Does gender influence subjective health reports? Evidence using anchoring vignettes.

Megan Todd, Office of Population Research and Woodrow Wilson School for Public and International Affairs, Princeton University
Jennifer Dowd, Hunter College, City University of New York and City University of New York Institute for Demographic Research

March 2011
DRAFT: do not cite

Abstract

Gender differences in health are well established in many contexts, with women often paradoxically reporting worse health but with lower rates of mortality. Self-rated health (SRH) measures are commonly used to quantify and explain differences in health status across groups, but population groups may differ in their understanding and interpretation of SRH survey questions. Anchoring vignettes have been introduced as one potential solution to this problem, whereby respondents' reports of their own health are "anchored" by their ratings of the health of a fictitious individual. This paper uses anchoring vignettes across six domains of health to consider how gender influences health ratings, taking advantage of randomization of the gender of the vignette characters in the Health and Retirement Study. We find evidence that health ratings are influenced by both respondent and vignette character gender. After adjusting for these differences, the gap between male and female SRH widened, suggesting that the poor health of women relative to men may be underestimated.

Background

Valid, reliable, and comparable indicators of health status are essential for measuring population health, determining the impact of health interventions, and understanding health disparities (Mathers, Murray et al. 2003). A growing demand for accountability and measurement in health programs highlights the need for consistent measures of health that are comparable through time and across populations. General self-rated health (SRH) on a 5-point scale (poor, fair, good, very good, or excellent), is a frequently used summary measure of health status. SRH is easy and inexpensive to determine in a survey, and is a consistent predictor of future morbidity and mortality in many contexts, even after controlling for various biological measures of health (Idler and Benyamini 1997; Deeg and Kriegsman 2003; DeSalvo, Bloser et al. 2005). Similar self-reported health measures on 5-point scales are also used to measure health in specific domains such as mobility or pain.

Self-reported health measures may be problematic in measuring health disparities if respondents from different population groups systematically understand and interpret survey questions differently (Mathers, Murray et al. 2003). Such systematic "reporting heterogeneity" (Lindeboom and van Doorslaer 2004) could lead to over- or underestimates of between-group health disparities if, for a given objective level of health, one group systematically reports better health than another group. Evidence of such reporting heterogeneity has been found along the dimensions of survey respondents' language (Bzostek, Goldman et al. 2007), nationality (Jürges 2007), socioeconomic status (SES) (Humphries and Van Doorslaer 2000; Shmueli 2003; Etile and Milcent 2006), and race/ethnicity (Boardman 2004; Spencer, Schulz et al. 2009).

Variation by gender is particularly paradoxical: while women on average report worse health in SRH surveys (Benyamini, Blumstein et al. 2003; Lindeboom and van Doorslaer 2004), they have greater survival rates at every age (Case and Paxson 2005). Could differences in reporting behavior by gender be partly responsible for these findings? As one example of reporting differences in global health, some studies have shown that men and women place different weights on particular domains of health when considering overall health, with women giving sleep and mental health greater consideration (Brunner 2006).

One approach to the problem of disentangling true health differences from reporting differences is the use of anchoring vignettes in a survey along with SRH questions. In an anchoring vignette, survey respondents are asked to read a short description of the health of a fictitious individual and rate the individual's health, using the same categories used to assess their own health. Health ratings for both respondents and fictitious vignette characters are domain specific, meaning

respondents rate health related to, for example, pain, mobility, and depression separately. Thus far, to our knowledge, vignettes for global SRH have not been attempted due to the challenge of constructing vignettes that would encompass all of the elements comprising an individual's overall health, but anchoring vignettes have recently been introduced in several surveys for specific health domains.

Because each survey respondent receives the same information about the vignette characters' health, variation between different respondent's ratings of the same vignette characters can provide a measure of relative rating "conservatism" that can then be used to adjust his or her own SRH. Specifically, a respondent's rating of the vignette characters is used to estimate the respondent's thresholds between reporting categories (cut-points). These cut-points are then used to adjust the respondent's SRH, thus purging the effect of reporting heterogeneity from the health measure.

The validity of the anchoring vignette approach depends on two keys assumptions. First, it must be assumed that all respondents interpret and understand the health of the vignette character in the same way. This assumption has been called "vignette equivalence" (Bago d'Uva, van Doorslaer et al. 2008). Second, it must be assumed that respondents evaluate their own health in the same way that they evaluate the health of the vignette characters. This assumption has been called "response consistency" (Bago d'Uva, van Doorslaer et al. 2008).

This paper takes advantage of a uniquely designed set of anchoring vignettes in which half of the sample was given vignette characters of one gender and half of the sample the other gender. We test three (non-mutually exclusive) hypotheses with regards to gender differences in reporting behaviors:

(1) "Respondent gender" hypothesis: Given identical vignettes, female respondents rate the health of the vignette character (of either gender) differently than male respondents do. That is, the health rating of a vignette character depends on the gender of the respondent, holding constant the gender of the vignette character. This hypothesis is represented by:

$HR(F,M) = HR(M,M) \neq HR(F,F) = HR(M,F)$

where HR(i,j) is the health rating of a vignette character of gender i by respondent of gender j.

If female respondents rate the health of all equivalent vignette characters worse than male respondents, this could mean that women have systematically stricter standards for rating health compared to men. If this type of reporting heterogeneity were found, it could affect the magnitude and/or direction of male/female health differences measured by self-reports.

