

Reporting Heterogeneity and Health Disparities Across Gender and Education Levels: Evidence From Four Countries

Teresa Molina¹

Published online: 24 February 2016
© Population Association of America 2016

Abstract I use anchoring vignettes from Indonesia, the United States, England, and China to study the extent to which differences in self-reported health across gender and education levels can be explained by the use of different response thresholds. To determine whether statistically significant differences between groups remain after adjusting thresholds, I calculate standard errors for the simulated probabilities, largely ignored in previous literature. Accounting for reporting heterogeneity reduces the gender gap in many health domains across the four countries, but to varying degrees. Health disparities across education levels persist and even widen after equalizing thresholds across the two groups.

Keywords Self-reported health · Anchoring vignettes · Gender health disparities · SES–health disparities

Introduction

Understanding health disparities across gender and education levels is crucial for informing policies aimed at reducing such inequalities and for understanding why life choices and outcomes (e.g., human capital investment, occupational choice, marriage, income, or life satisfaction) may differ across these groups. Valid measures of these health inequalities are required, and self-reported health is a relatively simple and widely available measure that can be used. Unfortunately, comparisons of self-reported health can be confounded by the use of different response scales across individuals. In this article,

Electronic supplementary material The online version of this article (doi:10.1007/s13524-016-0456-z) contains supplementary material, which is available to authorized users.

✉ Teresa Molina
tsmolina@usc.edu

¹ Department of Economics and USC-INET, University of Southern California, Kaprielian Hall 300, 3620 S. Vermont Avenue, Los Angeles, CA 90089, USA

I use anchoring vignettes to quantify the extent to which differences in reporting behavior may drive these differences across gender as well as differences across education levels. I draw on data from four countries: the Indonesian Family Life Survey (IFLS), the U.S. Health and Retirement Study (HRS), the English Longitudinal Study of Aging (ELSA), and the China Health and Retirement Longitudinal Study (CHARLS). All these surveys ask respondents to rate their health difficulties from 1 to 5 (where 1 represents the least severe problems and 5 represents the most severe problems) in six domains: mobility, pain, cognition, sleep, affect, and breathing. In addition, for each domain, all surveys ask respondents to rate the health of three hypothetical individuals in order to anchor the respondents' numerical self-reports. These anchoring vignettes allow me to adjust for the use of different response thresholds across gender and education levels using a hierarchical ordered probit (HOPIT) model, enabling comparisons that are not confounded by systematic reporting differences.

In most health domains across countries, I find that gender gaps are reduced after accounting for the use of different thresholds, although less drastically in Indonesia and the United States, where one-half of the domains still reveal significant gender differences after adjustment. In England and China, adjusting for thresholds completely eliminates the gender gap in the majority of domains. This elimination (or reduction) of significant gender differences after adjusting for response thresholds offers a partial explanation for one quite persistent puzzle that has emerged from studies of self-reported health: women have significantly worse self-reported health than men despite the fact that women have lower mortality rates (Case and Paxson 2005; Macintyre et al. 1999; Nathanson 1975; Strauss et al. 1993; Verbrugge 1989). The observed female disadvantage in self-reported health could be driven by their use of different response thresholds when evaluating a person's health. This is not the only possible explanation for the gender paradox¹ or the first time that this particular hypothesis has been proposed (Macintyre et al. 1999; Verbrugge 1989), but this article offers evidence that the use of different response thresholds across men and women can confound gender comparisons of self-reported health because women have a higher bar for considering someone "healthy."

The narrowing or elimination of gender gaps is not a mechanical result of the econometric exercise: when I repeat this analysis to compare individuals of different education levels, I find no evidence of existing differences shrinking. Across all four data sets, I find persistent education differences that do not diminish (and in most cases widen) after adjusting for the use of different thresholds. This finding adds further support to the large literature on the education health gradient,² emphasizing that if anything, differential reporting behavior may result in an underestimation of the strength of the link between education and health.

In addition to offering evidence on the role of reporting behavior in explaining gender and education gaps, this article contributes to the literature on anchoring vignettes by

¹ Mortality selection is one potential reason for the gender paradox, but Strauss et al. (1993) found that adjusting for it reduces but does not eliminate the gender gap in self-reported health. Case and Paxson (2005) found evidence that men and women face different distributions of chronic conditions; and for some conditions, the severity is worse for men than women. The combination of these two findings help explain why women, afflicted with more chronic conditions that are less fatal, may report worse health yet still live longer than men.

² See Cutler and Lleras-Muney (2006) and Grossman (2006) for reviews of the theory and empirical evidence and Vogl (2014) for a review specifically for developing countries.

expanding their use to within-country gender and education differences in four countries. Most of the early anchoring vignettes studies focused on cross-country comparisons: for example, political efficacy in China and Mexico (King et al. 2004) or work disability and life satisfaction in the United States and the Netherlands (Kapteyn et al. 2007, 2010). A more recent strand of literature has used vignettes and the HOPIT model to analyze within-country differences, particularly in self-reported health (Bago d'Uva et al. 2008a, b; Dowd and Todd 2011; Mu 2014). In these studies, any discussion of differences across gender or education levels is usually limited to a comparison of coefficients in a pooled HOPIT model, which allows gender and education to have only a level effect on latent health and response thresholds. Unlike existing work, I estimate the HOPIT model separately for men and women (and separately for more-educated and less-educated individuals) and then simulate self-report distributions using adjusted and unadjusted thresholds to allow for gender and education to change how other covariates affect health and reporting behavior. Kapteyn et al. (2007, 2010) and Mu (2014) all ran the HOPIT model separately for different countries or different regions, but this article is the first to conduct this exercise for gender and education levels. This article is also the first to calculate standard errors for a key estimate: the difference between the simulated proportion of individuals falling into the "healthiest" category in two different groups. Previously ignored in the literature, standard errors allow me to conclude whether groups are statistically different before and after allowing for the use of different response thresholds across groups.

Anchoring Vignettes

Many economic studies have turned to self-reported health measures as outcome variables (Finkelstein et al. 2012; Gertler and Gruber 2002; Maccini and Yang 2009; Manning et al. 1987; Strauss et al. 1993) because objective measures of health are often infeasible for measuring large populations or too narrow to capture the multidimensional nature of health. The particular type of measure studied in this article is a response to a question like, "Overall, in the last 30 days, how much pain or bodily aches did you have?," chosen from five options: none, mild, moderate, severe, or extreme. These self-reports are simple and may be better suited to capturing an individual's health as a whole than are objective measures that are more specific (e.g., blood pressure or BMI) or more extreme (e.g., mortality). Moreover, self-reported health is also strongly linked with objective measures of health. General self-reported health,³ which is slightly different from the measures used in this article, has been repeatedly shown to have a significant relationship with mortality, robust to the inclusion of a host of demographic and socioeconomic controls.⁴

³ General self-reported health is an answer to the question, "In general, how healthy do you feel?" I use domain-specific and not general self-reported health in this article because the standard vignettes have been designed for domain-specific health.

⁴ Idler and Benyamini (1997) reviewed 27 studies conducted in eight countries. With remarkable consistency, these studies showed that the coefficient on general self-rated health in regression on mortality remains significant even when other covariates and health status indicators are included. A more recent meta-analysis by DeSalvo et al. (2006) found that individuals who reported being in "poor" health have almost double the mortality risk of those who reported being in "excellent" health. This calculation included studies that controlled for various covariates, such as age and socioeconomic status (SES).

Despite their advantages, subjective scale measures have also long been the source of some controversy, due to potential differences in reporting behavior across groups. Dow et al. (1997), in their analysis of the effect of health care prices on health outcomes, highlighted that self-reported measures often suffer from reporting bias that is nonrandom, potentially correlated with variables such as income and healthcare usage. Clearly, self-reported measures of health that assign a quantitative value to how healthy one feels are not perfect measures of actual health. They also incorporate an individual's interpretation of the response choices: that is, what do mild, moderate, severe, and extreme really mean?

The idea that individuals may use different reporting thresholds in their self-reports is particularly problematic in comparisons across groups or individuals. The underlying problem is that it is impossible ascertain whether the observed differences are being driven by actual differences in health status or simply the use of different response scales—what King et al. (2004) referred to as “differential item functioning” (DIF), a term originally from the education testing literature.⁵ Also unclear is whether, across groups that appear similar, there exist differences that are masked by different response scales. In short, with systematically different response scales, one must first adjust for this DIF before any valid comparisons can be made. Methods recently developed to make these necessary adjustments involve the use of anchoring vignettes, introduced by King et al. (2004). These vignettes tell a brief story about a hypothetical person and ask respondents to evaluate the severity of the person's situation. For example,

*[John] can concentrate while watching TV, reading a magazine, or playing a game of cards or chess. Once a week he forgets where his keys or glasses are, but finds them within five minutes. Overall how much difficulty did [John] have remembering things?*⁶

A vignette like this one would help anchor respondents' answers to the question: “Overall in the last 30 days, how much difficulty did you have remembering things?” In general, vignettes offer insight into how people set their thresholds and therefore help adjust for differences in response scales.

A simple figure can summarize why comparisons based on subjective scales can be problematic and how anchoring vignettes can be used to address these issues. Figure 1, from King et al. (2004), shows two respondents: A and B. In panel A, $Self_1$ represents A's numerical response to a subjective question like, “How is your health in general?” $Self_2$, in Panel B, represents B's response to this same question. A naive comparison of these two numbers would lead to the conclusion that A is in better health than B. However, these figures also depict how A and B evaluate three hypothetical vignette individuals: Alison, Jane, and Moses. Even though A and B are faced with identical vignette descriptions, they evaluate the three vignettes very differently, indicating the use of potentially different response scales. Panel C shows what B's responses would look like if she had instead used A's response scale. This essentially boils down to

⁵ A test question with DIF is one that two people of the same ability but from different groups (races or genders, for example) have different probabilities of answering correctly.

