Survey Research Methods (2014)
Vol. 8, No. 1, pp. 1-17
ISSN 1864-3361
http://www.surveymethods.org

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

Hanna Grol-Prokopczyk
University at Buffalo

Anchoring vignettes are 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* (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. Specifically, respondents often appear to neglect instructions to treat vignette characters as age peers. Furthermore, when vignette characters' sex is matched to respondents' sex, interpretation of sex differences in rating style is rendered problematic. This study applies two experimental manipulations to a national American sample (n=1,765) to clarify best practices for enhancing response consistency. First, an analysis of two methods of highlighting vignette characters' age suggests that both yield better response consistency than previous, less prominent means. Second, a comparison of ratings of same- and opposite-sex vignette characters suggests that, with avoidable exceptions, the sex of the respondent rather than of the vignette character drives observed sex differences in rating style. Implications for interpretation and design of anchoring vignette studies are discussed. In addition, this study clarifies the importance of two additional measurement assumptions, *cross-respondent vignette equivalence* and *cross-character vignette equivalence*. It also presents empirical findings of significant sex, educational, and racial/ethnic differences in styles of rating health, and racial/ethnic differences in styles of rating political efficacy. These findings underscore the incomparability of unadjusted subjective self-ratings across demographic groups, and thus support the potential utility of the anchoring vignette method.

Keywords: anchoring vignettes; reporting heterogeneity; differential item functioning

1 Introduction

The past decade has seen a burgeoning of interest in anchoring vignettes as a tool for improving intergroup comparability of survey items. However, little research has empirically tested how to design anchoring vignettes so as to maximize their adherence to measurement assumptions, and thereby their validity – despite growing evidence that measurement assumptions are indeed sometimes violated. This article presents two sets of experimental findings to identify best practices for enhancing response consistency (one key measurement assumption) through depiction of vignette characters' age and sex. The article also tests vignette equivalence (the other key assumption) and clarifies interpretation of vignette-based analyses. In addition, the present empirical findings reveal substantial differences across demographic groups in how respondents use subjective response categories, supporting the potential utility of anchoring vignette-based analyses.

Contact information: Hanna Grol-Prokopczyk, University at Buffalo, Department of Sociology, 430 Park Hall, Buffalo, NY 14260, U.S.A. (hgrol@buffalo.edu)

2 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 "very good", for example, may represent the same level of health as another group's "excellent"; or some groups may be more sparing in use of a given category than are others. Phrased more formally, groups may differ in where they locate the intercategory cutpoints (thresholds) along the latent spectrum (see Figure 1 for schematic depiction). This phenomenon, termed "reporting heterogeneity" (e.g., Bago D'Uva, Lindeboom, O'Donnell, & Doorslaer, 2011) or "response-category differential item functioning" (DIF) (King, Murray, Salomon, & Tandon, 2004), can lead to bias in cross-group comparisons – and to highly implausible research findings. Sadana et al.'s (2002) comparison of self-rated health in 46 countries, for example, shows Indonesia, Nepal, and Peru to be among the five healthiest countries, while Spain and France fall among the bottom five. Similarly, unadjusted self-reports show residents of Kerala (the Indian state with the highest life expectancy) to be less healthy than residents of the rest of India, and Americans to be less healthy still (Sen, 2002). In both examples, rank-orderings of regions by subjective self-rated health are inconsistent with – indeed, opposite to –

orderings based on objective measures of health, suggesting that reporting heterogeneity may be quite substantial. Studies have found evidence of reporting heterogeneity in health self-ratings not only across nations (e.g., Iburg, Salomon, Tandon, & Murray, 2002; Jürges, 2007; Jylhä, Guralnik, Ferrucci, Jokela, & Heikkinen, 1998; Murray, Tandon, Salomon, Mathers, & Sadana, 2002; Zimmer, Natividad, Lin, & Chayovan, 2000) but across sexes (Grol-Prokopczyk, Freese, & Hauser, 2011), socioeconomic categories (Dowd & Zajacova, 2007), and races/ethnicities (Menec, Shooshtari, & Lambert, 2007; Shetterly, Baxter, Mason, & Hamman, 1996; Smith, 2003), and reporting heterogeneity appears to be a serious issue in other areas of research as well (e.g., political efficacy (King et al., 2004)).

Since the early 2000s, anchoring vignettes have been promoted as a "most promising" strategy for addressing reporting heterogeneity (Murray et al., 2002). 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 response categories that they use for their own self-rating. Since the same vignette is given to multiple respondents, the objective level of the trait is 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 response category cutpoints (τ's) for each group. By accounting for these different cutpoint locations, self-ratings can be statistically adjusted to be comparable across groups (e.g., King et al., 2004; King & Wand, 2007). Figure 1 presents the logic underlying the anchoring vignette method.

In the past decade, anchoring vignettes have appeared in numerous regional, national, and cross-national surveys, including the Survey of Health, Ageing and Retirement in Europe (SHARE), the Study on Global AGEing and Adult Health (SAGE), and the 70-country World Health Survey (WHS). They have been applied to domains as diverse as political efficacy, job satisfaction, women's autonomy, and binge drinking (Hopkins & King, 2010, 202203; Anchoring Vignettes web site: http://gking.harvard.edu/vign). The anchoring vignette method remains relatively new, however, and advancements continue to be made regarding how to test the method's measurement assumptions (e.g., Bago D'Uva et al., 2011; Datta Gupta, Kristensen, & Pozzoli, 2010; Rice, Silvana, & Smith, 2011; Soest, Delaney, Harmon, Kapteyn, & Smith, 2007) and how to optimize vignette wording and implementation (e.g., Grol-Prokopczyk et al., 2011; Hopkins & King, 2010).

3 Interrogating Measurement Assumptions

Clarifying Assumptions

As described in most writings on anchoring vignettes, two key measurement assumptions are required for the correct functioning of the method: *response consistency* (RC) and *vignette equivalence* (VE) (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 different standards than vignette characters). In the context of 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.

Vignette equivalence is used to mean 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. That is, while respondents may differ in how they understand and use response categories, they cannot differ in their understanding of the vignettes themselves (if both are allowed to vary, the model cannot be identified; see Bago D'Uva et al., 2011). VE would be violated if different respondents interpret the base vignette text in substantially different ways. For example, if an obese vignette character were considered *healthy* by residents of low-income countries (because they see obesity as a sign that the character has avoided starvation or food insufficiency), but were considered *unhealthy* by residents of high-income countries (because, e.g., they associate obesity with increased risk of diabetes or other health problems), then VE has been violated. In Figure 1, VE is indicated by depicting each vignette as a flat, horizontal line – i.e., each vignette represents the same absolute level of health for each of the three groups. (If VE were violated, as in the obesity example, the vignette line would not be flat, as it would cross one group's health spectrum at a different height than another's.)

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". (Most surveys, including SHARE and WHS, use this or very similar wording.)