(2) "Vignette gender" hypothesis: Given vignettes that are identical except for the gender of the vignette character, respondents (of both genders) interpret the health of a female vignette character and a male vignette character differently. In other words, the health rating of a vignette character

depends on the gender of the vignette character, holding constant the gender of the respondent. This hypothesis is represented by:

$HR(F,M) = HR(F,F) \neq HR(M,M) = HR(M,F)$

where HR(i,j) is the health rating of a vignette character of gender i by respondent of gender j.

If respondents systematically rate the health of male vignette characters more favorably, this would be consistent with more lenient health standards for men compared to women, meaning it takes a worse level of objective health for the respondent to downgrade the health of a man. This might be due to systematic beliefs that men are "tougher" or somehow are more resilient than women with the same level of observed health. If this kind of systematic belief leads a woman to rate her own health differently from how she rates the health of a male vignette, then the assumption of response consistency would be violated.

(3) "Gender empathy" hypothesis: Given vignettes that are identical except for the gender of the vignette character, respondents interpret the health of a vignette character of the same gender differently than a vignette character of the opposite gender. That is, the health rating of a vignette character depends on both the gender of the respondent and the gender of the vignette character. In symbols:

$HR(i,j) \neq HR(i,i)$

where HR(i,j) is the health rating of a vignette character of gender i by respondent of gender j, $i,j \in (M,F)$ and $i \neq j$.

If respondents rate the health of same-sex vignette characters differently than opposite-sex vignette characters, this could mean that respondents empathize more with vignette characters of the same gender. This would provide evidence in favor of using same-sex vignettes.

Providing evidence for or against these three hypotheses will help shed light on the proper design of vignette studies as well as whether measurement of sex differences in health based on self-reports is likely to be biased or not. The direction of results will also provide empirical evidence of how health assessments differ by gender.

Data and Methods

Data come from the U.S. Health and Retirement Study (HRS) 2006 Core, and the Disability Vignette Study (DVS), a mail survey conducted in 2007 by HRS. The HRS Core is a nationally representative survey of U.S. adults aged 50 and older and their spouses (regardless of age). DVS respondents were asked to rate the severity of their own health problems in six domains (pain, sleep,

mobility, memory, shortness of breath, and depression). They then read several short vignettes and rated the severity of the vignette characters' health problems in the same six domains. Two versions of the DVS were used, with the gender of the vignette characters changed between versions. Respondents who participated in an HRS 2006 Core self-interview were eligible for participation in the DVS. Those who had died prior to the start of the DVS, those who requested removal from the sample, and those who were participating in other HRS studies in 2007 were removed. Of 5,678 mailed questionnaires, 4,639 were returned (81.7%). DVS data were combined with HRS 2006 Core data including demographic information on sex, age, race/ethnicity, and educational attainment.

Measures

Survey respondents rated the severity of their own health problems and those of the fictitious vignette characters on a five point scale: 1=none, 2=mild, 3=moderate, 4=severe, and 5=extreme. Self-rated health in each domain was obtained from the following questions: "Overall, in the last 30 days, how much…."

- "pain or bodily aches did you have?" (pain)
- "difficulty did you have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?" (sleep)
- "of a problem did you have with moving around?" (mobility)
- "difficulty did you have with concentrating or remembering things?" (memory)
- "of a problem did you have because of shortness of breath?" (shortness of breath)
- "of a problem did you have with feeling sad, low, or depressed?" (depression)

Vignette questions were of the form: "Charles has pain in his knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, he feels uncomfortable when moving around, holding and lifting things. Overall, in the last 30 days, how much of bodily aches or pains did Charles have?" Three vignettes, representing three different levels of severity, were presented for each health domain.

Amongst our explanatory variables, age was measured continuously as of 12/31/2006. Race/ethnicity was categorized as white (reference), non-Hispanic black, Hispanic, and non-Hispanic other. Educational attainment was classified into four categories: less than high school (reference), high school, some college, and college or more.

Empirical strategy

We estimate two ordered probit models of respondents' ratings of vignette characters' health problems for each health domain. In the first model, a respondent's rating of the severity of a vignette

character's health problems depends on the gender of the respondent, the gender of the vignette character, and other demographic control variables (age, race/ethnicity, and educational attainment):

$$\Pr(h_{jv}=i) = \Pr(\kappa_{i-1} < \beta_1(gender_v) + \beta_2(gender_j) + \beta_4(\boldsymbol{D_j}) + \varepsilon_j \leq \kappa_i) \quad \text{(Model 1)}$$

where h_{jv} is person j's rating of the health problem of vignette character v, i is the severity level from 1=none to 5=extreme, $\kappa_1,\ldots,\kappa_4$ are cut-points, and $\boldsymbol{D}_j$ is the vector of person j's demographic control variables (age, race/ethnicity, and educational attainment).

The second model adds an interaction term between the gender of respondent and the gender of the vignette character:

$$\Pr(h_{jv}=i) = \Pr(\kappa_{i-1} < \beta_1(gender_v) + \beta_2(gender_j) + \beta_3(gender_v * gender_j) + \beta_4(\boldsymbol{D_j}) + \varepsilon_j \leq \kappa_i) \quad \text{(Model 2)}$$

where h_{jv} is person j's rating of the health problem of vignette character v, i is the severity level from 1=none to 5=extreme, $\kappa_1,\ldots,\kappa_4$ are cut-points, and $\boldsymbol{D}_j$ is the vector of person j's demographic control variables age, race/ethnicity, and educational attainment.