⁶ This vignette is from the cognition domain and is used in all four data sets in this article. See Online Resource 1 (section A1) for complete list of vignettes.

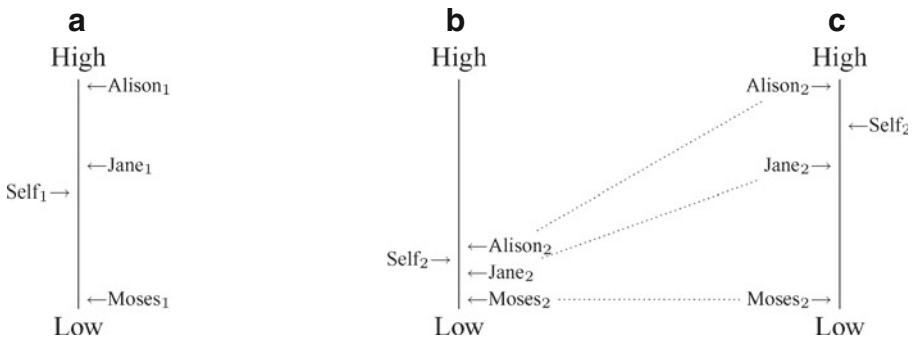


Fig. 1 Comparing subjective scales (from King et al. (2004))

aligning B's vignette evaluations with A's and comparing $Self_1$ and $Self_2$ on the new scale. Comparing panel A and panel C shows that B is actually in better health than A but has a higher bar for defining what is "healthy."

Anchoring vignettes allow inferences about respondents' internal response scales that are otherwise completely unobservable to the researcher. When comparing two groups of individuals, one can use the scale in one group as a benchmark to make valid comparisons. The validity of these comparisons hinges on two important assumptions: (1) *response consistency*, which means that respondents use the same response scales when evaluating themselves and evaluating others; and (2) *vignette equivalence*, which asserts that the way respondents interpret the scenarios and questions are independent of their individual characteristics. In other words, respondents differ only in the thresholds they use—not in how they interpret the question. In the next section, I discuss what both of these assumptions mean in the context of the econometric model.

Response consistency would not hold if, for some reason, the respondents held the hypothetical individuals to a different standard than their own. For example, King et al. (2004) suggested that response consistency in their study of political efficacy would be violated if respondents felt inferior to the people in vignettes and set a higher bar for what it means to have "a lot of say" in the government. Both King et al. (2004) and van Soest et al. (2011) tested for response consistency by using objective measures and found strong evidence to support response consistency. Unfortunately, tests like these are possible only when relevant objective measures, which map directly to the unobserved latent variable, exist.⁷ Although the validity of this assumption may depend on the particular context of the vignettes, I argue that the straightforward nature of the vignettes in this article make this a reasonable assumption for the self-reported health setting. The individuals described in the vignettes in this article suffer from common ailments that are undoubtedly somewhat familiar to respondents in all countries. This familiarity, combined with the fact that health is an issue that these elderly respondents deal with everyday—unlike the political issues in King et al. (2004)—makes it unlikely that respondents would hold the vignette individuals to a different standard or use a different scale to evaluate them.

⁷ For example, King et al. (2004) used vision tests to validate subjective scale questions about vision impairment, and van Soest et al. (2011) used actual counts of alcoholic drinks to validate subjective questions about the severity of drinking problems.

The second assumption, vignette equivalence, would not hold if there are systematic differences in the way respondents interpret the questions or vignettes, which is more likely when dealing with abstract concepts. Because vignettes are brief, vignette equivalence may also be violated if respondents fill in any gaps by making assumptions to create a complete picture. These assumptions are likely to vary by person and are problematic if correlated with individual characteristics. Fortunately, all the vignettes used in this article are straightforward and deal with tangible, familiar concepts. However, because of their brevity, they may be slightly open to interpretation.

Because of the dearth of objective measures that map directly to my domain-specific health variables of interest, as well as the strong support in the literature for the validity of response consistency (Grol-Prokopczyk et al. 2015; King et al. 2004; van Soest et al. 2011), I take this first assumption as given. However, I test for vignette equivalence by using methods proposed by Bago d'Uva et al. (2011).

Econometric Model

To separately identify the effect of individual characteristics on true health from their effect on reporting thresholds, I use the same econometric model used in Kapteyn et al. (2007) and Kapteyn et al. (2010). For each health dimension d , I model the subjective response of an individual i , Y_{di} , in the following ordered response equation, where Y_{di} ranges from 1 (least severe) to 5 (most severe). Y_{di} is determined by a latent variable Y_{di}^* , which is a function of individual respondent characteristics and an error term. For simplicity, I drop the subscript d in the model exposition but analyze a separate model for each health domain in the empirical section.

$$Y_i^* = \mathbf{X}_i\beta + \varepsilon_i; \quad (1)$$

ε_i is $N(0, \sigma_\varepsilon)$, ε_i independent of \mathbf{X}_i , and the other error terms in the model.

$$Y_i = j \text{ if } \tau_i^{j-1} < Y_i^* \leq \tau_i^j, \quad j = 1, \dots, 5. \quad (2)$$

$$\tau_i^0 = -\infty, \quad \tau_i^5 = \infty, \quad \tau_i^1 = \gamma^1 \mathbf{X}_i + u_i, \quad \tau_i^j = \tau_i^{j-1} + e^{\gamma^j \mathbf{X}_i}, \quad j = 2, 3, 4; \quad (3)$$

u_i is $N(0, \sigma_u^2)$ and is independent of \mathbf{X}_i and the other error terms in the model.

What sets this model apart from a normal ordered response model is that the thresholds τ_i^j vary across individuals. These thresholds are also a function of individual characteristics and an unobserved individual effect, u_i , which allow individuals with identical \mathbf{X} characteristics to have different response scale thresholds. The individual-specific thresholds, τ_i^j , are the essence of DIF.

Given data on self-reported health and individual characteristics only, identifying β and γ^1 separately is impossible (but γ^j for $j > 1$ is identified through the nonlinearity of the exponential function). For this, I use the three vignette evaluations given by each respondent for each health domain. The vignette responses (of individual i to vignette

number l for domain d) can be modeled in a similar ordered response framework. Again, the d subscript is omitted. In this article, $l = 1, 2, 3$.

$$Y_{li}^* = \theta_l + \varepsilon_{li}; \quad (4)$$

ε_{li} is $N(0, \sigma_v)$, ε_{li} independent of \mathbf{X}_i and the other error terms in the model.

$$Y_{li} = j \text{ if } \tau_i^{j-1} < Y_{li}^* \leq \tau_i^j, \quad j = 1, \dots, 5. \quad (5)$$

The nonnegative exponential function in threshold Eq. (3) ensures that $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$. Its nonlinearity ends up identifying the γ_j coefficients for $j > 1$. The results in this article use the exponential function to define the gaps between different thresholds, as in Eq. (3). In Online Resource 1, however, I also test the sensitivity of these results by replacing the exponential in Eq. (3) with a square, as follows:

$$\tau_i^0 = -\infty, \quad \tau_i^5 = \infty, \quad \tau_i^1 = \gamma^1 \mathbf{X}_i + u_i, \quad \tau_i^j = \tau_i^{j-1} + (\gamma^j \mathbf{X}_i)^2, \quad j = 2, 3, 4. \quad (3a)$$

I also explore the possibility of using a linear specification for the threshold equations in Online Resource 1. The results remain remarkably consistent across alternate functional forms. This is true for all domains and all four data sets.

The model's first crucial assumption, response consistency, means that the thresholds τ_i in Eq. (3) are used for both the self-reports (Eqs. (1) and (2)) and the vignette responses (Eqs. (4) and (5)). Given that vignette responses Y_{li}^* depend only on individual characteristics through their influence on the thresholds τ_i , it is possible to identify γ and θ vectors from Eqs. (4) and (5). Here, θ_l is a vignette fixed effect that, together with an unobserved individual error ε_{li} , completely determines the latent variable for vignette evaluations, Y_{li}^* .

The assumption of vignette equivalence implies that θ_l is constant across all individuals, and the unobserved error is uncorrelated with individual characteristics. That is, individual characteristics do not affect the perceived underlying severity of the each vignette. Respondent characteristics can affect evaluations of vignettes only through their effect on thresholds. This leads naturally to a test of vignette equivalence, which involves including respondent characteristics \mathbf{X}_i in vignette Eq. (4). I discuss this vignette equivalence check in section A5 of Online Resource 1. Like Bago d'Uva et al. (2011) (who developed this test) and Grol-Prokopczyk et al. (2015) (who applied the same methods), I find evidence that vignette equivalence is not always satisfied. However, adjusting the model to allow for violations does not significantly change my coefficient estimates and therefore my conclusions.

Data

I use data from the 2007 wave of the IFLS (Strauss et al. 2009); the 2007 Disability Vignette Study mail survey from the HRS (HRS 2014); the 2006–2007 wave of the ELSA (Marmot et al. 2014); and the first wave of the CHARLS, conducted in 2011

(Zhao et al. 2013). Each of these four data sets includes the following domain-specific self-reported health questions:

Overall in the last 30 days. . .