Anchoring vignette studies rarely acknowledge, however, that matching vignette characters' demographic traits to respondents' demographic traits may put the method's key measurement assumptions into conflict: response consistency is presumably enhanced, since the vignette characters more closely resemble the respondent, but vignette equivalence may be jeopardized, since respondents are no longer all receiving identical vignettes. To deal with this tension, existing vignette studies seem to assume axiomatically (and tacitly) that vignette characters differing in sex, age, or "background" represent identical absolute levels of a trait. Characters' demographic characteristics can thus be manipulated without risk of violating VE. This is, indeed, a crucial assumption, since without it vignette-adjusted self-ratings could not be compared across male and female respondents, or across respondents of different ages and back-

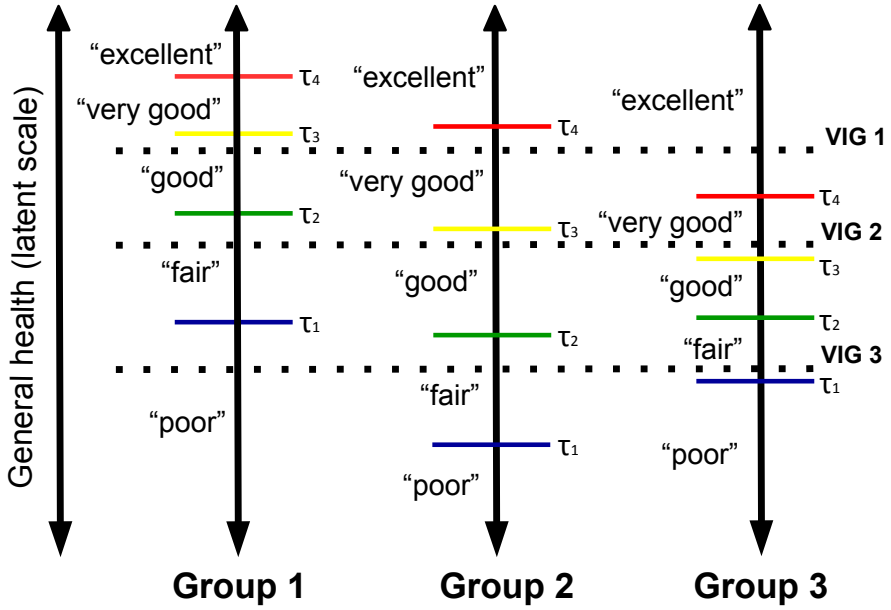

Figure 1. Using Anchoring Vignettes to Estimate Reporting Heterogeneity.

Reporting heterogeneity: Groups may differ in how they use subjective ordered response categories, leading to incomparability of responses across groups. Here, members of Group 1 use systematically higher intercategory cutpoints (τ_1 through τ_4) when rating their health than do members of Group 2, while respondents in Group 3 show a relative compression of cutpoints. A level of health rated "good" in Group 1 might be considered "very good" in Group 2 and "excellent" in Group 3.

Anchoring vignettes: Anchoring vignettes measure and statistically adjust for such reporting heterogeneity. Here, respondents receive three vignettes (dotted 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 cutpoints (τ's) for each group, which are then statistically controlled for to permit unbiased comparisons of self-rated health. See King et al. 2004 and the Anchoring Vignettes web site (http://gking.harvard.edu/vign/) for more information.

grounds.

There appear, then, to be *two* kinds of vignette equivalence assumed in anchoring vignette studies. 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. For precision, we might call this *cross-respondent vignette equivalence*. The second kind of VE, introduced here, postulates that respondents 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 character. We can term this assumption *cross-character vignette equivalence*. That is, if respondent R rates a male character's health differently than an otherwise identical female character's health, it is because R uses different cutpoints for the two sexes, not because R sees them as having different absolute levels of health. Phrased in reference to Figure 1, this means that respondents may use different τ's when rating male versus female vignette characters, but they perceive both to be at the same absolute location along the latent spectrum (i.e., the horizontal line representing the vignette is at the same height in both cases).

Testing Measurement Assumptions

Testing measurement assumptions in anchoring vignette studies is challenging, since neither intercategory cutpoints nor perceived absolute levels of a trait can be observed directly (indeed, in survey research in general, one typically must assume one of these in order to estimate the other). When objective measures of the trait of interest are available, a number of tests of RC are possible, e.g., comparing the cutpoints derived from vignette ratings with cutpoints derived from self-ratings paired with objective measures (as as described by Bago D'Uva et al., 2011; Grol-Prokopczyk et al., 2011; Soest et al., 2007. More generally, vignette validity may be assessed by testing whether vignette-adjusted

self-ratings align more closely with objective measures than unadjusted ratings). However, since anchoring vignettes are most often used when objective measures of the trait are *unavailable* (indeed, vignettes are often presented as a simpler or more affordable alternative to objective measures), such tests are often not possible. Instead, RC is often assessed indirectly, based on the plausibility of vignette-adjusted findings (e.g., King et al., 2004). In the present data, lack of objective measures necessitates such indirect assessment, as described in the discussion of age-related response consistency below.

Cross-respondent VE is most often tested by checking whether respondents consistently rank-order the vignettes in a series (e.g., Murray et al., 2003; Rice et al., 2011; King et al., 2004, 199-200), since if respondents interpret vignettes the same way, they should also rank-order them the same way. Generally, these are "benefit-of-the-doubt" calculations, meaning that ties in ranking are assumed to resolve consistently with the expected order (Murray et al., 2003, 376). Violations of expected ordinal rankings are especially worrisome if they are systematically patterned, as this suggests genuine multidimensionality rather than random measurement error. This study tests cross-respondent VE in this manner.

As mentioned, cross-*character* VE is typically assumed in vignette studies, not explicitly tested. Indeed, it is difficult to think of how to test cross-character VE in a context in which (as demanded by the anchoring vignette method) intercategory cutpoints are allowed to vary. That is, to use the example of male and female vignette characters, one must assume either 1) the same perceived absolute health across male and female characters, while allowing cutpoints for rating the two to differ, or 2) the same cutpoints for rating the two, while allowing their perceived absolute health to differ. For the most part, this article joins other vignette studies in making the former assumption (i.e., the assumption of cross-character VE), and thus presents its experiments primarily as tests of whether *cutpoint locations* differ when rating male versus female characters. However, the same experiments could in fact be interpreted as tests of cross-character VE, *if* one instead assumes equality of cutpoints. In interpreting the results of sex differences in vignette ratings below, both these possibilities are discussed.

Even when cross-character VE is assumed, the matching of vignette characters' and respondents' demographic characteristics requires further consideration, as it may pose methodological and/or interpretational problems. This is discussed in the following sections.

4 Age and Sex of Vignette Characters

How to Improve Age-Related Response Consistency?

Recent findings from anchoring vignette-based studies suggest that respondents often neglect instructions to treat vignette characters as age peers, leading to a violation of response consistency. Grol-Prokopczyk et al. (2011), for example, find that older adults in the Wisconsin Longitudinal Study (WLS) rate general health vignettes more "health-pessimistically" (i.e., using more negative response categories) than do younger adults. Not only is this inconsistent with the predictions of previous literature (e.g., Groot, 2000; Idler, 1993; Doorslaer & Gerdtham, 2003), but it leads to the implausible result that, when self-rated health is adjusted based on vignette ratings, health appears *not* to 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 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. (While the former approach is a minor variation on wording used in other surveys, the latter approach appears to be an innovation of the current study.) Do either or both approaches improve age-related response consistency relative to previous studies? Though, as mentioned, response consistency cannot be measured directly with the present data due to lack of objective measures, the tested vignettes might provisionally be considered successful if they avoid the significant, negative coefficients for higher age dummies found in prior studies.