A finding that $\beta_1 \neq 0$ would mean that female respondents rate male and female vignette characters differently; this would support the vignette gender hypothesis. $\beta_2 \neq 0$ would mean that male and female respondents rate female vignette characters differently, providing evidence for the respondent gender hypothesis. $\beta_3 \neq 0$ would mean male respondents rate male vignette characters differently from female respondent rating female vignettes, and would be consistent with the gender empathy (or dis-empathy, depending on the sign) hypothesis.

Because gender differences in ratings could differ by the respondents' age or race/ethnicity, we also test for interaction effects with these variables. In the following four models, we individually allow interaction with the gender of the vignette character and the gender of the respondent:

$$\Pr(h_{jv}=i) = \Pr(\kappa_{i-1} < \beta_1(gender_v) + \beta_2(gender_j) + \beta_{I1}(gender_j * age_j) + \beta_4(\boldsymbol{D_j}) + \varepsilon_j \leq \kappa_i) \quad \text{(Model I1)}$$

$$\Pr(h_{jv}=i) = \Pr(\kappa_{i-1} < \beta_1(gender_v) + \beta_2(gender_j) + \beta_{I2}(gender_j * race/ethnicity_j) + \beta_4(\boldsymbol{D_j}) + \varepsilon_j \leq \kappa_i) \quad \text{(Model I2)}$$

$$\Pr(h_{jv}=i) = \Pr(\kappa_{i-1} < \beta_1(gender_v) + \beta_2(gender_j) + \beta_{I3}(gender_v * age_j) + \beta_4(\boldsymbol{D_j}) + \varepsilon_j \leq \kappa_i) \quad \text{(Model I3)}$$

$$\Pr(h_{jv}=i) = \Pr(\kappa_{i-1} < \beta_1(gender_v) + \beta_2(gender_j) + \beta_{I4}(gender_v * race/ethnicity_j) + \beta_4(\boldsymbol{D_j}) + \varepsilon_j \leq \kappa_i) \quad \text{(Model I4)}$$

where h_{jv} is person j's rating of the health problem of vignette character v, i is the severity level from 1=none to 5=extreme, $\kappa_1,\ldots,\kappa_4$ are cut-points, and $\boldsymbol{D}_j$ is the vector of person j's demographic control variables age, race/ethnicity, and educational attainment.

If $\beta_{I1} \neq 0$ or $\beta_{I2} \neq 0$, this would suggest that the respondent gender hypothesis is relatively more strongly supported among respondents of a particular age or race/ethnicity. If $\beta_{I3} \neq 0$ or $\beta_{I4} \neq 0$, this would similarly provide stronger or weaker evidence for the vignette gender hypothesis depending on the respondents' characteristics.

Finally, we assess the size of potential biases in self-reported severity of health problems that are due to the reporting differences documented above by comparing a naive ordered probit model of self-rated severity of health problems to two adjusted models: 1) an ordered probit model that controls for the variables found above to predict ratings of vignette characters' health problems, and 2) a HOPIT model that adjusts for reporting heterogeneity by allowing variation in the cut-points between reporting categories. For more detail on this model, see (Bago d'Uva, O'Donnell et al. 2008; Bago d'Uva, van Doorslaer et al. 2008). The HOPIT model jointly estimates the cut-points between reporting categories based on the respondents' ratings of the severity level of vignette characters' health problems and the severity level of the respondent's own health problems based on these cut-points. The thresholds for reporting extreme, severe, moderate, mild or no problem in each health domain were allowed to vary by covariates. The naïve ordered probit model includes only characteristics of the respondent as covariates (respondent gender, age, educational attainment, and race/ethnicity). Both the adjusted probit and HOPIT models include the following right hand side variables: respondent gender, vignette gender, respondent gender * age, vignette gender * race/ethnicity, age, educational attainment, and race/ethnicity.

Main results

Table 1 shows descriptive statistics of the DVS respondents. Table 2 displays respondents' mean self-rated severity of health problems by gender, and the test statistic from an equality of means t test. In the domains of pain, sleep, shortness of breath, and depression, there is a significant difference between the male and female self-rated severity with female respondents reporting greater severity of health problems related to pain, sleep and depression, and male respondents reporting more severe health problems in the domain of shortness of breath.

Results from probit models of the reported severity of vignette characters' health problems are shown in Table 3. The coefficient for male vignette character β_1 is significant at the 5% level in the domain of pain ($p=0.045$ in model 2), sleep ($p<0.001$ in both models), mobility ($p=0.015$ in model 1), and shortness of breath ($p<0.001$ in both models). For vignettes identical except for the gender of the vignette character, the health problems of male vignette characters are rated by female respondents as more severe in the domains of sleep, mobility, and shortness of breath; the health problems of male vignette characters are rated by female respondents as less severe in the domain of pain. These results support the vignette gender hypothesis for female respondents, but the direction of the results differs

by domain. The "men are tough" story is supported only in the domain of pain; female vignette characters are the "tough" ones in the domains of sleep, mobility and shortness of breath.

The coefficient for male compared to female respondent gender β_2 is significant in the domain of sleep (p=0.045 in model 1) and memory (p=0.029 in model 2). In particular, male respondents rate a female vignette character's sleep problems as being less severe than that rating given by female respondents; male respondents rate a given set of memory problems more severely than female respondents do. These findings support the respondent gender hypothesis in two out of six health domains, but not in a consistent direction. Male respondents appear to have more lenient health standards in the domain of sleep, but female respondents are more lenient in rating health problems related to memory.