1. How much of a problem did you have with moving around?
2. How much pain or bodily aches did you have?
3. How much difficulty did you have remembering things?
4. How much difficulty did you have with sleeping, such as falling asleep, waking up frequently during the night, or waking up too early in the morning?
5. How much of a problem did you have with feeling sad, low, or depressed?
6. How much of a problem did you have because of shortness of breath?

In addition to these questions, all four surveys include the exact same set of three vignettes per health domain (see section A1 in Online Resource 1 for a list all of the vignettes). The inclusion of all six of the same health domains and the use of identical vignettes across the four data sets make this combination of data sets particularly appealing. Moreover, unlike several other surveys that also include vignettes, all these data sets either focus on the elderly or have a large enough sample of elderly individuals to estimate the HOPIT model separately for different subgroups within the elderly population, which is the group likely to be the most familiar with the health problems discussed in the vignettes. Focusing on this narrow (and arguably more relevant) age range allows me to hone in on sources of reporting heterogeneity other than age.

Answers to the health status questions and anchoring vignettes form the outcome variables of interest for this analysis: domain-specific Y_i , Y_{1i} , Y_{2i} , and Y_{3i} in the HOPIT model. For the explanatory variables X_i , I purposely focus on a simple set of variables in order to facilitate comparisons across the data sets: gender, age, and education levels. Specifically, I create two age dummy variables (for those aged 56–70 and those older than 70, leaving those 55 and younger as the omitted category) and a dummy variable for males. Because I eventually split each sample into high- and low-education groups, I define different education dummy variables for each data set in order to have groups that are large enough (see upcoming Table 1 for category descriptions).

Although all data sets include the same self-report questions and anchoring vignettes, there are some important differences in the way the information was collected. For example, the IFLS and CHARLS were in-person surveys, while the ELSA and HRS involved written questionnaires for the vignettes. The appendix contains more information about the individual data sets.

Summary Statistics

Table 1 lists summary statistics for all four data sets, including only individuals who responded to the self-report and three vignette evaluations for at least one of the domains and who were not missing any of the other covariates of interest. Each survey represents one cross section of data, with the IFLS and HRS sampled in 2007, the ELSA sampled during 2006 and 2007, and the CHARLS sampled in 2011. For the IFLS and CHARLS, the sample sizes

Table 1 Summary statistics

| | IFLS (1) | HRS (2) | ELSA (3) | CHARLS (4) |
|--|-------------------|-------------------|--------------------|--------------------|
| Age | 52.000 (9.618) | 63.760 (9.046) | 65.800 (10.300) | 59.420 (10.040) |
| 1(Male) | 0.535 (0.499) | 0.453 (0.498) | 0.461 (0.499) | 0.466 (0.499) |
| 1(High Education Group) ^a | 0.218 (0.413) | 0.281 (0.449) | 0.358 (0.480) | 0.364 (0.481) |
| 1(Medium Education Group) ^b | 0.436 (0.496) | 0.570 (0.495) | 0.225 (0.418) | 0.214 (0.410) |
| Mobility Self-Report | 1.430 (0.848) | 1.742 (0.910) | 1.644 (0.944) | 1.296 (0.801) |
| Pain Self-Report | 1.815 (1.027) | 2.366 (0.871) | 2.288 (0.932) | 1.872 (1.116) |
| Cognition Self-Report | 1.687 (0.989) | 1.834 (0.776) | 1.801 (0.815) | 1.711 (1.011) |
| Affect Self-Report | 1.678 (1.034) | 2.309 (0.922) | 2.278 (1.044) | 1.727 (1.058) |
| Sleep Self-Report | 1.473 (0.896) | 1.777 (0.876) | 1.583 (0.836) | 1.476 (0.877) |
| Breathing Self-Report | 1.282 (0.727) | 1.450 (0.772) | 1.408 (0.782) | 1.336 (0.786) |
| Average Pairwise Correlation | .39 | .42 | .34 | .34 |
| Year of Vignette Survey | 2007 | 2007 | 2006–2007 | 2011 |
| Number of Observations | 3,058 | 4,158 | 2,192 | 3,630 |

Notes: Standard errors are shown in parentheses. Self-reports are reported on a scale from 1 to 5, with 1 representing the least and 5 the most severe health difficulties. All data are weighted using individual cross-sectional sampling weights provided by each data set to make summary statistics representative of the United States for the HRS, England for the ELSA, China for the CHARLS, and the 13 IFLS provinces in Indonesia for the IFLS.

^a IFLS: high school graduates; HRS: college graduates; ELSA: A-levels and above; CHARLS: junior high and above.

^b IFLS: primary but not high school; HRS: high school but not college; ELSA: any degree lower than A-levels; CHARLS: primary but not junior high.

reported here are much larger than the sample sizes in each individual domain because individuals responded to only two domains each.⁸

Although *t* tests are not reported here, large and significant differences exist across all four countries that arise from differences in survey parameters, covariate distributions within each country, or a combination thereof. For instance, the HRS and ELSA samples are older, on average, which could be partly due to the higher life expectancies

⁸ See the [appendix](#) for more detail.

in these two countries but is likely driven primarily by the higher age threshold for inclusion in these data sets: 50, compared with 40 in the IFLS and 45 in the CHARLS.⁹ Rather than drop all IFLS and CHARLS respondents younger than 50, I include everyone and control for age in order to retain as many observations as possible. The longer life expectancy of females relative to males is reflected in the fact that less than one-half of the population is male in all samples except the IFLS (which is also the youngest sample). This disproportionate female share is particularly apparent in the older HRS and ELSA samples, which have significantly higher female proportions than the other two—again, most likely an artifact of the survey design but potentially also generated by demographic differences across countries.

The education statistics must be interpreted with caution because, as described earlier, the “high education,” “medium education,” and “low education” category definitions differ across the samples and are roughly equivalent to using the 75th percentile as the high education cutoff. Keeping this in mind, large differences in the levels of educational attainment across countries clearly emerge. More than 80 % of the American sample are high school graduates; this figure is less than one-quarter for Indonesian respondents, an older cohort in a developing country. In the CHARLS sample, less than 10 % of the sample graduated from high school. More than one-third (36 %) of the ELSA sample received their A-levels or higher, which is a slightly more advanced qualification than high school graduation in the United States.

Table 1 also lists the self-report means for each health domain, and the average of all pairwise correlations between self-reports for different domains. The correlations are positive but weak for all four data sets. For IFLS and CHARLS respondents, all self-report means fall between 1 (“no difficulty”) and 2 (“mild difficulty”). Pain and (to a lesser extent) cognition appear to be the most serious afflictions for these two groups. The U.S. sample reports the worst health on average across all domains; pain and affect appear to be the most serious problems for this group. These are also the two most serious afflictions for the ELSA sample, whose self-report averages are almost on the same level as those of the HRS. Given the significant differences in covariates across groups, the different formats and languages of the surveys—and of course, the possibility of different response thresholds across countries—it is difficult to use these raw differences in self-reports to draw any conclusions about the relative true health levels of these countries.¹⁰

Table 2 reports the responses to the hypothetical vignettes for each sample and each domain. I report the domain-specific sample size at the bottom of each column. Here, I number the vignettes in order of increasing intended severity based on the IFLS sample and questionnaire.¹¹ In all samples, the average perceptions of severity are

⁹ The HRS, ELSA, and CHARLS are all aging data sets focused on the elderly, while the IFLS is a household survey that interviews all members of a sample household. The vignettes in the IFLS, however, were targeted only to those 40 and older.

¹⁰ See Molina (2014).

¹¹ The vignettes in the IFLS are grouped by domain and within each domain appear to be ordered with the least severe vignettes at the beginning and the most severe at the end. For most domains, the ordering is quite clear, while domains like cognition and sleep are more open to interpretation. However, the data confirm that the relative severity perceived by IFLS respondents is consistent with the ordering of vignettes in the interview.

Table 2 Vignette responses

| | Mobility | Pain | Cognition | Sleep | Affect | Breathing |
|------------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| IFLS | | | | | | |
| Vignette 1 | 2.352 (1.047) | 2.525 (1.006) | 2.536 (1.000) | 2.712 (1.018) | 2.508 (0.966) | 2.794 (1.064) |
| Vignette 2 | 2.843 (1.065) | 2.726 (0.971) | 2.884 (1.050) | 3.058 (1.042) | 3.025 (1.002) | 3.330 (1.056) |
| Vignette 3 | 3.520 (1.081) | 3.457 (1.076) | 3.175 (1.093) | 3.396 (1.094) | 3.703 (1.175) | 3.758 (1.142) |
| Number of observations | 1,003 | 1,027 | 1,018 | 1,122 | 944 | 996 |
| HRS | | | | | | |
| Vignette 1 | 2.461 (0.722) | 1.902 (0.652) | 1.948 (0.735) | 3.030 (0.721) | 2.567 (0.693) | 3.092 (0.769) |
| Vignette 2 | 3.708 (0.817) | 3.187 (0.739) | 2.796 (0.769) | 3.852 (0.837) | 3.357 (0.762) | 3.973 (0.804) |
| Vignette 3 | 3.834 (0.802) | 3.790 (0.775) | 3.776 (0.759) | 3.858 (0.780) | 4.532 (0.761) | 4.382 (0.767) |
| Number of observations | 4,118 | 4,123 | 4,127 | 4,126 | 4,113 | 4,119 |
| ELSA | | | | | | |
| Vignette 1 | 2.485 (0.770) | 1.967 (0.569) | 2.098 (0.680) | 2.994 (0.718) | 2.627 (0.709) | 3.197 (0.789) |
| Vignette 2 | 3.616 (0.878) | 3.035 (0.733) | 2.888 (0.745) | 3.649 (0.890) | 3.274 (0.777) | 3.865 (0.816) |
| Vignette 3 | 3.860 (0.796) | 3.902 (0.785) | 3.690 (0.834) | 3.582 (0.778) | 4.318 (0.840) | 4.434 (0.808) |
| Number of observations | 2,115 | 2,145 | 2,121 | 2,148 | 2,088 | 2,085 |
| CHARLS | | | | | | |
| Vignette 1 | 1.758 (0.902) | 2.080 (0.784) | 1.826 (0.873) | 2.333 (0.930) | 2.107 (0.863) | 2.708 (1.085) |
| Vignette 2 | 2.393 (1.067) | 2.075 (0.792) | 2.504 (0.927) | 3.167 (1.163) | 2.730 (0.937) | 3.454 (1.060) |
| Vignette 3 | 3.532 (0.991) | 3.263 (0.940) | 2.626 (1.058) | 3.054 (0.979) | 3.822 (1.075) | 3.933 (1.095) |
| Number of observations | 1,067 | 1,045 | 1,136 | 1,155 | 1,116 | 1,082 |