How to Interpret Sex Differences in Vignette Ratings?

While many surveys consistently sex-match vignettes, some, for ease of administration, field the same set of mixed-sex vignettes to all respondents (e.g., WHS and SAGE[1]), while others randomly assign each vignette character's sex (e.g., Kapteyn, Smith, & Soest, 2007; 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 ratings of vignettes.

As documented by sociolinguists since Lakoff (1973), men and women may use language differently (with these differences varying by socioeconomic and cultural context (Eckert & McConnell-Ginet, 2013; Wardhaugh, 2011). Lakoff notes, for example, that in American English certain adjectives (e.g., "lovely" or "adorable" (1973, 51)) are used

[1] 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).

much more often by women than by men, and more generally, women more frequently use positive or emphatic adjectives. These and other gender differences in language use could lead to different use of response categories on surveys. In addition, there may be domain-specific reasons to expect reporting heterogeneity by sex. For example, as described by Courtenay (2000, 1389), men are often socially constructed as the "stronger" sex, and are expected to minimize their complaints of ill health. Men may thus be more sparing than women in their use of categories such as "poor" to rate their health.

Cultural gender norms might also lead *vignette characters'* sex to affect vignette ratings. For example, referring again to Courteney's (2000) account of "hegemonic masculinity", male vignette characters who acknowledge having health problems, perhaps especially "unmanly" health problems such as pain, may be rated as more unhealthy than female characters with identical complaints.[2] If male and female characters are indeed rated using different standards, then use of opposite-sex vignettes in surveys 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 is 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. As long as one assumes cross-character VE, neither case would invalidate vignette-based adjustments (since, when sexes are matched, response consistency should not be threatened), but clarifying the interpretation of such ambiguous scenarios would be of theoretical importance, and would have practical application 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.

This study clarifies such issues through an experiment randomly assigning respondents to receive same-sex or opposite-sex vignettes. The experimental data are used to compare two idealized scenarios, depicted visually in Figure 2. In Scenario 1 (left side of Figure 2), respondent's sex, but not vignette character's sex, drives observed sex differences in rating style. In this scenario, sex differences in vignette ratings are truly a reflection of women's and men's different styles of evaluation (i.e., there is truly reporting heterogeneity); proxy ratings of opposite-sex family members will be biased due to these different evaluation styles (though such bias could be corrected for with properly designed anchoring vignettes); matching vignette characters' sex to respondents' sex is optional (since it has no bearing on response consistency; respondents use the same intercategory cutpoints no matter what the sex of the vignette character); and results from sex-matching and non-sex-matching designs can be unproblematically compared. In Scenario 2 (right side of Figure 2), only vignette character's sex, not respondent's sex, affects vignette ratings.

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 (since, e.g., women rate men the same way that men rate men); and matching vignette characters' sex to respondents' sex is *crucial* for response consistency, since respondents will use different τ's when rating themselves than when rating opposite-sex characters. Sex-matching vignettes would be essential, not optional, in this scenario. (The possibility that both respondents' and vignette characters' sex affect vignette ratings, perhaps interactively, is also considered and discussed in the results section.)

Other Sociodemographic Differences in Health-Rating Style

In addition to conducting the above two experiments, this study fields general health and political efficacy vignettes to a nationally-representative American sample, and thus provides an opportunity to identify sociodemographic differences (e.g., across age groups, race/ethnicities, or educational groups) in use of response categories when evaluating these domains. (Previous fieldings of the general health vignettes were limited to a racially, geographically, and age-limited sample (Grol-Prokopczyk et al., 2011).) The study can thus directly assess whether some demographic groups are more "health-pessimistic" in subjective health reports than others, as suggested by previous research (e.g., Hispanics compared to whites (Shetterly et al., 1996; Menec et al., 2007; Turner & Avison, 2003)), or whether groups differ in their propensity to use extreme response categories, within or across substantive domains (see, e.g., Smith (2003, 82), on African-Americans and Hispanics using extreme categories more often than whites).

5 Data and Methods

Data

Data collection was sponsored by Time-sharing Experiments for the Social Sciences (TESS) (http://www.tessexperiments.org/), and fielded by Knowledge Networks (http://www.knowledgenetworks.com/). Knowledge Networks recruits respondents to its nationally-representative

[2] The remainder of this section assumes that such differences in ratings of male and female characters reflect different use of intercategory cutpoints, rather than a violation of cross-character VE. Of course, violation of cross-character VE would also be a threat to vignette validity (and a less surmountable one; sex-matching of characters to respondents would then yield incomparable adjusted scores for men and women). Fortunately, as mentioned earlier, the sex-matching experiment presented below can also be interpreted as a test of cross-character VE, and this interpretation is discussed in the results section.

[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 is completely collinear with respondents' sex or with vignette severity.

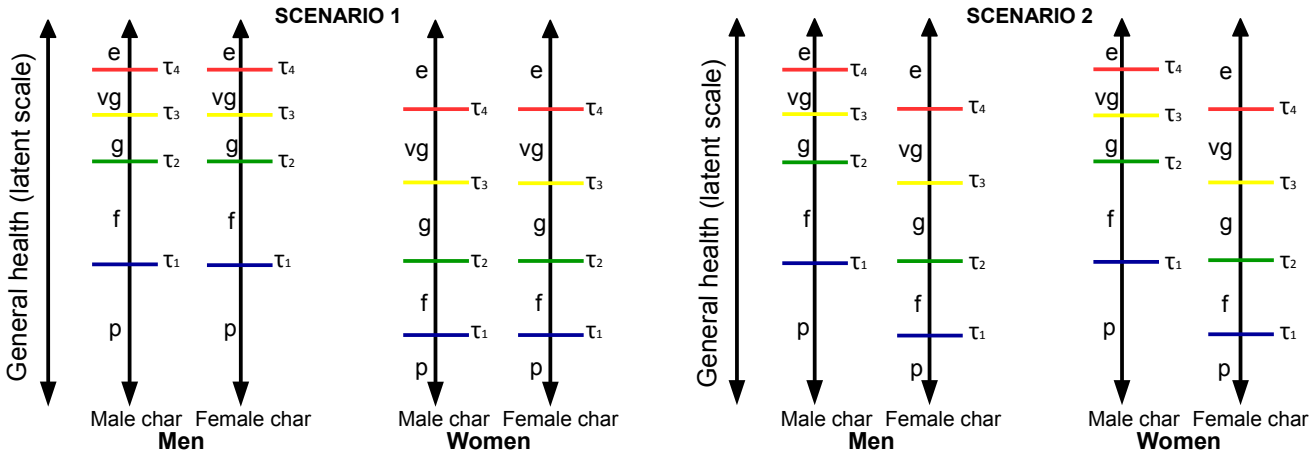

Figure 2. Possible Explanations for Sex Differences in Vignette Ratings.