The coefficient on the interaction of the genders of the vignette character and the respondent β_3 is only marginally significant in the domain of depression (p=0.052), with male respondents rating the depression problems of male vignette characters as less severe compared to male respondents rating the depression of female vignette characters. This suggests that, with the exception of depression, male respondents do not give different ratings for same-sex versus opposite sex vignette characters. Looking at the coefficient for male vignette character in Model 2, which represents female respondents with male vignette characters compared to female respondents with female vignette characters, we see that female respondents do in fact rate same-sex vignette characters differently than opposite-sex vignette characters in three of the six domains (pain, sleep, and shortness of breath). Thus the effects for male vignette characters generally being rated with more severe health problems for the same given health compared to female vignette characters seems to be driven primarily by female rather than male respondents.

Abridged results from the models with age and race/ethnicity interactions are shown in Figures 1 and 2. (Complete results are shown in Appendix Table A1.) The male respondent * age interaction is significant in the domains of pain (p=0.004), memory (p=0.003), shortness of breath (p=0.014), and depression (p=0.025). For male respondents, increasing age corresponds to a lower severity rating in all four of these domains. For female respondents, increasing age acts in the same direction for shortness of breath and memory, but acts in opposite direction for pain and memory (Figure 1). The male vignette character * Hispanic interaction is significant in the domains of sleep (p=0.031) and shortness of breath (p=0.024). Both white and Hispanic respondents rate the sleep problems of male vignette characters as more severe than those of female vignette characters, but Hispanic respondents do so to a greater degree. While white respondents rate the shortness of breath problems of male vignette characters as

more severe than those of female vignette characters, Hispanic respondents rate male vignette characters' shortness of breath problems as less severe than those female vignette characters (Figure 2). No other interaction variables were significant.

Overall, we find support for two of our three hypotheses: 1) While on average respondents have different reporting standards for male versus female vignette characters in four of six health domains, these results are primarily due to reporting differences for female versus male respondents. Rather than "empathy" for female vignette characters, female respondents seem tougher on female vignette characters, giving them less severe ratings for the same fixed level of health compared to male vignette characters. The magnitude and direction of these differences depends on respondent's Hispanic race/ethnicity in the domains of sleep and shortness of breath. 2) Holding fixed the sex of the vignette character, we find differences in reporting standards between male and female respondents in two of our six health domains (sleep and memory). When we allow for the differential male/female effect of respondent age on rating, we find sex differences in four domains.

Adjustment for reporting differences

Figure 3 shows the coefficient for male respondent from three different models of self-reported severity of health problems. In the naïve ordered probit model, male respondents report less severe problems with sleep and depression, but more severe problems with shortness of breath. The additional control variables in the adjusted probit model cause an increase in magnitude of the negative coefficient on male in the domain of sleep; respondent gender becomes insignificant for shortness of breath and depression. The negative coefficient on memory becomes larger and significant. These results show that male respondents generally rate their own health problems as less severe than female respondents do, and this effect is stronger after controlling for additional variables. Finally, the adjustment for reporting differences of the HOPIT model causes the negative coefficient on shortness of breath to become significant, and increases the magnitude of the negative coefficients in the domains of sleep and memory. This adjustment further strengthens the male advantage in self-rated health. These results suggest that failing to adjust for reporting differences may underestimate the gender differences in health status. Rather than explaining the puzzle of women's lower self-rated health despite greater survival, these results appear to deepen the puzzle. These findings are consistent with the argument that women suffer more from specific health conditions, despite their longevity (Case and Paxson 2005), and suggest that the degree to which women suffer from these conditions may have actually been underestimated.

Conclusion

In this study, we tested for gender differences in health reporting behavior by examining the severity rating assigned to fictitious vignette characters by DVS respondents. We asked whether respondents systematically rate the severity of vignette characters' health problems differently depending on vignette character gender (the "vignette gender" hypothesis), respondent gender (the "respondent gender" hypothesis), or an interaction of respondent gender * vignette character gender (the "gender empathy" hypothesis). We find support for the vignette gender hypothesis in four domains of health, which is driven primarily by female respondents. The respondent gender hypothesis is supported in two domains, and the gender empathy hypothesis is supported only for women in three of six domains, but in the direction of gender dis-empathy. We additionally find some evidence of interaction effects by age and race/ethincity. Vignette gender plays a different role among Hispanic respondents in two domains as compared to white respondents, and the direction and magnitude of the effect of respondent gender depends on respondent age.

Many current studies that utilize anchoring vignettes give respondents same-sex vignettes (e.g. Hopkins and King 2010). Our results support the need for this practice mainly due to the reporting differences of female respondents for same versus opposite sex vignettes. The unique HRS vignette study design, with half the sample receiving vignettes of one gender and half the other, allowed interesting insight into gender perceptions and health. In general, female respondents were more likely to rate the health of female vignette characters better compared to male characters for the same fixed level of health, with the exception of pain for which the opposite was true. Male respondents in this sense were the most equality minding, with the same reporting standards for male versus female vignette characters. Whether women have expectations that women should "tough it out" or simply that women can manage the same symptoms better than a man remains for future work to elucidate.