Notes: Standard errors are shown in parentheses. Vignettes are evaluated on a scale from 1 to 5, with 1 representing the least severe and 5 the most severe health difficulties. All data are weighted using individual cross-sectional sampling weights provided by each data set to make summary statistics representative of the United States for the HRS, England for the ELSA, China for the CHARLS, and the 13 IFLS provinces in Indonesia for the IFLS.

generally in accord with the intended relative levels. With the exception of the sleep domain (which is one of the least straightforward of all vignette domains) for the ELSA and CHARLS samples and the pain domain for the CHARLS,

the first vignette is, on average, rated healthier than the second, which in turn is rated healthier than the third.¹²

As shown in Figs. 2 and 3 in the appendix, there are substantial within-country differences in self-reported health across gender and education. For all data sets, at least three domains show significantly different distributions for men and women, and in at least four domains, highly educated and less-educated individuals have significantly different distributions. I investigate these differences using the HOPIT model discussed earlier, which I estimate using the methods described in the following section.

Estimation Strategy

Estimating the Model

I use maximum likelihood to estimate the model described in the Econometric Model section. Details about the estimation procedure, as well as the likelihood function, can be found in section A2 of Online Resource 1. I estimate the model separately for each data set and health domain given that common response scales across health domains is a strong assumption (Kapteyn et al. 2007). To simulate distributions by subgroup, I also estimate the model separately for males and females, and then for high-education and pooled medium- and low-education individuals (which I refer to for the remainder of the article as the “lower-education” category). For the gender analysis, my specification includes the following in the vector \mathbf{X}_i : two age dummy variables, one dummy variable for high education, and one for medium education, which essentially breaks down the sample into three groups, where the omitted category is the lower-education group. I also include interactions between the age and education dummy variables. For the education analysis, \mathbf{X}_i includes the age dummy variables, a male dummy variable, and the age-gender interactions.¹³

Simulating Distributions and Standard Errors for Predicted Probabilities

Using the coefficients from the separately estimated models, I simulate the distribution of self-reports for the separate groups in several ways. I simulate the distribution of domain-specific self-reported health separately for males (high-education individuals) using their own thresholds, females (lower-education individuals) using their own thresholds, and then males (high-education) using female (lower-education) thresholds. As a summary measure for each simulated distribution, I calculate the simulated proportion of males and females (or high- and lower-education groups) who fall into the healthiest category. Therefore, to analyze the differences between groups, I can look

¹² In these three exceptions, the differences in average ratings are very small in magnitude. Note that my arbitrarily chosen ordering is irrelevant to the estimation of the model because the θ_{ij} , which capture the actual ordering of perceived severity, are directly estimated.

¹³ In Online Resource 1, I estimate both an ordered probit model and a HOPIT model on the entire IFLS sample to illustrate importance of accounting for reporting heterogeneity. For pooled analyses of the HRS, ELSA, and CHARLS vignettes, see Dowd and Todd (2011), Bago d’Uva et al. (2011), and Mu (2014), respectively. Dowd and Todd (2011) and Bago d’Uva et al. (2011) used the same data I use here, whereas Mu (2014) used the pilot wave of the CHARLS. I use a slightly different specification from these studies.

at two estimates. The first is the difference between the simulated proportion of males and females (or high- vs. lower-education groups) in the healthiest category, calculated using their own group's coefficients estimated from the model. The second comparison is the difference between the simulated proportion of healthy males predicted using female thresholds and the simulated proportion of healthy females using female thresholds. This can be thought of as a DIF-adjusted gender comparison, and an analogous analysis can be conducted to compare high- and lower-education groups. This DIF-adjusted comparison illustrates how different the two groups would be if they used the same reporting thresholds.

In previous literature that has conducted these simulations, most analysis and interpretation has been conducted by simply comparing the distributions calculated using own-group thresholds and then the same thresholds for both groups. Without standard errors, however, it is difficult to draw definitive conclusions about how much the thresholds matter and whether significant differences still exist after adjustment. In order to conduct statistical inference, I analytically calculate standard errors for the two differences described earlier. See section A3 of Online Resource 1 for greater detail about the derivations of all the formulas used.

Results

Simulations

In this section, I discuss the simulation results by gender and by education for each of the four data sets. Table 3 reports the results of various simulations that compare males with females. Each panel summarizes the results from a different data set, and each column represents a different domain. Every cell in the table reports the same summary measure of the simulated distribution: the proportion of individuals (in the given subgroup, either in the raw data or simulated using the specified parameters) that fall into the healthiest category (corresponding to a self-report response of 1).

In Table 3, the first row for each survey simply reports the proportion of ones in the raw data for men's self-reports, and the last row reports the proportion among women. These reflect the same numbers represented graphically in Fig. 2 in the appendix. The second row for each survey uses the coefficients estimated using the male-specific HOPIT model to simulate the distribution of self-reports. Taking the explanatory variables for males as given, I use the male-specific coefficients to predict the proportion of the male sample in each self-report category and report the proportion in the healthiest category. The fourth row conducts the same exercise for the female sample. Row 3 is the most informative. These calculations once again take the *male* explanatory variables and β coefficients as given, but instead use the *female* thresholds (γ coefficients) to predict the distribution of self-reports among men. This approach essentially predicts what the male distribution would look like if they had the same thresholds as women.

In the IFLS and ELSA data, the third row narrows the gap between males (row 2) and females (row 4) in all domains. In the HRS, the gap is narrowed for cognition,

Table 3 Simulated proportion falling in healthiest category, by gender

| | Mobility (%) (1) | Pain (%) (2) | Cognition (%) (3) | Sleep (%) (4) | Affect (%) (5) | Breathing (%) (6) |
|---|---------------------|-----------------|----------------------|------------------|-------------------|----------------------|
| IFLS | | | | | | |
| 1 Male sample raw data | 76.16 | 54.95 | 65.25 | 66.96 | 78.53 | 85.28 |
| 2 Male sample using male thresholds | 75.15 | 54.99 | 61.87 | 66.03 | 77.10 | 84.04 |
| 3 Male sample using female thresholds | 71.00 | 54.31 | 59.06 | 62.10 | 76.10 | 82.18 |
| 4 Female sample using female thresholds | 69.74 | 44.98 | 53.18 | 54.09 | 67.50 | 84.46 |
| 5 Female sample raw data | 71.70 | 44.67 | 54.15 | 55.08 | 68.04 | 85.60 |
| HRS | | | | | | |
| 1 Male sample raw data | 51.07 | 14.96 | 38.43 | 21.32 | 50.81 | 68.82 |
| 2 Male sample using male thresholds | 53.16 | 17.73 | 40.45 | 25.92 | 53.64 | 70.37 |
| 3 Male sample using female thresholds | 58.77 | 18.60 | 33.76 | 26.20 | 44.76 | 66.29 |
| 4 Female sample using female thresholds | 52.54 | 15.40 | 36.14 | 21.30 | 44.18 | 70.96 |
| 5 Female sample raw data | 50.99 | 12.29 | 35.10 | 17.25 | 42.06 | 69.09 |
| ELSA | | | | | | |
| 1 Male sample raw data | 64.64 | 24.19 | 43.39 | 34.07 | 65.59 | 76.19 |
| 2 Male sample using male thresholds | 65.40 | 25.28 | 45.44 | 36.92 | 67.96 | 78.05 |
| 3 Male sample using female thresholds | 63.38 | 18.59 | 36.73 | 34.94 | 57.49 | 67.17 |
| 4 Female sample using female thresholds | 60.49 | 18.13 | 42.07 | 24.23 | 56.26 | 73.56 |
| 5 Female sample raw data | 59.56 | 17.36 | 40.59 | 22.93 | 54.63 | 71.90 |
| CHARLS | | | | | | |
| 1 Male sample raw data | 85.80 | 59.02 | 70.03 | 66.94 | 75.36 | 83.00 |
| 2 Male sample using male thresholds | 86.37 | 57.35 | 64.33 | 68.01 | 75.19 | 82.25 |
| 3 Male sample using female thresholds | 85.03 | 50.89 | 61.44 | 59.83 | 70.15 | 86.76 |
| 4 Female sample using female thresholds | 82.61 | 48.69 | 52.79 | 52.71 | 66.45 | 78.35 |
| 5 Female sample raw data | 83.55 | 52.03 | 57.45 | 54.53 | 65.93 | 76.50 |

Notes: Individual cross-sectional sampling weights are used. Proportions are calculated using coefficients from a HOPIT specification with the following explanatory variables: two age dummy variables, 1(High Ed), 1(Medium Ed), and all age-education interactions.

affect, and breathing, but widened in mobility, pain, and sleep. In the CHARLS, the gender gap is close to eliminated in the pain domain and is narrowed in several others. In general, the significance of the reductions or increases that take place is unclear.