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.

(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 receive Internet access and 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.)

The current Web-based survey was fielded in June 2010 to 2,750 Knowledge-Panel respondents, of whom 1,771 (64.4%), completed the survey. Of these, six respondents who did not answer any vignette questions were dropped, leaving an analytic sample of 1,765. Non-response for individual vignette questions ranged from .45%-1.25%. Table 1 presents sample characteristics.

Experimental Design

Each respondent received seven English-language vignettes: four general health vignettes and three political efficacy vignettes. These vignette series calibrate key measures in health research and political science, respectively, and have been used in prior studies of anchoring vignettes (Grol-Prokopczyk et al., 2011; Hopkins & King, 2010; King & Wand, 2007). Vignette ratings were reverse-coded to permit intuitive interpretation, i.e., so higher ratings indicate better health or greater political efficacy. The order of the two vignette series, and of items within them, was randomly determined for each respondent. Vignette texts are shown in Appendix A.

The experiment had a 2×2 design, with random assignment to each treatment condition. Half of respondents received vignettes with male names, and half with female names. (To encourage response consistency, names in the vignettes were selected from the top-ten most common names on the 1990 U.S. Census (U.S. Census Bureau, 1995).) Furthermore, half of respondents received vignettes specifying each character's exact age (the "explicit age" condition), where this was the multiple of 5 nearest the respondent's own age; half received vignettes where characters' age was suggested only implicitly in opening instructions (e.g., "What follows are descriptions of the health of some people your age").

Table 2 presents opening instructions for the "explicit age" and "no explicit age" conditions. Ideally, to maximize comparability of findings across surveys, the current experiment would have exactly replicated the wording from the WLS. However, the WLS vignettes appeared at the end of a health module in phone survey, and thus included wording that would not make sense in a stand-alone Internet survey (e.g., "Earlier we asked you to rate your own health overall"; "Now I am going to describe . . . "). Nonetheless, where possible, phrasing from the WLS was replicated.[4]

Analytic Strategy

As described earlier, cross-respondent vignette equivalence was tested by calculating how many respondents cor-

[4] The full instructions from the Wisconsin Longitudinal Study read as follows: "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."

Table 1 Descriptive Statistics for Analytic Sample

	Proportion or Mean	N
Gender		
Female	0.51	898
Male	0.49	867
Age in years	48.76 (SD: 16.69)	1,765
Education		
Less than high school	0.11	194
High school degree	0.28	499
Some college	0.30	521
Bachelor's degree or higher	0.31	551
Household income (in $)		
Less than 24,999	0.20	359
25,000 to 49,999	0.26	458
50,000 to 84,999	0.28	490
85,000 or higher	0.26	458
Marital status		
Currently married	0.52	921
Separated/Divorced/Widowed	0.18	326
Never married	0.21	376
Cohabiting	0.08	142
Race/ethnicity		
White, non-Hispanic	0.77	1,353
Black, non-Hispanic	0.09	151
Hispanic	0.08	139
Other, including two or more races	0.07	121

Table 2 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.

rectly rank-ordered vignettes in each series, and checking whether deviations from expected orderings appeared random or systematic.

Next, ordered probit models were used to identify factors predicting differences in ratings of vignettes. Specifically, vignette ratings were regressed on key demographic variables (sex, age, education, income, marital status, and race/ethnicity) and on flags of experimental conditions. 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 as independent variables.

In addition, "hopit" (hierarchical ordered probit) models were used to identify factors predicting differences in intercategory cutpoint locations (as described in Rabe-Hesketh & Skrondal, 2002; King et al., 2004, 198).[5] 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:

$$v_{ij} = k \text{ if } \tau_i^{k-1} \leq V^*_{ij} < \tau_i^k; -\infty = \tau_i^0 < \tau_i^1 < \ldots \tau_i^K = \infty. \quad (1)$$

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 the present case, comprising measures of sex, age, education, income, marital status, and race/ethnicity) and γ^k represents the corresponding parameters:

$$\begin{aligned} \tau_i^1 &= \gamma^{1\prime} Z_i \\ \tau_i^k &= \tau^{k-1} + e^{\gamma k' Z_i},\ k = 2, \ldots, K. \end{aligned} \quad (2)$$

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

6 Results

Table 3 shows mean ratings of the general health and political efficacy vignettes. The standard deviation for health vignette 4 (describing the least healthy vignette character) is noticeably smaller than for other health vignettes (0.66 versus 0.82-0.88), suggesting a possible floor effect of response categories. Nonetheless, consistent with the assumption of cross-respondent vignette equivalence, ratings of both series decrease/increase monotonically in the expected direction.

In addition, 89.42% of respondents rank-ordered the health vignettes in a manner consistent with the expected ordering. Given that some variation in vignette ordering is expected due to measurement error, these data appear reasonably consistent with the assumption of cross-respondent VE. A lack of any systematic pattern in the observed misorderings (not shown) is also consistent with measurement error, rather than systematic, alternate understandings of vignettes (Rice et al., 2011). In the case of the political efficacy vignettes, however, the percentage of respondents adhering to the expected ordering is lower, despite the smaller number of vignettes: 79.26%. Moreover, the majority of these misorderings (73%) were due to a reversal in ratings of the Level 1 and Level 2 political efficacy vignettes. One may speculate that specific portions of the vignette texts explains this relatively high level of disagreement in ordinal rankings. While the character in the Level 1 vignette makes no effort to contact his or her local elected official, due to hopelessness about receiving help, the character in the Level 2 vignette writes a letter to the local elected official but gets a form letter in reply. Perhaps some respondents find form letters *even more* offensive than complete political non-responsiveness, leading to the frequent inversion of these two vignette scores. While there is no strict cut-off for what proportion of rank order violations ought be considered problematic, it does appear that the present political efficacy vignettes, despite their prior use (Hopkins & King, 2010), are not ideally worded to maximize cross-respondent VE.

Table 4 presents results of ordered probit regressions of vignette ratings (pooled within each series) on experimental conditions and key demographic variables.[6] Regressions of individual (rather than pooled) vignette ratings on the same variables yielded similar results except where noted below. Analyses including interactions between respondent's and character's sex were conducted, but because the interaction term was never statistically significant, it was excluded from presented analyses. Referring to Table 4, results from the two experimental manipulations are now presented, followed by findings regarding demographic predictors of differences in rating style.

Age Experiment Results

As Table 4 shows, 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 vignette series, whether analyzed individually or pooled. Furthermore, the problem of age-related response inconsistency reported in

[5] Some authors refer to this as a "chopit" model (e.g., Rabe-Hesketh & Skrondal, 2002). More often, however, "chopit" – with the "c" standing for "compound" – refers to cases where multiple ratings of each vignette allow for calculation of individual-level random effects. This is not the case in the present models, which are thus referred to simply as "hopit" models.