Table 1: Descriptive Statistics of DVS Respondents

	Mean or Proportion	Standard Deviation
Male	39.58%	
Age (as of end of year 2006)	65.17	10.63
Age: 30-39	0.28%	
Age: 40-49	3.61%	
Age: 50-59	33.81%	
Age: 60-69	31.73%	
Age: 70-79	19.82%	
Age: 80-89	9.42%	
Age: 90-99	1.28%	
Age: 100 +	0.04%	
Race/ethnicity: non-Hispanic white	77.63%	
Race/ethnicity: non-Hispanic black	11.28%	
Race/ethnicity: Hispanic	8.39%	
Race/ethnicity: non-Hispanic other	2.70%	
Education: < high school	16.32%	
Education: high school	34.67%	
Education: some college	23.74%	
Education: college +	25.27%	
N	4,626	

Table 2: Mean self-rated severity of health problems by respondent gender

	Female mean	Male mean	T statistic
Pain	2.40	2.33	2.60
Sleep	2.34	2.19	5.17
Mobility	1.77	1.79	-0.63
Memory	1.89	1.88	0.26
Shortness of Breath	1.46	1.51	-2.24
Depression	1.84	1.68	6.05

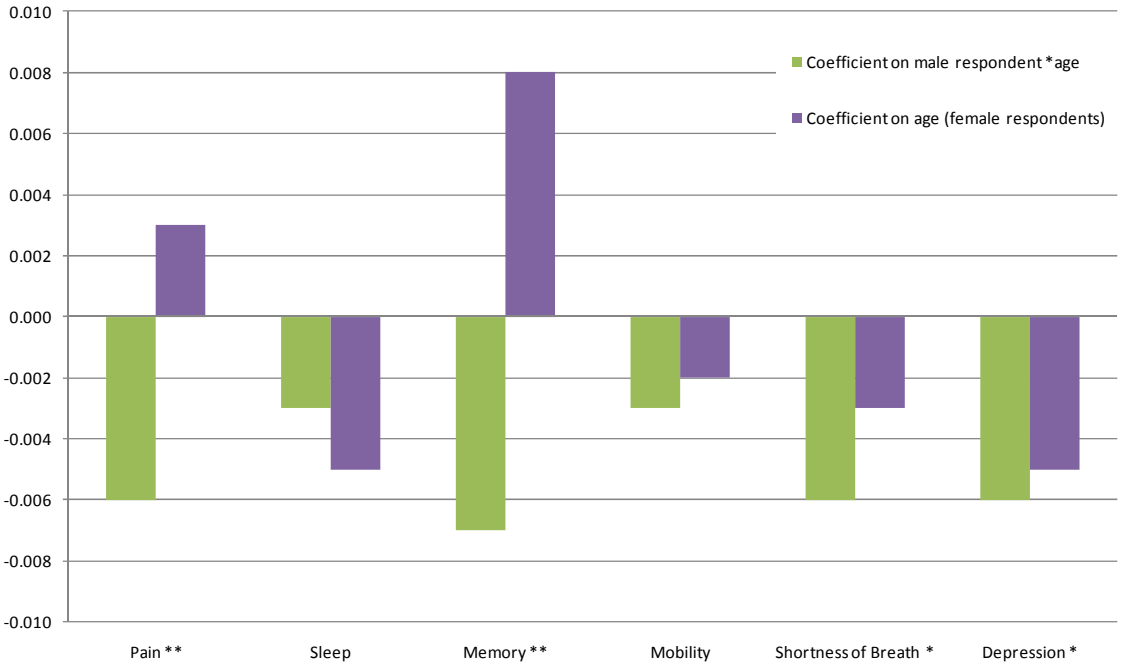
Note: Self-rated severity of health problems is rated on a five point scale where 1=none, 2=mild, 3=moderate, 4=severe, 5=extreme.

Table 3: Ordered Probit Model of Respondents' Rating of the Severity of Vignette Characters' Health Problems

	Pain		Sleep		Memory		Mobility		Shortness of Breath		Depression	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value
Male respondent	0.023	0.004	-0.051	-0.023	0.038	0.072	0.003	-0.006	-0.017	-0.014	-0.009	0.040
	(0.333)	(0.906)	(0.045)	(0.517)	(0.113)	(0.029)	(0.904)	(0.871)	(0.506)	(0.693)	(0.741)	(0.281)
Male vignette character	-0.043	-0.058	0.116	0.139	-0.012	0.015	0.058	0.051	0.174	0.176	-0.019	0.020
	(0.056)	(0.045)	(<0.001)	(<0.001)	(0.590)	(0.607)	(0.015)	(0.094)	(<0.001)	(<0.001)	(0.424)	(0.505)
Male respondent * Male vignette character		0.037		-0.058		-0.069		0.017		-0.007		-0.099
		(0.424)		(0.254)		(0.142)		(0.730)		(0.895)		(0.052)
N	13,818	13,818	13,821	13,821	13,823	13,823	13,810	13,810	13,818	13,818	13,811	13,811