Table 4, which summarizes the results of this same analysis conducted instead to compare high-education with lower-education individuals, shows a more universal pattern across countries. Across the overwhelming majority of domains and data sets, using the same thresholds for both groups does not narrow the education gap—and in fact, seems to widen it. In all domains for the IFLS and HRS and at least four domains in the CHARLS and ELSA, the numbers in row 3 are of larger magnitude than those in row 2, indicating that the proportion of high-education individuals falling into the

Table 4 Simulated proportion falling in healthiest category, by education level

| | Mobility (%) (1) | Pain (%) (2) | Cognition (%) (3) | Sleep (%) (4) | Affect (%) (5) | Breathing (%) (6) |
|---|------------------------|--------------------|-------------------------|---------------------|----------------------|-------------------------|
| IFLS | | | | | | |
| 1 High-ed. sample raw data | 84.39 | 55.40 | 61.67 | 68.13 | 74.69 | 86.53 |
| 2 High-ed. sample using high-ed. thresholds | 81.49 | 57.55 | 58.82 | 67.67 | 74.28 | 84.02 |
| 3 High-ed. sample using lower-ed. thresholds | 89.49 | 69.78 | 68.72 | 68.63 | 81.01 | 88.80 |
| 4 Lower-ed. sample using lower-ed. thresholds | 70.58 | 48.13 | 57.48 | 58.46 | 72.47 | 84.18 |
| 5 Lower-ed. sample raw data | 71.52 | 48.58 | 59.68 | 59.46 | 73.44 | 85.11 |
| HRS | | | | | | |
| 1 High-ed. sample raw data | 64.23 | 18.13 | 45.72 | 23.65 | 54.10 | 79.89 |
| 2 High-ed. sample using high-ed. thresholds | 64.51 | 21.01 | 46.07 | 27.13 | 55.70 | 79.88 |
| 3 High-ed. sample using lower-ed. thresholds | 70.90 | 30.50 | 64.45 | 45.42 | 63.12 | 90.03 |
| 4 Lower-ed. sample using lower-ed. thresholds | 48.23 | 14.37 | 34.59 | 22.53 | 45.70 | 67.12 |
| 5 Lower-ed. sample raw data | 45.86 | 11.70 | 33.06 | 17.32 | 42.89 | 64.70 |
| ELSA | | | | | | |
| 1 High-ed. sample raw data | 70.45 | 24.73 | 49.08 | 28.28 | 61.65 | 81.27 |
| 2 High-ed. sample using high-ed. thresholds | 70.95 | 25.54 | 50.65 | 30.20 | 63.22 | 82.13 |
| 3 High-ed. sample using lower-ed. thresholds | 68.05 | 29.47 | 61.55 | 46.33 | 79.62 | 84.34 |
| 4 Lower-ed. sample using lower-ed. thresholds | 58.19 | 19.28 | 39.91 | 30.51 | 61.01 | 72.13 |
| 5 Lower-ed. sample raw data | 57.03 | 18.15 | 37.79 | 27.99 | 58.62 | 69.67 |
| CHARLS | | | | | | |
| 1 High-ed. sample raw data | 92.04 | 64.31 | 75.31 | 71.37 | 80.44 | 86.79 |
| 2 High-ed. sample using high-ed. thresholds | 91.22 | 62.09 | 68.50 | 70.49 | 80.24 | 87.83 |
| 3 High-ed. sample using lower-ed. thresholds | 87.91 | 63.15 | 63.25 | 72.17 | 85.82 | 92.96 |
| 4 Lower-ed. sample using lower-ed. thresholds | 80.29 | 47.18 | 50.99 | 52.99 | 64.28 | 75.51 |
| 5 Lower-ed. sample raw data | 80.07 | 49.94 | 55.13 | 53.08 | 63.86 | 75.33 |

Notes: “Lower-Ed” pools both the medium and low education categories. Individual cross-sectional sampling weights are used. Proportions are calculated using coefficients from a HOPIT specification with the following explanatory variables: two age dummy variables, 1(Male), and all age-gender interactions.

healthiest category increases when predicted using the same thresholds as lower-education individuals. This result happens because high-education individuals usually have a lower first threshold: although they may be healthier than lower-education individuals, they are also less likely to categorize themselves or others as having no difficulty with a particular health problem,¹⁴ resulting in an understatement of differences across education levels.

¹⁴ A specific example is discussed in more detail in Online Resource 1, section A4.1.

Standard Errors for Simulated Probabilities

The preceding discussion about the importance of response thresholds is based on simply comparing one simulated proportion with another, without considering statistical significance. Not only are the simulated proportions calculated from estimated parameters, but they are also calculated using the distribution of covariates in a sample of the true population. For many comparisons, including some of the education comparisons discussed here, standard errors may be less important because definitive conclusions can be drawn without them. For the domains where significant education differences existed in the raw data, if adjusting for DIF widens the difference between the proportion of high-education and lower-education individuals that fall into the healthiest category, it is clear that the use of different thresholds at the very least does nothing to explain the education gap—and at most, it masks even larger differences.

However, certain types of analysis, such as that of the gender gap, require more subtlety. For instance, in the sleep domain of the IFLS, where using female thresholds to predict male distributions appeared to narrow the gender gap slightly but not completely (dropping the male proportion of 66 % to 62 %, bringing it closer to but still somewhat higher than the female proportion of 54 %), it is unclear whether males and females remain significantly different even after the same thresholds are used. The opposite problem exists with, for example, the mobility domain of the HRS, where the groups seemed similar initially but diverged when the same thresholds were used. This second issue is also relevant to some education comparisons, for which differences appeared trivial to begin with and widened after the DIF adjustment.

To assess the statistical significance of the differences between subgroups, before and after accounting for thresholds, I calculate standard errors for two differences: (1) the difference between the male (high-education) proportion in the healthiest category, predicted using male (high-education) thresholds, and the female (lower-education) proportion in the healthiest category, predicted using female (lower-education) thresholds (row 2 minus row 4 in Tables 3 and 4); (2) the difference between the male (high-education) proportion in the healthiest category, predicted using female (lower-education) thresholds, and the female (lower-education) proportion using female (lower-education) thresholds: row 3 minus row 4 of Tables 3 and 4. The formulas for the estimated variances are in Online Resource 1 (section A3, Eq. (A11) for the gender differences, and Eq. (A12) for the education differences).

In Tables 5 and 6, I report (respectively) gender and education differences, along with their respective standard errors and *t* statistics, for differences calculated using group-specific thresholds and differences calculated using the same thresholds for both subgroups. Each panel represents a different data set, and each row represents a different domain. Perhaps the most informative comparisons to make are between columns 3 and 6. Those comparisons indicate whether significant differences between gender and education exist before adjustment for DIF and after adjustment for DIF.

The gender results reported in Table 5 reveal an important role for reporting behavior in explaining the gender gap, particularly in the ELSA and CHARLS. In the ELSA, five domains show significant differences before adjustment, but only one (sleep) remains significant after the same thresholds are used to simulate the probabilities. In

Table 5 Standard errors and *t* statistics for simulated gender differences

| Domain | Using Different Thresholds | | | Using Same Thresholds | | |
|---------------|----------------------------|-----------|---------------------------|--------------------------|-----------|---------------------------|
| | Gender Difference (1) | SE (2) | <i>t</i> Statistic (3) | Gender Difference (4) | SE (5) | <i>t</i> Statistic (6) |
| IFLS | | | | | | |
| Mobility | 0.0541 | 0.0339 | 1.5972 | 0.0126 | 0.0445 | 0.2841 |
| Pain | 0.1001 | 0.0327 | 3.0640** | 0.0934 | 0.0396 | 2.3550* |
| Cognition | 0.0868 | 0.0334 | 2.5970** | 0.0588 | 0.0379 | 1.5497 |
| Sleep | 0.1193 | 0.0316 | 3.7710** | 0.0801 | 0.0332 | 2.4120* |
| Affect | 0.0960 | 0.0350 | 2.7470** | 0.0860 | 0.0427 | 2.0140* |
| Breathing | -0.0042 | 0.0308 | -0.1356 | -0.0227 | 0.0376 | -0.6048 |
| HRS | | | | | | |
| Mobility | 0.0062 | 0.0199 | 0.3098 | 0.0623 | 0.0306 | 2.0360* |
| Pain | 0.0233 | 0.0122 | 1.9060 [†] | 0.0319 | 0.0180 | 1.7740 [†] |
| Cognition | 0.0431 | 0.0183 | 2.3500* | -0.0238 | 0.0340 | -0.6997 |
| Sleep | 0.0462 | 0.0150 | 3.0850** | 0.0490 | 0.0235 | 2.0820* |
| Affect | 0.0946 | 0.0191 | 4.9490** | 0.0058 | 0.0394 | 0.1482 |
| Breathing | -0.0059 | 0.0208 | -0.2863 | -0.0467 | 0.0453 | -1.0308 |
| ELSA | | | | | | |
| Mobility | 0.0491 | 0.0218 | 2.2570* | 0.0290 | 0.0331 | 0.8738 |
| Pain | 0.0715 | 0.0178 | 4.0110** | 0.0046 | 0.0227 | 0.2037 |
| Cognition | 0.0337 | 0.0221 | 1.5284 | -0.0533 | 0.0440 | -1.2122 |
| Sleep | 0.1270 | 0.0202 | 6.2870** | 0.1072 | 0.0275 | 3.9000** |
| Affect | 0.1170 | 0.0211 | 5.5420** | 0.0123 | 0.0497 | 0.2473 |
| Breathing | 0.0449 | 0.0194 | 2.3080* | -0.0640 | 0.0518 | -1.2336 |
| CHARLS | | | | | | |
| Mobility | 0.0376 | 0.0552 | 0.6820 | 0.0242 | 0.0593 | 0.4076 |
| Pain | 0.0866 | 0.0507 | 1.7080 [†] | 0.0220 | 0.0502 | 0.4387 |
| Cognition | 0.1153 | 0.0679 | 1.6980 [†] | 0.0865 | 0.0679 | 1.2748 |
| Sleep | 0.1531 | 0.0535 | 2.8590** | 0.0712 | 0.0524 | 1.3599 |
| Affect | 0.0874 | 0.0504 | 1.7350 [†] | 0.0370 | 0.0597 | 0.6198 |
| Breathing | 0.0390 | 0.0524 | 0.7448 | 0.0841 | 0.0565 | 1.4876 |

Notes: “Gender Difference” is the difference between the proportion of males in the healthiest category and the proportion of females in the healthiest category. Simulated proportions are calculated using coefficients from a HOPIT specification with the following explanatory variables: two age dummy variables, 1(High Education Group), 1(Medium Education Group), and all age-education interactions. Standard errors are calculated analytically using the formulas in Online Resource 1, section A3.