[6] These models do not meet the parallel regression assumption (i.e., independent variables' effects are not constant across all binary pairings of response categories). Nonetheless, these models constitute a largely accurate summary of findings, providing parameter estimates consistent in direction and statistical significance with those from binary response models (not shown to conserve space; available upon request). Furthermore, the hopit model in Appendix B *does* show 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 latent spectrums of health and political efficacy.

Table 3 Mean Ratings of Anchoring Vignettes

	Vignette 1	Vignette 2	Vignette 3	Vignette 4
General Health	4.17	3.10	1.98	1.48
	(0.85)	(0.88)	(0.82)	(0.66)
Political Efficacy	2.16	2.32	2.95	n/a
	(0.82)	(0.78)	(0.78)	

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

Datta Gupta et al. (2010) and Grol-Prokopczyk et al. (2011) – in which older adults gave more negative ratings of health vignettes – was not replicated, even when the analysis was restricted to white, non-Hispanic respondents aged 60 and above to better resemble the WLS sample (as used by Grol-Prokopczyk et al. (2011); results not shown). This improvement may reflect the present study's more succinct instructions (which use 28 words in the "no explicit age" condition, compared to 65 in the WLS), the fielding by web rather than telephone (which permits respondents to reread instructions), or greater respondent fatigue in the WLS (in which vignettes appeared in 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; both options appear to overcome previously reported problems with age-related response consistency. 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.

Sex Experiment Results

Table 4 indicates that, while vignette character's sex has no significant effect on ratings of political efficacy, male characters elicit lower health ratings than do female ones ($\beta = -.060$; $p = .024$). 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 ($\beta = -.107$; $p = .045$), and in a pooled analysis excluding this vignette, the relationship is no longer statistically significant ($\beta = -.042$; $p = .175$). Character's sex may be relevant in this vignette but not others because it mentions "headaches," which disproportionately affect women (e.g., Fillingim, King, Ribeiro-Dasilva, Rahim-Williams, & Riley, 2009; Kroenke & Spitzer, 1998, 152) and which, as a form of pain, may violate the masculine ideology described by Courtenay (2000). 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 conditions, and thus appear less likely to elicit such gendered associations. This finding is consistent with Angelini, Cavapozzi, and Paccagnella (2011), who find that vignettes mentioning back pain and depression – both substantially more common among women than men (Fillingim et al., 2009; Wetzel, 1994) – "are considered less severe for a woman than for a man". Such effects appear to be independent of the respondent's sex (indicated by the aforementioned lack of interaction between respondent's and character's sex).

There are, however, main effects of respondents' sex: women give systematically higher ratings of health than men ($\beta = .143$; $p < .001$), with a similar though weaker effect found for political efficacy. (This significant sex difference was found for all individual vignettes except health 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.) Perhaps such "positivity bias" reflects the gendered nature of American English described above, in which women are more inclined to use emphatic or positive adjectives (Lakoff, 1973).

Regardless, these findings show that in the tested domains, sex differences in vignette ratings are driven primarily by respondents', not vignette characters', sex. Thus, referring to Figure 2 above, Scenario 1 is more strongly supported than Scenario 2. Previous reports that women are more "health-optimistic" than men thus appear correct, and not mere artifacts of vignette sex-matching (though for the headache vignette, sex differences may be exaggerated by sex matching). These findings suggest that, as long as conditions with gendered associations are avoided in vignettes, matching character's to respondent's sex is not essential for response consistency, and studies that differ in their sex-matching practices can be fairly compared.

It was noted earlier that this same experiment could serve as a test of cross-character vignette equivalence, if one assumes that the same response category cutpoints are used when rating male and female vignette characters. Under this assumption, cross-character VE would be demonstrated by finding no significant differences in ratings of male versus female vignette characters. The present results are, thus, reassuring: for the most part, male and female vignette characters did *not* elicit different ratings. The sole exception was, as mentioned, the headache vignette. It seems likely that replacing mention of headaches with an alternate (non-gendered) health conditions could correct this problem. Violation of cross-character VE thus does not seem to be a serious problem in the present analysis.

Table 4 Ordered Probit Regression of Vignette Ratings on Demographic Variables

	General Health series	Political Efficacy series
Male vignette character	−0.060*	0.049
	(0.027)	(0.030)
Explicit mention of character's age	0.030	0.024
	(0.027)	(0.030)
Female respondent	0.143***	0.062*
	(0.027)	(0.030)
Age		
30-44	0.004	−0.003
	(0.045)	(0.051)
45-59	−0.021	0.104*
	(0.044)	(0.050)
60 and above	−0.070	0.061
	(0.048)	(0.054)
Education		
Less than high school degree	−0.164**	0.099
	(0.049)	(0.055)
Some college	0.062	0.033
	(0.035)	(0.040)
Bachelor's degree or higher	0.143***	0.157***
	(0.037)	(0.042)
Household income (in $)		
25,000 to 49,999	−0.109**	−0.090*
	(0.040)	(0.046)
50,000 to 84,999	−0.052	−0.094
	(0.042)	(0.048)
85,000 or higher	−0.085	−0.085
	(0.045)	(0.051)
Marital status		
Separated/Divorced/Widowed	0.002	−0.032
	(0.039)	(0.044)
Never married	−0.116**	−0.014
	(0.040)	(0.045)
Cohabiting	−0.055	0.024
	(0.053)	(0.060)
Race/ethnicity		
Black, non-Hispanic	−0.423***	0.397***
	(0.050)	(0.056)
Hispanic	−0.279***	0.349***
	(0.052)	(0.059)
Other, including two or more races	−0.080	0.120*
	(0.053)	(0.061)
N	1,757	1,749
Pseudo R-squared	0.290	0.074

Notes: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, 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.

Group Differences in Rating Style

Several respondent characteristics besides sex predicted substantively large differences in vignette ratings. 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 ($\beta = 0.143$, $p < 0.001$ for health; $\beta = 0.157$, $p < 0.001$ 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 = -0.116$, $p = 0.005$). Respondent's age did not

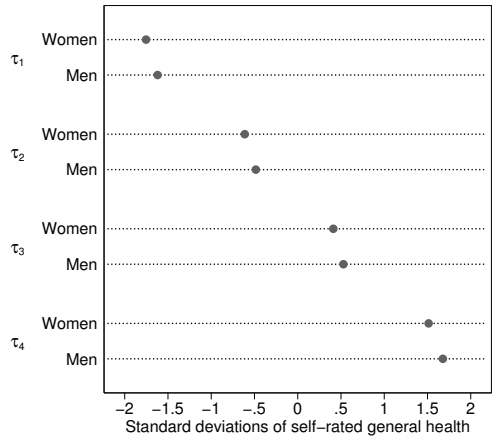

Figure 3. Estimated Cutpoints for General Health by Sex. Intercategory cutpoints ($\tau_1 - \tau_4$) were estimated by applying hopit model coefficients (Appendix B) to the analytic sample. (These cutpoints represent dividing lines between poor/fair, fair/good, good/very good, and very good/excellent health, respectively.)