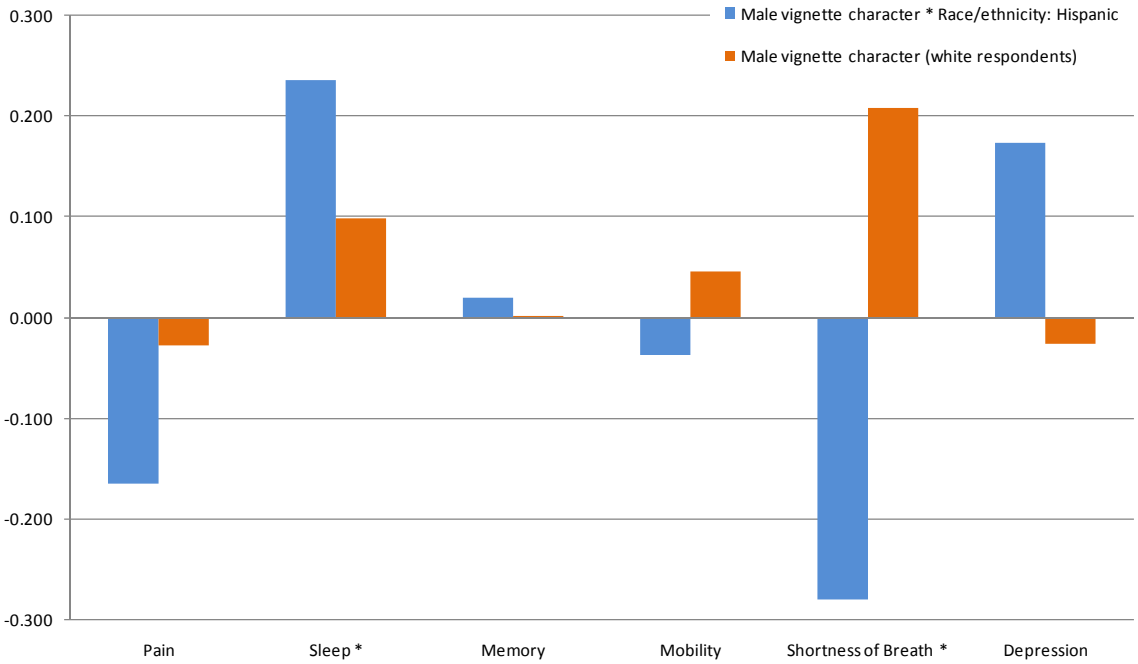
Note: models control for age, educational attainment, and race/ethnicity. Robust standard errors are clustered at the respondent level. Health problems in each of the six domains shown were rated on a 5-point scale where 1=none, 2=mild, 3=moderate, 4=severe, and 5=extreme.

Figure 1: Coefficients on age from ordered probit models of the severity rating of vignette characters health problems, by gender



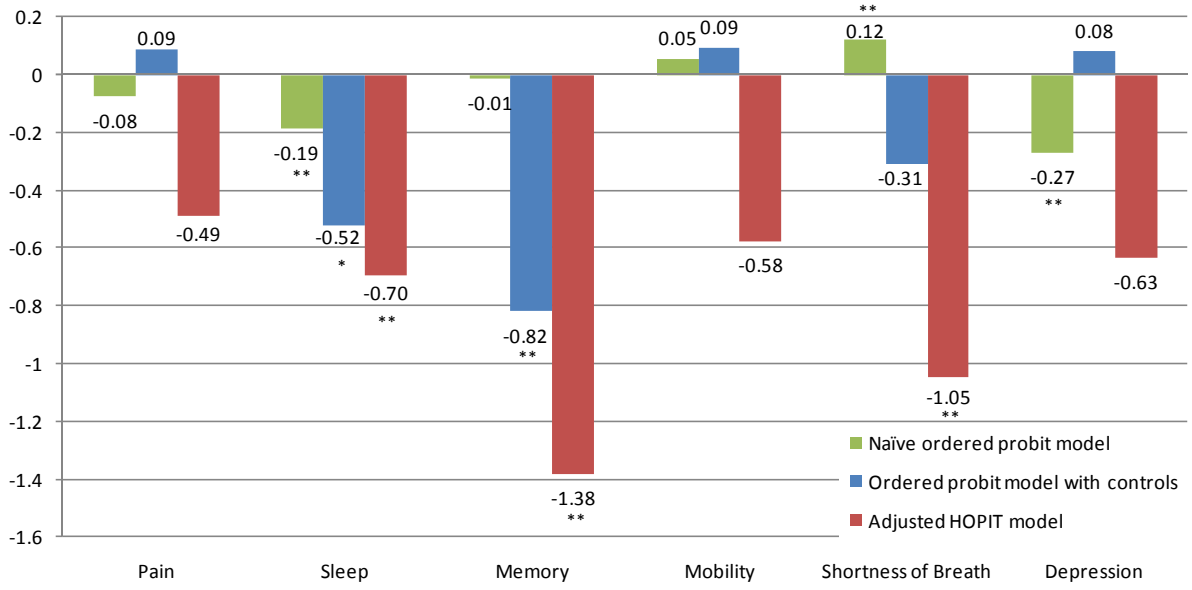
Note: Coefficients from the ordered probit models in Table A1 (Model I1) are shown. Models are run separately for each health domain, and include the following covariates: male respondent, male vignette character, male respondent * age, age, race/ethnicity, and educational attainment. * indicates the coefficient on male respondent * age is significant at the 5% level; ** indicates significance at the 1% level.

Figure 2: Coefficients on vignette character gender from ordered probit models of the severity rating of vignette characters health problems, by race/ethnicity



Note: Coefficients from the ordered probit models in Table A1 (Model I4) are shown. Models are run separately for each health domain, and include the following covariates: male respondent, male vignette character, male vignette character * Hispanic, age, race/ethnicity, and educational attainment. * indicates the coefficient on male vignette character * Hispanic is significant at the 5% level; ** indicates significance at the 1% level.