[†] $p < .10$; * $p < .05$; ** $p < .01$

the CHARLS data, four domains start out with differences significant at the 10 % level, but none remain significant after I adjust for DIF. For these two data sets, reporting differences are clearly driving the majority of the significant gender differences that show up in naive comparisons.

On the other hand, in the IFLS, the differences in pain, sleep, and affect remain significant even after adjustment, although all the differences are narrowed. In the HRS, significant differences in mobility, pain, and sleep remain even after I adjust for thresholds. Interestingly, the significant difference in the mobility domain arises only after I adjust for thresholds, suggesting that DIF in this case distorts naive comparisons by masking existing differences instead of generating spurious ones. It is surprising that the English and Chinese appear more similar (in terms of the absence of gender differences after adjustment) than the English and Americans or the Indonesians and Chinese, which represent pairings of countries at more similar stages of economic development.

Nevertheless, the narrowing or elimination of gender gaps as a general result is broadly consistent with findings from studies that analyzed biomarkers and other objective health measures from these data sets. For example, in CHARLS data, the magnitude of the female disadvantage in hypertension, diabetes, depression, and cognition measures is much smaller than the magnitude of their disadvantage in self-reported health (Zhao et al. 2012). For cognition specifically, Lei et al. (2013) found that the significant female disadvantage in objective measures is almost completely explained (for mental intactness) or completely explained (for episodic memory) by differences in education levels.

Crimmins et al. (2010) looked at gender differences in the prevalence of various conditions in HRS and ELSA data and found that women are significantly more likely to have certain disabling conditions (like arthritis or depressive symptoms) than men. Although this conclusion is consistent with my result that HRS gender differences remain significant after adjustment, it seems contradictory to the result that most ELSA gender differences do disappear after adjustment. However, each domain self-report potentially takes into account a number of conditions: some conditions that afflict women more (hypertension and functional limitations) as well as conditions that are more prevalent among men (heart problems, stroke, and diabetes). As a result, the significance, sign, and magnitude of a gender difference in self-reported health is partly driven by the relative severities and prevalences of the two sets of conditions. In the United States, for example, there is a much higher prevalence of hypertension and functional limitation than in the ELSA (Crimmins et al. 2010), which could explain why, for example, women are significantly worse off than men with regard to the pain domain in the HRS but not in the ELSA.¹⁵ Potential explanations aside, this discussion highlights an important point: what is captured by self-reported health is not necessarily the same as what is captured by more objective measures like disease prevalence rates.

Table 6 tells a more straightforward story. On the whole, education differences in reporting behavior appear to be masking larger underlying differences between the two groups. In the IFLS, although only three domains show significant education differences before adjustment, using the same thresholds to adjust for DIF reveals significant differences in an additional domain (cognition). Similarly, in the ELSA data, unadjusted significant differences exist only in four, but significant differences in the adjusted proportions exist in all six. For the HRS, significant differences are found both before and after adjustment in all six

¹⁵ Although hypertension itself may not result in more pain, related conditions, such as obesity or inactivity, might. I thank an anonymous reviewer for making this point.

Table 6 Standard errors and *t* statistics for simulated differences: Education

| Domain | Using Different Thresholds | | | Using Same Thresholds | | |
|---------------|-----------------------------|-----------|---------------------------|-----------------------------|-----------|---------------------------|
| | Education Difference (1) | SE (2) | <i>t</i> Statistic (3) | Education Difference (4) | SE (5) | <i>t</i> Statistic (6) |
| IFLS | | | | | | |
| Mobility | 0.1091 | 0.0463 | 2.3540* | 0.1892 | 0.0494 | 3.8270** |
| Pain | 0.0942 | 0.0414 | 2.2770* | 0.2165 | 0.0538 | 4.0220** |
| Cognition | 0.0134 | 0.0399 | 0.3365 | 0.1124 | 0.0457 | 2.4580* |
| Sleep | 0.0921 | 0.0381 | 2.4160* | 0.1017 | 0.0428 | 2.3750* |
| Affect | 0.0181 | 0.0438 | 0.4138 | 0.0854 | 0.0525 | 1.6272 |
| Breathing | -0.0016 | 0.0385 | -0.0404 | 0.0463 | 0.0416 | 1.1124 |
| HRS | | | | | | |
| Mobility | 0.1629 | 0.0231 | 7.0420** | 0.2267 | 0.0374 | 6.0600** |
| Pain | 0.0664 | 0.0150 | 4.4270** | 0.1613 | 0.0308 | 5.2430** |
| Cognition | 0.1148 | 0.0213 | 5.3840** | 0.2986 | 0.0510 | 5.8580** |
| Sleep | 0.0460 | 0.0177 | 2.5990** | 0.2289 | 0.0357 | 6.4100** |
| Affect | 0.0999 | 0.0224 | 4.4530** | 0.1742 | 0.0476 | 3.6620** |
| Breathing | 0.1276 | 0.0239 | 5.3280** | 0.2291 | 0.0383 | 5.9840** |
| ELSA | | | | | | |
| Mobility | 0.1275 | 0.0221 | 5.7820** | 0.0986 | 0.0336 | 2.9340** |
| Pain | 0.0627 | 0.0187 | 3.3440** | 0.1019 | 0.0320 | 3.1870** |
| Cognition | 0.1074 | 0.0226 | 4.7580** | 0.2164 | 0.0509 | 4.2500** |
| Sleep | -0.0031 | 0.0209 | -0.1489 | 0.1582 | 0.0303 | 5.2250** |
| Affect | 0.0221 | 0.0220 | 1.0018 | 0.1860 | 0.0406 | 4.5870** |
| Breathing | 0.1000 | 0.0198 | 5.0620** | 0.1221 | 0.0405 | 3.0170** |
| CHARLS | | | | | | |
| Mobility | 0.1093 | 0.0661 | 1.6550 [†] | 0.0762 | 0.0689 | 1.1066 |
| Pain | 0.1491 | 0.0584 | 2.5510* | 0.1597 | 0.0623 | 2.5630* |
| Cognition | 0.1751 | 0.0915 | 1.9140 [†] | 0.1226 | 0.0834 | 1.4709 |
| Sleep | 0.1751 | 0.0758 | 2.3090* | 0.1918 | 0.0843 | 2.2760* |
| Affect | 0.1597 | 0.0579 | 2.7580** | 0.2155 | 0.0671 | 3.2110** |
| Breathing | 0.1233 | 0.0637 | 1.9340 [†] | 0.1746 | 0.0706 | 2.4710* |

Notes: “Education Difference” is the difference between the proportion of high-education individuals in the healthiest category and the proportion of lower-education individuals in the healthiest category. Simulated proportions are calculated using coefficients from a HOPIT specification with the following explanatory variables: two age dummy variables, 1(Male), and all age-gender interactions. Standard errors are calculated analytically using the formulas in Online Resource 1, section A3.

[†] $p < .10$; * $p < .05$; ** $p < .01$

domains. The CHARLS shows significant differences in all six domains before adjustment, but for mobility and cognition, the differences narrow and become insignificant after adjustment for DIF. Despite this, across all data sets (including

CHARLS), education differences are generally quite large and persistent. For pain and sleep, all data sets show significant differences across education levels after reporting heterogeneity is accounted for.

Conclusion

Anchoring vignettes are a vital tool that can be used to account for reporting bias in subjective scale measures. Ignoring DIF underestimates the differences in health across education levels in Indonesia, the United States, England, and China because educated individuals have a higher bar for considering someone healthy. If individuals' evaluations of health are based partially on comparisons with peers, perhaps it is because more-educated people are surrounded by more-educated and healthier peers and therefore have a tendency to consider themselves (and hypothetical individuals) relatively less healthy.¹⁶ If schooling directly affects one's knowledge about health and disease, then more-educated individuals may be simply more aware of potential threats to health or may be more knowledgeable about the consequences of certain symptoms.

The result that education disparities in health can be underestimated by reporting heterogeneity is consistent with previous literature that used the same data sets (Bago d'Uva et al. 2011; Dowd and Todd 2011) as well as with studies of elderly health in different countries (Bago d'Uva et al. 2008a). However, the universality of this finding should not be overstated: it does not appear to be true in younger populations (Bago d'Uva et al. 2008b) or for variables other than domain-specific self-reported health. Using general self-reported health instead of the domain-specific health that I use here, Grol-Prokopczyk et al. (2011) found that the education gap actually diminishes after adjustment. For work disability, the results are mixed (Angelini et al. 2011; Kapteyn et al. 2007).