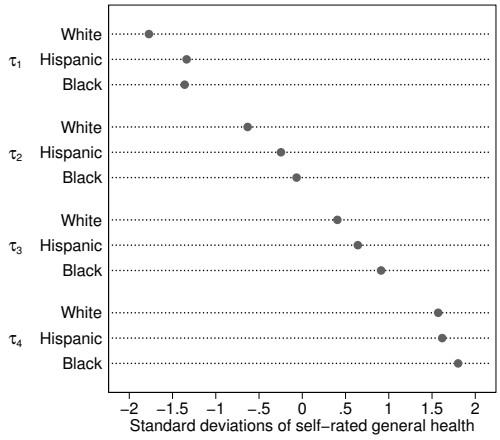

Figure 5. Estimated Cutpoints for General Health by Race/Ethnicity. Intercategory cutpoints ($\tau_1 - \tau_4$) were estimated by applying hopit model coefficients (Appendix B) to the analytic sample. (These cutpoints represent dividing lines between poor/fair, fair/good, good/very good, and very good/excellent health, respectively.)

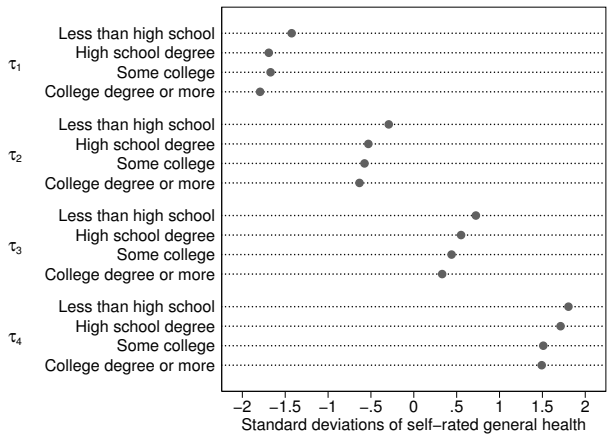

Figure 4. Estimated Cutpoints, General Health, by Education. Intercategory cutpoints ($\tau_1 - \tau_4$) were estimated by applying hopit model coefficients (Appendix B) to the analytic sample. (These cutpoints represent dividing lines between poor/fair, fair/good, good/very good, and very good/excellent health, respectively.)

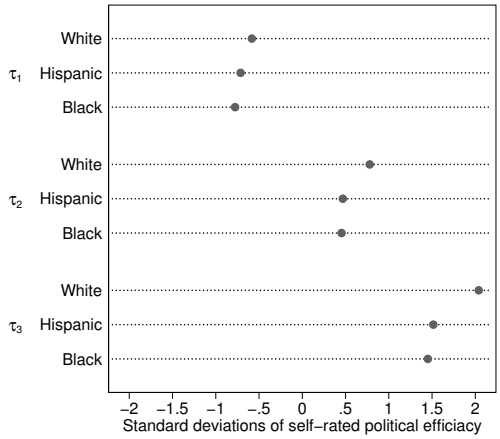

Figure 6. Estimated Cutpoints for Political Efficacy by Race/Ethnicity Intercategory cutpoints ($\tau_1 - \tau_4$) were estimated by applying hopit model coefficients (Appendix B) to the analytic sample. (These cutpoints represent dividing lines between no say/little say, little say/some say, and some say/a lot of say in government, respectively.)

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 ($\beta = .104$, $p = .037$).

The largest coefficients in each model came from race/ethnicity indicator variables. 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.

Appendix B presents a hopit model that uses differences in vignette ratings to estimate intercategory cutpoint locations by group. Because cutpoints beyond the first defy straightforward interpretation (since they are based additively on previous cutpoints and involve exponentiation of coefficients – see Equation 1 above), estimated cutpoint locations are here presented visually. Figures 3-6 show intercategory cutpoints by sex, education, and race/ethnicity for health ratings, and by race-ethnicity for political efficacy. (Differences in political efficacy cutpoints by sex and education are trivial, and thus not pictured.)

Numerical axis units in these graphs are standard deviations (Std. Dev.) of the relevant self-rating (health or political efficacy). Thus, the figures show that women's intercategory cutpoints for rating health are approximately 0.15 Std. Dev. units lower than men's; that college-degree holders' cutpoints are roughly 0.35 Std. Dev. units lower than high school non-completers'; and that differences between white and black cutpoints average 0.4 Std. Dev. units (and sometimes reach 0.6 units). (The 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 differences presented here are extremely large, they do represent non-trivial sources of measurement bias, which could lead to incorrect or misleading research findings.

Anchoring vignette studies can document and adjust for reporting heterogeneity. However, they do not in themselves explain *why* groups differ in their rating styles. Nonetheless, the present findings invite some speculation on this topic. Discussions of reporting heterogeneity often describe it as a phenomenon resulting from different local norms. For example, residents of the Indian state of Kerala typically witness less mortality and morbidity on a daily basis than residents of the (medically and educationally much less developed) state of Bihar (Sen, 2002); as a result, residents of Kerala appear to have higher standards for "good health". Because of such higher cutpoints, a given absolute level of health is likely to be rated more negatively by residents of Kerala than residents of Bihar. Generalizing from this example, one might expect that higher socioeconomic status would predict lower ratings of a given level of health.

In this light, two of the present findings appear puzzling: that higher education predicts *higher* ratings of health vignettes, and that racial/ethnic minority status predicts *lower* ratings. However, group rating styles may reflect more than just differences in group averages or distributions in the phenomena of interest. For example, in the United States, less educated and/or non-white respondents may be less likely to have health insurance; fear of inability to obtain or afford adequate medical treatment may thus lead them to rate more negatively a given level of health impairment. Similarly, the same demographic groups may be more likely to have jobs involving physical labor; poor physical health might then be rated more negatively because it is more likely to jeopardize their ability to work. Taking such factors into consideration, the present findings may not be counterintuitive after all. Of course, this discussion remains speculative, since the present data do not permit testing of the above hypotheses.

The present findings do, however, suggest that a complex host of factors can affect groups' styles of rating health (and other phenomena). One reason anchoring vignettes are useful is precisely because these factors – and the complex ways they may interact – are often not obvious in advance.

7 Discussion

The experimental findings and theoretical clarifications presented in this article yield a number of concrete recommendations for how to design anchoring vignettes so as to minimize violations of measurement assumptions, and thus to maximize vignette validity.

The first experimental manipulation suggests that use of clear opening instructions that highlight vignette characters' ages, or explicit mention of vignette characters' ages in each vignette, both appear to improve age-related response consistency relative to prior studies. (Proof of this is admittedly not definitive, given mode and wording differences between current and prior fieldings of the vignettes. Nonetheless, the current vignettes show none of the clear violations of age-related RC reported in earlier studies; e.g., Datta Gupta et al., 2010; Grol-Prokopczyk et al., 2011.) Given the demonstrated possibility of challenges in this area, the author provisionally recommends explicitly mentioning the character's age in each vignette, in case contextual factors, such as respondent fatigue, lead to poor attention to opening instructions.