Figure 3: Coefficient on male respondent in models of self-rated severity of health problems, before and after adjustment for reporting heterogeneity



Note: * indicates significance at the 5% level; ** indicates significance at the 1% level respectively. A positive coefficient corresponds to an increased severity of health problems; negative coefficients indicate lower severity. The naive ordered probit model estimates rating of own health problems on respondent gender, age, educationational attainment, and race/ethnicity. The ordered probit model with controls estimates rating of own health problems on respondent gender, vignette gender, respondent gender * age, vignette gender * race/ethnicity, age, educational attainment, race/ethnicity. The HOPIT model includes the same covariates, with individual cut-points allowed to vary based with covariates.

References

Bago d'Uva, T., O. O'Donnell, et al. (2008). "Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans." Int. J. Epidemiol. **37**(6): 1375-1383.

Bago d'Uva, T., E. van Doorslaer, et al. (2008). "Does Reporting Heterogeneity Bias the Measurement of Health Disparities?" Health Economics **17**: 15.

Bago d'Uva, T., E. van Doorslaer, et al. (2008). "Does reporting heterogeneity bias the measurement of health disparities?" Health Economics **17**(3): 351-375.

Benyamini, Y., T. Blumstein, et al. (2003). "Gender Differences in the Self-Rate Health-Mortality Association: Is it Poor Self-Rated Health That Predicts Mortality or Excellent Self-Rated Health That Predicts Survival?" The Gerontologist **43**(3): 10.

Boardman, J. D. (2004). "Health pessimism among black and white adults: the role of interpersonal and institutional maltreatment. ." Social Science & Medicine **59**(12): 11.

Brunner, R. (2006). "Understanding Gender Factors Affecting Self-Rated Health." Gender Medicine **3**(4): 3.

Bzostek, S., N. Goldman, et al. (2007). "Why do Hispanics in the USA report poor health?" Social Science & Medicine 65(5): 14.

Case, A. and C. Paxson (2005). "Sex Differences in Morbidity and Mortality." Demography **42**: 26.

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

Deeg, D. and D. M. Kriegsman (2003). "Concepts of Self-Rated Health: Specifying the Gender Difference in Mortality Risk." The Gerontologist **43**(3): 11.

DeSalvo, K., N. Bloser, et al. (2005). "Mortality Prediction with a Single General Self-Rated Health Question. A Meta-Analysi." Journal of General Internal Medicine **21**(3): 9.

Etile, F. and C. Milcent (2006). "Income-related reporting heterogeneity in self-assessed health: Evidence from France." Health Economics **15**(9): 17.

Hopkins, D. J. and G. King (2010). "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." Public Opinion Quarterly **74**(2): 22.

Humphries, K. H. and E. Van Doorslaer (2000). "Income-related health inequality in Canada." Social Science & Medicine **50**(5): 9.

Idler, E. L. and Y. Benyamini (1997). "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." Journal of Health and Social Behavior **38**(1): 17.

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

Lindeboom, M. and E. van Doorslaer (2004). "Cut-point shift and index shift in self-reported health." Journal of Health Economics **23**(6): 17.

Mathers, C. D., C. J. Murray, et al. (2003). "Population Health Metrics: Crucial Inputs to the Development of Evidence for Health Policy." Population Health Metrics **1**(6): 4.

Shmueli, A. (2003). "Socio-economic and demographic variation in health and in its measures: the issue of reporting heterogeneity. ." Social Science & Medicine **57**(1): 10.

Spencer, S., R. Schulz, et al. (2009). "Racial differences in self-rated health at similar levels of physical functioning: An examination of health pessimism in the health, aging, and body composition study." Journal of Gerontology: Social Sciences **64**(1): 8.

Appendix

Table A1: Ordered Probit Model of Respondents' Rating of the Severity of Vignette Characters' Health Problems, with Interactions