This article's conclusions about gender differences are slightly less uniform than its education results. Although significant differences between males and females remain in three of the six domains for the IFLS and HRS even after adjustment for thresholds, accounting for thresholds in England and China completely eliminates significant differences between males and females in all but one domain (sleep in the ELSA). Overall, however, reporting differences across gender are clearly important, given that gender gaps are narrowed after adjustment in the majority of domains for all data sets except the HRS.

Previous vignette studies have found that both male and female respondents rate a given vignette condition as more severe when the hypothetical vignette individual is female (Kapteyn et al. 2007). Together with the results of this article, these findings suggest that the gender of the *object* of evaluation—regardless of whether a hypothetical individual or one's own self—plays a role in shaping the elicited evaluations of health. Separating the effect of the respondent's gender from the effect of the object's gender is outside the scope of this work,¹⁷ but existing research suggests that the gender of the respondent matters much more than the gender of the vignette individual (Grol-

¹⁶ See Dowd and Todd (2011) for a more detailed discussion.

¹⁷ Although some studies have been able to include vignette gender as a variable in the vignette latent variable equation, I do not have this information for all four data sets.

Prokopczyk 2014). What I can conclude from this analysis is that irrespective of the reasons for their use of different thresholds, males and females in the ELSA, CHARLS, and to a lesser extent, IFLS, would report much more similar levels of health if they used the same thresholds.

The narrowing of the gender gap after adjusting for reporting heterogeneity provides empirical support for the hypothesis that differential reporting behavior may play a partial role in the gender puzzle discussed in the Introduction. Males are more stoic in their evaluations of health, which leads to overstated differences between the self-reports of each gender that are not aligned with differences in objective measures.¹⁸ Although this finding holds true across the majority of domain-data set combinations in this article, it is a partial explanation at best. Gender gaps fail to narrow after adjustment, not only in several HRS health domains in this article, but also in other vignette studies that used different measures of health (Angelini et al. 2011; Grol-Prokopczyk et al. 2011; Kapteyn et al. 2007).

Education disparities in self-reported health appear to reflect true (and, if anything, understated) differences in health. Although overstated in some contexts, gender inequalities also exist (particularly in the HRS). Both of these findings emphasize the importance of pinning down the causal mechanisms linking health, gender, education, and related life outcomes. They also highlight how crucial it is to consider reporting heterogeneity when comparing self-reported measures. Fortunately, the increasing availability of anchoring vignettes in surveys across the globe is making it easier to avoid relying on naive, distorted comparisons of self-reported health.

Acknowledgments I am grateful to John Strauss, Arie Kapteyn, Jeff Nugent, Geert Ridder, and Hashem Pesaran for invaluable feedback and guidance throughout various stages of this project. Many thanks to USC seminar participants and attendees of the 2015 annual meeting of the Population Association of America for helpful comments. I acknowledge funding from the USC Provost's Ph.D. Fellowship, the USC Dornsife INET graduate student fellowship, and the Oakley Endowed Fellowship.

Appendix: Description of Data Sets

Indonesian Family Life Survey (IFLS)

I use the 2007 wave of the IFLS, an ongoing longitudinal household survey of individuals in 13 of the 27 Indonesian provinces, representative of 80 % of the Indonesian population. This article uses information from the individual-level demographic and health status modules. IFLS 4 also randomly chose 2,500 households to participate in the health vignette module. In selected households, all adults over 40 were asked the six domain-specific health questions. Crucially, the IFLS included three anchoring vignettes per health domain in addition to the above self-reports. Although all vignette households were asked all of the self-reported health questions, time constraints meant that each vignette household was assigned to respond to anchoring

¹⁸ These heterogeneous reporting styles are likely related to the tendency of women to incorporate a wider range of nonphysical factors into self-reports (Benyamini et al. 2000) or societal expectations that consider males the tougher gender (Courtenay 2000).

vignettes for only two randomly chosen domains of the six, leaving between 1,100–1,300 individuals per domain. During the interview, the interviewers read aloud a vignette like the one described in the Anchoring Vignettes section (see Online Resource 1 for a list all of the vignettes). The interviewers then repeated the domain-relevant question from the list of self-report questions in the Data section (of course, replacing the word “you” with the name of the hypothetical vignette person). The gender of the hypothetical individuals, depicted through their names, was randomized at the household level. Answers to the health status questions and anchoring vignettes form the outcome variables of interest for this analysis.

Purposely focusing on a set of simple explanatory variables in order to facilitate comparisons with the three other data sets, I use gender, age, and education levels. Specifically, I create one dummy variable for males, one for high school graduates, and one for those who completed primary but not high school.

Health and Retirement Study (HRS)

Since 1992, the HRS has interviewed a representative sample of Americans older than 50, reinterviewing the original sample and adding new cohorts every two years. In 2007, an “off-year” between two main interview years, the Disability Vignette Study (DVS) was sent out as a mail survey to a subsample, of which 81.7 % (more than 4,000) responded. This survey included the exact same anchoring vignettes for the same six domains found in the IFLS vignette modules, except with American instead of Indonesian names. Unlike the IFLS, two versions of the questionnaires, which ordered the questions differently and used different genders for the hypothetical individuals, were used.

I combine data from this off-year study with data from the most recent main survey prior to it, which took place in 2006. From the 2006 interviews, I obtain the basic explanatory variables: age, gender, and educational attainment. Because the vast majority of HRS respondents are high school graduates, I use college graduation as my “high-education” group and high school graduates (who have not completed college) as my “medium-education” group.

English Longitudinal Study of Aging (ELSA)

Similar to the HRS, the ELSA is a longitudinal panel of individuals aged over 50 living in England (Marmot et al. 2014). Since 2002, the representative sample, which was initially drawn from the Health Survey for England, has been reinterviewed every two years. The ELSA sample was also refreshed at Waves 3, 4, and 6. I use data from the third wave, collected during 2006 and 2007, which included self-completion vignette questionnaires that were handed out to a randomly selected one-third of the sample (and completed by almost 2,500 individuals). Individuals were asked to rate their own health in the six domains and then to respond to the same vignettes found in the IFLS and HRS. Unlike the other data sets, which randomized the genders of vignette individuals in varying ways, the ELSA had only one version of the questionnaire, which had the same names (and thus genders) assigned to the same questions for all respondents. The vignette genders alternated throughout the questionnaire, with one-half of the vignette individuals assigned female names and the other one-half male names.

Along with respondent age and gender, I use degree qualifications as my education variable because precise years of schooling are not included in this survey. The “high-education” category includes those who have received their A-levels or higher, while the “medium-education” category includes all qualifications lower than A-levels. This leaves those with no qualifications as the low-education group.

China Health and Retirement Longitudinal Study (CHARLS)¹⁹

Finally, I also use data from the first wave of the CHARLS, conducted in 2011 (Zhao et al. 2013). Very similar to the other two longitudinal aging studies described earlier (the HRS and ELSA), the CHARLS has interviewed a representative sample of over 17,000 Chinese residents aged 45 and older and plans to follow up with the respondents every two years. The CHARLS is one of very few Chinese surveys that include domain-specific self-reports and vignette questions, which are asked as part of the full in-person interview for a random subsample of households. Like in the IFLS, each vignette household is randomly assigned to two of the six domains, resulting in around 1,100 to 1,300 respondents per domain. The genders of the hypothetical individuals are also randomized at the household level.

Because high school graduation rates for this sample are so low (less than 10 %), I use junior high school completion as my “high-education” cutoff and primary school completion as the boundary between the medium- and low-education groups.

Self-report Distributions

Figures 2 and 3 explore within-country differences across gender and education. Figure 2 depicts the distribution of self-report responses by gender for each data set separately. On each domain graph, I report the p value corresponding to the Pearson chi-squared statistic for the test of the null hypothesis that the distribution of the responses is the same for males and females. In the IFLS and CHARLS, for pain, cognition, affect, and sleep, males and females have significantly different self-report distributions, with males disproportionately falling in the healthiest category.

In the HRS, male and female distributions in the cognition, affect, and sleep domains are significantly different. In the ELSA, the domains that exhibit significant gender differences are pain, sleep, and affect. Like in the IFLS and CHARLS, males fall into the healthiest category more often than females.

Figure 3 shows even more drastically different distributions of self-reports, this time between high-education and lower-education groups (for which I pool the medium- and low-education categories). In virtually all domains in all four samples (with the exception of cognition and affect in the IFLS), the distributions are significantly different, with the higher education group disproportionately represented in the healthiest categories.

¹⁹ CHARLS is conducted by the National School of Development (China Center for Economic Research) at Beijing University. See <http://charls.ccer.edu.cn/charls/> for more detail.