Results of the second experimental manipulation suggest that survey designers should strive to avoid mentioning conditions with highly gendered connotations or distributions in anchoring vignettes. Otherwise, differences in ratings of vignettes may result from differences in the sex of the depicted vignette character, rather than true differences in respondents' use of response categories (i.e., true reporting heterogeneity). Phrased in different terms, if a vignette featuring a male character is seen as representing a different

absolute level of a trait than an otherwise-identical vignette featuring a female character, then *cross-character vignette equivalence* has been violated, and anchoring vignette-based adjustments are seriously undermined. In the present study, a health vignette that mentioned "headaches" – a condition substantially more common among women than men – created such a problematic scenario: male characters with "mild headaches" were rated as less healthy than female characters with "mild headaches". Since it may be difficult to predict contexts in which characters' sex affects interpretation of or ratings of vignettes, researchers might consider conducting experiments similar to the present one as part of their pilot work.

When gendered conditions are avoided in vignette texts, however (as in the other six vignettes tested here), vignette ratings appear to genuinely reflect differences in how men and women use response categories. In this case, matching vignette characters' sex to respondents' sex is optional (see discussion of Scenario 1 above), and findings from studies differing in sex-assignment practices can fairly be compared.

This article may be among the first to highlight the tension between the key measurement assumptions of the anchoring vignette method, as demonstrated by the example of sex-matching: efforts to increase response consistency by making vignette characters resemble respondents may undermine cross-character VE – and hence the vignette method as a whole – by unintentionally making vignettes represent fundamentally different levels of the trait of interest. That is, there may be a serious cost to matching vignette characters' and respondent's personal characteristics. This article has focused on matching by sex and age, since omitting information about a vignette character's sex is often not an option (if only for linguistic reasons – i.e., the need to choose male or female pronouns), and since at least in studies of health, age is also often treated as an essential characteristic. However, most personal traits (whether demographic, psychological, personality-related, etc.) *can* be omitted from vignette texts. While the present study is not equipped to recommend omission of all such traits on empirical grounds, it does raise the possibility that efforts to increase RC by describing vignette characters' personal characteristics may have deleterious effects on cross-respondent VE, and hence vignette validity overall. Pending evidence to the contrary, researchers may wish to avoid elaborate attempts to make vignette characters resemble respondents (even if computer-based surveys make such matching increasingly feasible).

Results from the tests of *cross-respondent VE* show that many details of vignette wording – not only those describing personal features of vignette characters – affect adherence to measurement assumptions. While a large majority of respondents correctly rank ordered the health vignettes, a smaller portion correctly ordered the political efficacy vignettes, with a systematic pattern of misorderings in the latter case suggesting a violation of cross-respondent VE. In particular, the misorderings seemed likely to result from the ambiguous interpretation of "form letters": some respondents may see such letters as better than no communication at all, while others find them *more* aggravating than utter silence. To reiterate, careful, low-level examination of the details of vignette wording is necessary to avoid such multidimensionality and maximize vignette validity.

In addition to testing measurement assumptions, this study confirms and extends previous empirical findings of non-trivial reporting heterogeneity across key demographic groups. In particular, especially in the context of general health, more educated respondents give higher ratings than less educated ones, and women give slightly higher (more "optimistic") ratings than men. Such findings suggest that studies based on unadjusted self-ratings will be biased. In addition, evidence of genuine sex differences in rating styles suggests that proxy ratings given by opposite-sex respondents are likely biased due to men and women's different evaluation styles, and thus should be interpreted with caution – or adjusted statistically, potentially using anchoring vignettes.

In addition, in both tested domains, differences in rating style across racial/ethnic groups are strikingly large, with non-whites more "pessimistic" in the context of health (a finding consistent with previous studies, e.g., Shetterly et al. (1996)), and more "optimistic" in the context of political efficacy. Unfortunately, while anchoring vignettes can document reporting heterogeneity, they do not explain it; the present data thus permitted only speculation about *why* groups differ in their rating styles. (Other research techniques, e.g., cognitive interviewing, might be used to address this question.)

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

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 such reporting heterogeneity. At the same time, the study shows that creating well-functioning anchoring vignettes is no trivial enterprise: great attention to detail is required to design vignettes that fulfill their potential.

Acknowledgments

This research uses data collected by Time-sharing Experiments for the Social Sciences (NSF Grant 0818839, Jeremy Freese and Penny Visser, Principal Investigators), and was 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.

Data collection was approved by the UW–Madison Social and Behavioral Sciences Institutional Review Board (protocol SE-2009-0278).

The author thanks Jeremy Freese and Robert M. Hauser for helpful comments about this project.

References

Angelini, V., Cavapozzi, D., & Paccagnella, O. (2011). Dynamics of work disability reporting in Europe. *Journal of the Royal Statistical Society: Series A (Statistics and Society)*, *174*(3), 621-638.

Bago D'Uva, T., Lindeboom, M., O'Donnell, O., & Doorslaer, E. van. (2011). Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. *The Journal of Human Resources*, *46*(4), 875-906.

Courtenay, W. (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, N., Kristensen, N., & Pozzoli, D. (2010). External validation of the use of vignettes in cross-country health studies. *Economic Modelling*, *27*, 854-865.

Doorslaer, E. van, & Gerdtham, U.-G. (2003). Does inequality in self-assessed health predict inequality in survival by income? evidence from swedish data. *Social Science and Medicine*, *57*, 1621–1629.

Dowd, J., & Zajacova, A. (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.

Eckert, P., & McConnell-Ginet, S. (2013). *Language and gender, second edition*. Cambridge: Cambridge University Press.

Fillingim, R., King, C., Ribeiro-Dasilva, M., Rahim-Williams, B., & Riley, J. (2009). Sex, gender, and pain: A review of recent clinical and experimental findings. *The Journal of Pain*, *10(5)*, 447–485.

Grol-Prokopczyk, H., Freese, J., & Hauser, R. (2011). Using anchoring vignettes to assess group differences in self-rated health. *Journal of Health & Social Behavior*, *52(2)*, 246-261.

Groot, W. (2000). Adaptation and scale of reference bias in self-assessments of quality of life. *Journal of Health Economics*, *19*, 403-420.

Hopkins, D., & King, G. (2010). Improving anchoring vignettes: designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, *74(2)*, 201–222.

Iburg, K., Salomon, J., Tandon, A., & Murray, C. (2002). Cross-population comparability of physician-assessed and self-reported measures of health. In *Summary measures of population health: Concepts, ethics, measurement and applications* (p. 433-448). Geneva: World Health Organization.

Idler, E. (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)*, 289-300.

Jylhä, M., Guralnik, J., Ferrucci, L., Jokela, J., & Heikkinen, E. (1998). Is self-rated health comparable across cultures and genders? *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *53(3)*, 144-152.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics*, *16(2)*, 163-178.

Kapteyn, A., Smith, J., & Soest, A. van. (2007). Vignettes and self-reports of work disability in the united states and the netherlands. *The American Economic Review*, *97(1)*, 461-473.

King, G., Murray, C., Salomon, J., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of survey research. *American Political Science Review*, *98(1)*, 191-207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: evaluating and selecting anchoring vignettes. *Political Analysis*, *15(1)*, 46-66.