	Pain				Sleep				Memory			
	(I1)	(I2)	(I3)	(I4)	(I1)	(I2)	(I3)	(I4)	(I1)	(I2)	(I3)	(I4)
	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value
Male respondent	0.440	0.011	0.023	0.022	0.157	-0.066	-0.051	-0.052	0.472	0.030	0.037	0.037
	(0.003)	(0.674)	(0.339)	(0.347)	(0.327)	(0.015)	(0.045)	(0.041)	(0.002)	(0.234)	(0.117)	(0.120)
Male vignette character	-0.044	-0.043	0.127	-0.028	0.116	0.115	0.077	0.099	-0.012	-0.013	-0.262	0.002
	(0.052)	(0.056)	(0.367)	(0.259)	(<0.001)	(<0.001)	(0.611)	(<0.001)	(0.612)	(0.586)	(0.066)	(0.939)
Male respondent * Age	-0.006				-0.003				-0.007			
	(0.004)				(0.189)				(0.003)			
Male respondent * Race/ethnicity: non-Hispanic black		0.105				-0.017				0.021		
		(0.211)				(0.853)				(0.808)		
Male respondent * Race/ethnicity: Hispanic		0.032				0.181				0.061		
		(0.729)				(0.109)				(0.524)		
Male respondent * Race/ethnicity: non-Hispanic other		-0.069				0.066				-0.012		
		(0.620)				(0.667)				(0.939)		
Male vignette character * Age			-0.003				0.001				0.004	
			(0.221)				(0.799)				(0.076)	
Male vignette character * Race/ethnicity: non-Hispanic black				-0.035				0.001				-0.101
				(0.658)				(0.994)				(0.215)
Male vignette character * Race/ethnicity: Hispanic				-0.165				0.235				0.019
				(0.069)				(0.031)				(0.844)
Male vignette character * Race/ethnicity: non-Hispanic other				0.100				-0.117				-0.160
				(0.472)				(0.432)				(0.286)
Race/ethnicity: non-Hispanic black	0.284	0.249	0.285	0.303	-0.023	-0.019	-0.023	-0.024	0.200	0.193	0.201	0.249
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.595)	(0.731)	(0.604)	(0.698)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Race/ethnicity: Hispanic	0.198	0.185	0.198	0.276	-0.092	-0.163	-0.092	-0.215	0.167	0.142	0.167	0.155
	(<0.001)	(0.003)	(<0.001)	(<0.001)	(0.102)	(0.020)	(0.102)	(0.008)	(0.001)	(0.024)	(0.001)	(0.026)
Race: non-Hispanic other	0.164	0.193	0.165	0.117	-0.082	-0.110	-0.081	-0.020	0.017	0.022	0.020	0.100
	(0.019)	(0.045)	(0.019)	(0.270)	(0.275)	(0.245)	(0.276)	(0.846)	(0.817)	(0.815)	(0.788)	(0.379)
Age	0.003	0.000	0.002	0.000	-0.005	-0.006	-0.006	-0.006	0.008	0.005	0.003	0.005
	(0.055)	(0.794)	(0.323)	(0.803)	(0.002)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.042)	(<0.001)
N	13,818	13,818	13,818	13,818	13,821	13,821	13,821	13,821	13,823	13,823	13,823	13,823

	Mobility				Shortness of Breath				Depression			
	(I1)	(I2)	(I3)	(I4)	(I1)	(I2)	(I3)	(I4)	(I1)	(I2)	(I3)	(I4)
	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value	Beta/P-value
Male respondent	0.205	-0.009	0.003	0.002	0.386	-0.007	-0.017	-0.018	0.357	-0.005	-0.008	-0.009
	(0.189)	(0.739)	(0.906)	(0.925)	(0.020)	(0.797)	(0.513)	(0.495)	(0.029)	(0.846)	(0.752)	(0.718)
Male vignette character	0.058	0.058	0.115	0.046	0.173	0.175	0.001	0.208	-0.019	-0.020	0.216	-0.026
	(0.016)	(0.014)	(0.446)	(0.070)	(<0.001)	(<0.001)	(0.995)	(<0.001)	(0.441)	(0.410)	(0.153)	(0.306)
Male respondent * Age	-0.003				-0.006				-0.006			
	(0.191)				(0.014)				(0.025)			
Male respondent * Race/ethnicity: non-Hispanic black		0.010				-0.118				-0.063		
		(0.911)				(0.209)				(0.525)		
Male respondent * Race/ethnicity: Hispanic		0.137				0.070				0.067		
		(0.201)				(0.581)				(0.550)		
Male respondent * Race/ethnicity: non-Hispanic other		-0.026				-0.135				-0.080		
		(0.849)				(0.416)				(0.646)		
Male vignette character * Age			-0.001				0.003				-0.004	
			(0.703)				(0.263)				(0.120)	
Male vignette character * Race/ethnicity: non-Hispanic black				0.093				-0.065				-0.101
				(0.285)				(0.460)				(0.267)
Male vignette character * Race/ethnicity: Hispanic				-0.037				-0.280				0.173
				(0.733)				(0.024)				(0.108)
Male vignette character * Race/ethnicity: non-Hispanic other				0.170				-0.133				0.122
				(0.211)				(0.414)				(0.482)
Race/ethnicity: non-Hispanic black	0.074	0.070	0.075	0.027	0.076	0.118	0.077	0.110	0.069	0.091	0.070	0.118
	(0.088)	(0.188)	(0.086)	(0.664)	(0.089)	(0.031)	(0.086)	(0.088)	(0.134)	(0.096)	(0.132)	(0.071)
Race/ethnicity: Hispanic	-0.031	-0.085	-0.031	-0.015	0.035	0.007	0.034	0.167	-0.084	-0.110	-0.085	-0.175
	(0.564)	(0.238)	(0.565)	(0.832)	(0.573)	(0.931)	(0.584)	(0.030)	(0.127)	(0.105)	(0.122)	(0.041)
Race: non-Hispanic other	-0.077	-0.066	-0.077	-0.158	-0.008	0.049	-0.011	0.055	-0.016	0.018	-0.020	-0.080
	(0.257)	(0.465)	(0.259)	(0.088)	(0.920)	(0.642)	(0.891)	(0.624)	(0.850)	(0.880)	(0.817)	(0.536)
Age	-0.002	-0.003	-0.003	-0.003	-0.003	-0.005	-0.006	-0.005	-0.005	-0.007	-0.005	-0.007
	(0.113)	(0.003)	(0.053)	(0.003)	(0.062)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.002)	(<0.001)
N	13,810	13,810	13,810	13,810	13,818	13,818	13,818	13,818	13,811	13,811	13,811	13,811

Note: models control for age, educational attainment, and race/ethnicity. Robust standard errors are clustered at the respondent level. Health problems in each of the six domains shown were rated on a 5-point scale where 1=none, 2=mild, 3=moderate, 4=severe, and 5=extreme.