***	.144	-.393**	.129
Cutpoint 2 (Fair-Good / Little say-Some say)				
Female respondent	.000	.037	-.027	.031
Age 30-44	.070	.064	.024	.052
Age 45-59	.065	.064	.016	.052
Age 60 and above	.236***	.066	.071	.055
Less than high school degree	-.024	.062	.022	.057
Some college	-.026	.047	.055	.041
Bachelor's degree or higher	.019	.049	.015	.043
HH Income: $25,000-$49,999	.022	.052	-.016	.049
HH Income: $50,000-$84,999	-.019	.057	.075	.051
HH Income: $85,000 or higher	.002	.060	.071	.055
Separated/Divorced/Widowed	-.121*	.053	.064	.046
Never married	-.084	.055	.001	.046
Cohabiting	-.057	.072	-.172*	.068
Black, non-Hispanic	.182**	.064	-.077	.064
Hispanic	.025	.070	-.097	.067
Other, including two+ races	-.143	.079	.025	.063
Constant	.058	.082	.225**	.073
Cutpoint 3 (Good-Very good / Some say-A lot of say)				
Female respondent	.012	.037	.037	.036
Age 30-44	.039	.063	.027	.062
Age 45-59	.123	.063	.095	.059
Age 60 and above	.110	.067	.045	.064
Less than high school degree	-.031	.065	-.135*	.068
Some college	-.059	.047	.030	.049
Bachelor's degree or higher	-.093	.051	.084	.050
HH Income: $25,000-$49,999	.033	.054	.039	.056
HH Income: $50,000-$84,999	.060	.057	.087	.059
HH Income: $85,000 or higher	-.063	.063	.010	.064
Separated/Divorced/Widowed	.025	.054	-.009	.053
Never married	.151**	.054	-.040	.053

Cohabiting	-.012	.076	.113	.068
Black, non-Hispanic	-.102	.071	-.201**	.067
Hispanic	-.167*	.075	-.146*	.069
Other, including two+ races	-.017	.072	-.115	.073
Constant	-.044	.084	.098	.086
Cutpoint 4 (Very good-Excellent)				
Female respondent	-.038	.039		
Age 30-44	.040	.065		
Age 45-59	.048	.066		
Age 60 and above	-.004	.070		
Less than high school degree	-.054	.077		
Some college	-.061	.052		
Bachelor's degree or higher	.013	.052		
HH Income: $25,000-$49,999	-.048	.061		
HH Income: $50,000-$84,999	-.121	.064		
HH Income: $85,000 or higher	-.160*	.067		
Separated/Divorced/Widowed	-.121*	.059		
Never married	-.128*	.059		
Cohabiting	-.156	.080		
Black, non-Hispanic	-.258**	.087		
Hispanic	-.166*	.083		
Other, including two+ races	-.025	.078		
Constant	.314**	.091		

Note: * $p < .05$, ** $p < .01$, *** $p < .001$, two-tailed. Omitted reference categories: "Male respondent," "Age 18 to 29," "High school degree," "Less than $24,999," "Currently married," and "White, non-Hispanic." Parameterization for cutpoints above the first involves exponentiation, as shown in Equation (1) in main text.

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