Kroenke, K., & Spitzer, R. (1998). Gender differences in the reporting of physical and somatoform symptoms. *Psychosomatic Medicine*, *60*, 150-155.

Lakoff, R. (1973). Language and woman's place. *Language in Society*, *2(1)*, 45-80.

Menec, V., Shooshtari, S., & Lambert, P. (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, C., Tandon, A., Salomon, J., Mathers, C., & Sadana, R. (2002). New approaches to enhance cross-population comparability of survey results. In *Summary measures of population health: Concepts, ethics, measurement and applications* (p. 421-431). Geneva: World Health Organization.

Murray, C., Özaltin, E., Tandon, A., Salomon, J., Sadana, R., & Chatterji, S. (2003). Empirical evaluation of the anchoring vignette approach in health surveys. In *Health systems performance assessment: Debates, methods and empiricism.* (p. 369-399). Geneva: World Health Organization.

Rabe-Hesketh, S., & Skrondal, A. (2002). *Estimating chopit models in gllamm: Political efficacy example.* Retrieved 14/03/2014, from http://www.gllamm.org/chopit.pdf

Rice, N., Silvana, R., & Smith, P. (2011). Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *The European Journal of Health Economics*, *12(2)*, 141-162.

Sadana, R., Mathers, C., Lopez, A., Murray, C., & Iburg, K. (2002). Comparative analyses of more than 50 household surveys on health status. In *Summary measures of population health: Concepts, ethics, measurement and applications* (p. 369-386). Geneva: World Health Organization.

Sen, A. (2002). Health: perception versus observation. *BMJ*, *324*, 860-861.

Shetterly, S., Baxter, J., Mason, L., & Hamman, R. (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, T. (2003). Developing comparable questions in cross-national surveys. In *Cross-cultural survey methods* (p. 69-91). Hoboken: John Wiley & Sons.

Soest, A. van, Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. (2007). *Validating the use of vignettes for subjective threshold scales.* IZA Discussion Paper 2860. Retrieved 14/05/2014, from ftp://repec.iza.org/RePEc/Discussionpaper/dp2860.pdf

Turner, R., & Avison, W. (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. (1995). *Genealogy data: Frequently occurring surnames from Census 1990 – names files.* Retrieved 21/02/2014, from http://www.census.gov/genealogy/www/data/1990surnames/names_files.html

Verdes, E. (2011). *Personal communication.* (E-mail, 10 March)

Wardhaugh, R. (2011). *An introduction to sociolinguistics, sixth edition*. Hoboken: John Wiley & Sons.

Wetzel, J. (1994). Depression: Women-at-risk. In *Women's health and social work: Feminist perspectives* (p. 85-108). Binghamton: The Haworth Press.

Zimmer, Z., Natividad, J., Lin, H., & Chayovan, N. (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 3	[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 2	[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. 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	−0.133***	0.037	−0.043	0.043
Age 30-44	−0.075	0.062	−0.027	0.069
Age 45-59	−0.085	0.061	−0.143*	0.069
Age 60 and above	−0.161*	0.066	−0.144	0.074
Less than high school degree	0.199**	0.065	−0.072	0.077
Some college	0.000	0.049	−0.094	0.055
Bachelor's degree or higher	−0.114*	0.051	−0.201**	0.059
HH Income: $25,000-$49,999	0.082	0.055	0.087	0.064
HH Income: $50,000-$84,999	0.061	0.059	−0.010	0.068
HH Income: $85,000 or higher	0.137*	0.062	0.014	0.073
Separated/Divorced/Widowed	0.109*	0.054	−0.029	0.064
Never married	0.130*	0.055	0.026	0.062
Cohabiting	0.129	0.072	0.097	0.082
Black, non-Hispanic	0.349***	0.068	−0.232**	0.083
Hispanic	0.336***	0.070	−0.193*	0.086
Other, including two+ races	0.194**	0.073	−0.100	0.086
Constant	−0.718***	0.144	−0.393**	0.129
Cutpoint 2 (Fair-Good / Little say-Some say)				
Female respondent	0.000	0.037	−0.027	0.031
Age 30-44	0.070	0.064	0.024	0.052
Age 45-59	0.065	.064	0.016	0.052
Age 60 and above	0.236***	0.066	0.071	0.055
Less than high school degree	−0.024	0.062	0.022	0.057
Some college	−0.026	0.047	0.055	0.041
Bachelor's degree or higher	0.019	0.049	0.015	0.043
HH Income: $25,000-$49,999	0.022	0.052	−0.016	0.049
HH Income: $50,000-$84,999	−0.019	0.057	0.075	0.051
HH Income: $85,000 or higher	0.002	0.060	0.071	0.055
Separated/Divorced/Widowed	−0.121*	0.053	0.064	0.046
Never married	−0.084	0.055	0.001	0.046
Cohabiting	−0.057	0.072	−0.172*	0.068
Black, non-Hispanic	0.182**	0.064	−0.077	0.064
Hispanic	0.025	0.070	−0.097	0.067
Other, including two+ races	−0.143	0.079	0.025	0.063
Constant	0.058	0.082	0.225**	0.073

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, 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 2 in main text

Appendix B. 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 3 (Good-Very good / Some say-A lot of say)				
Female respondent	0.012	0.037	0.037	0.036
Age 30-44	0.039	.063	0.027	0.062
Age 45-59	0.123	0.063	0.095	0.059
Age 60 and above	0.110	0.067	0.045	0.064
Less than high school degree	−0.031	0.065	−0.135*	0.068
Some college	−0.059	0.047	0.030	0.049
Bachelor's degree or higher	−0.093	0.051	0.084	0.050
HH Income: $25,000-$49,999	0.033	0.054	0.039	0.056
HH Income: $50,000-$84,999	0.060	0.057	0.087	0.059
HH Income: $85,000 or higher	−0.063	0.063	0.010	0.064
Separated/Divorced/Widowed	0.025	0.054	−0.009	0.053
Never married	0.151**	0.054	−0.040	0.053
Cohabiting	−0.012	0.076	0.113	0.068
Black, non-Hispanic	−0.102	0.071	−0.201**	0.067
Hispanic	−0.167*	0.075	−0.146*	0.069
Other, including two+ races	−0.017	0.072	−0.115	0.073
Constant	−0.044	0.084	0.098	0.086
Cutpoint 4 (Very good-Excellent)				
Female respondent	−0.038	0.039		
Age 30-44	0.040	0.065		
Age 45-59	0.048	0.066		
Age 60 and above	−0.004	0.070		
Less than high school degree	−0.054	0.077		
Some college	−0.061	0.052		
Bachelor's degree or higher	0.013	0.052		
HH Income: $25,000-$49,999	−0.048	0.061		
HH Income: $50,000-$84,999	−0.121	0.064		
HH Income: $85,000 or higher	−0.160*	0.067		
Separated/Divorced/Widowed	−0.121*	0.059		
Never married	−0.128*	0.059		
Cohabiting	−0.156	0.080		
Black, non-Hispanic	−0.258**	0.087		
Hispanic	−0.166*	0.083		
Other, including two+ races	−0.025	0.078		
Constant	0.314**	0.091		

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, 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 2 in main text