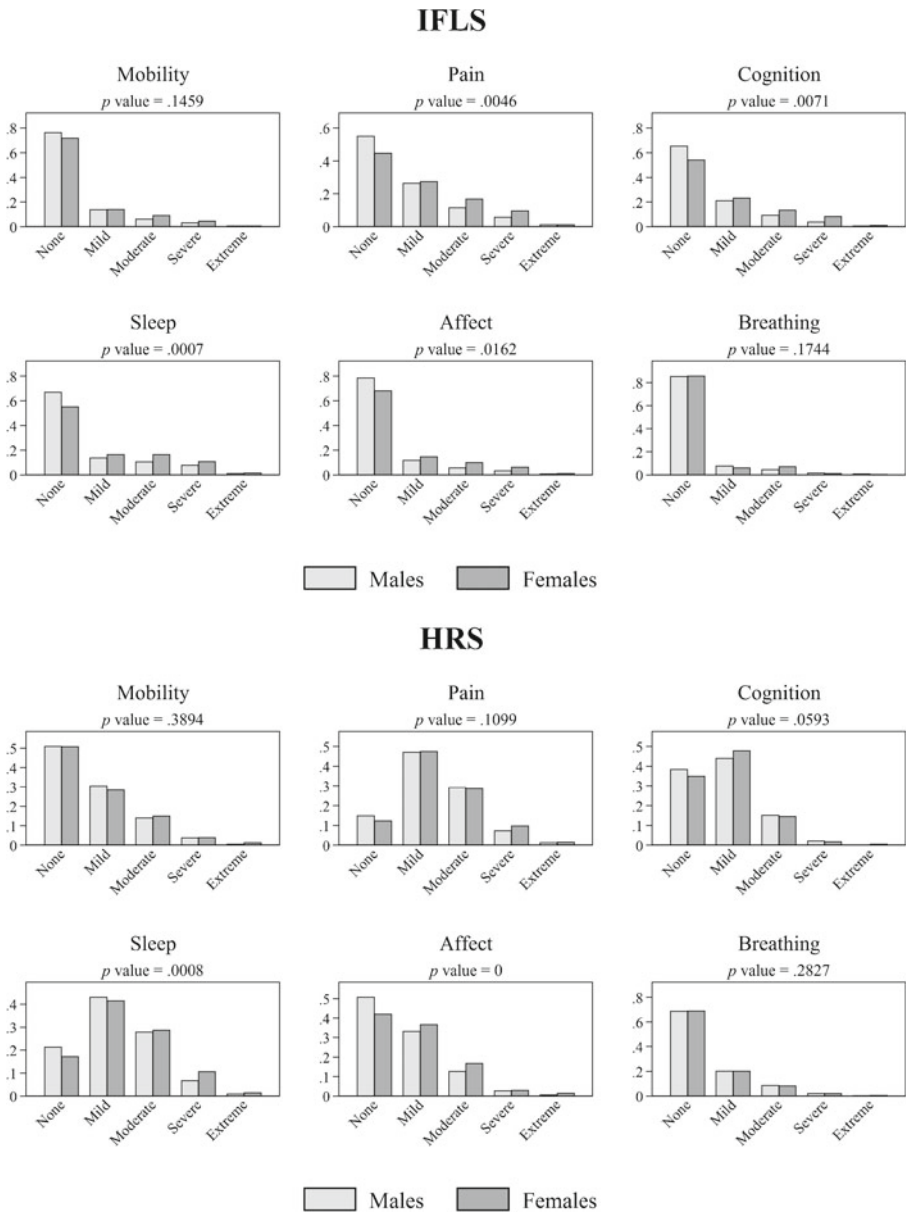
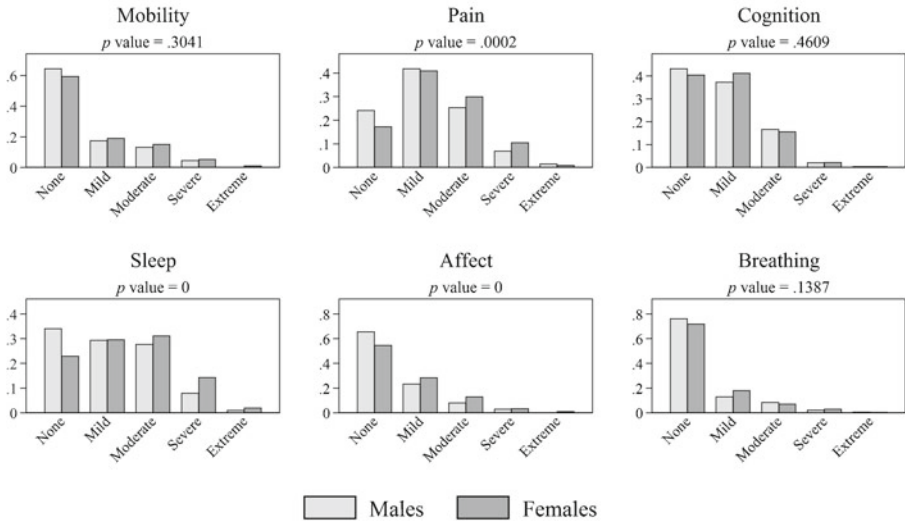


Fig. 2 Distribution of self-reports by gender

ELSA



CHARLS

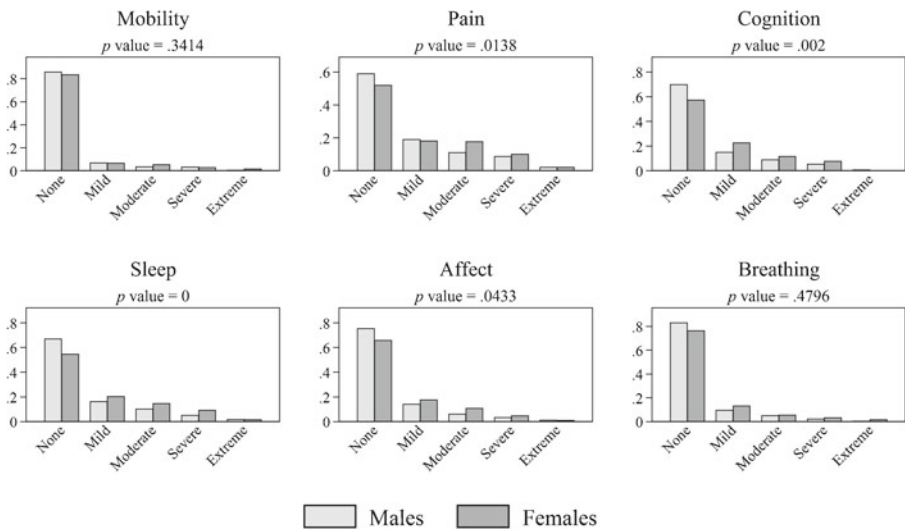


Fig. 2 (continued)

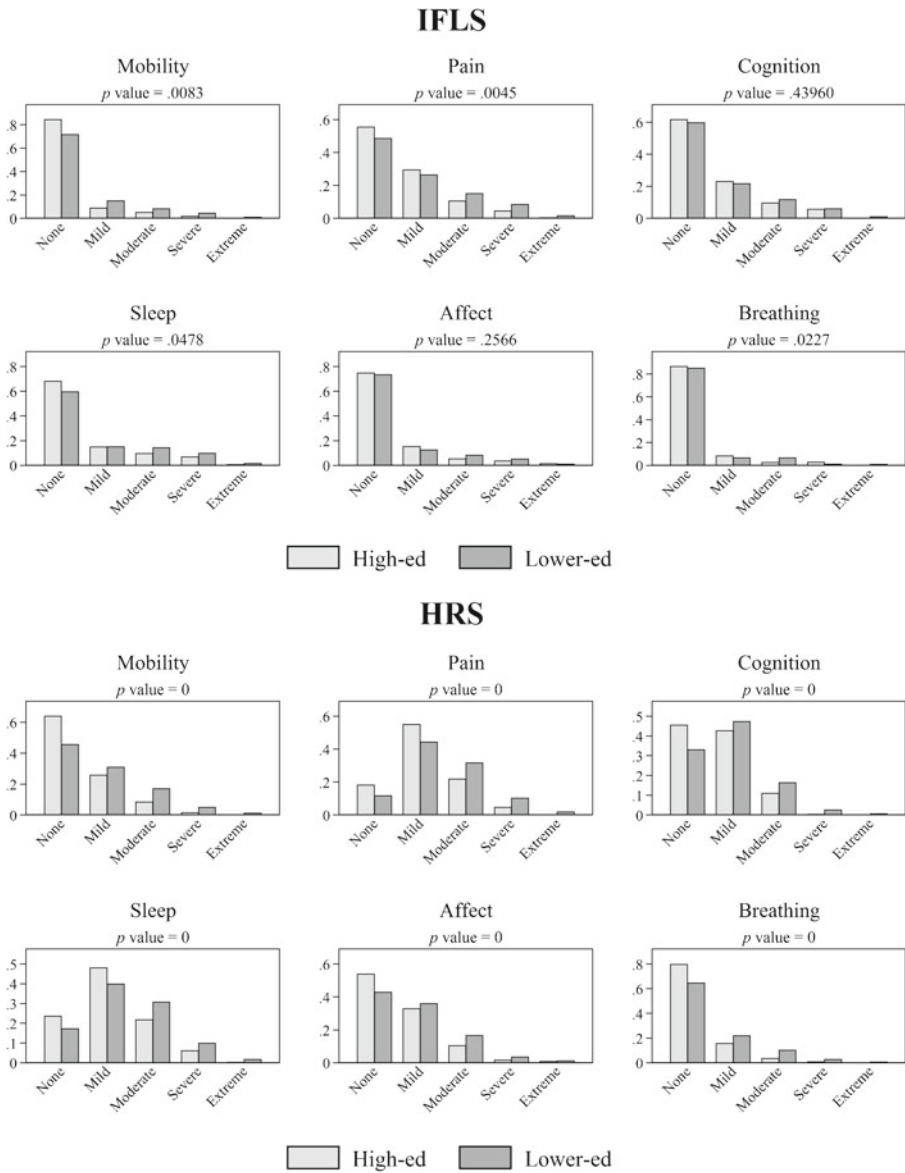
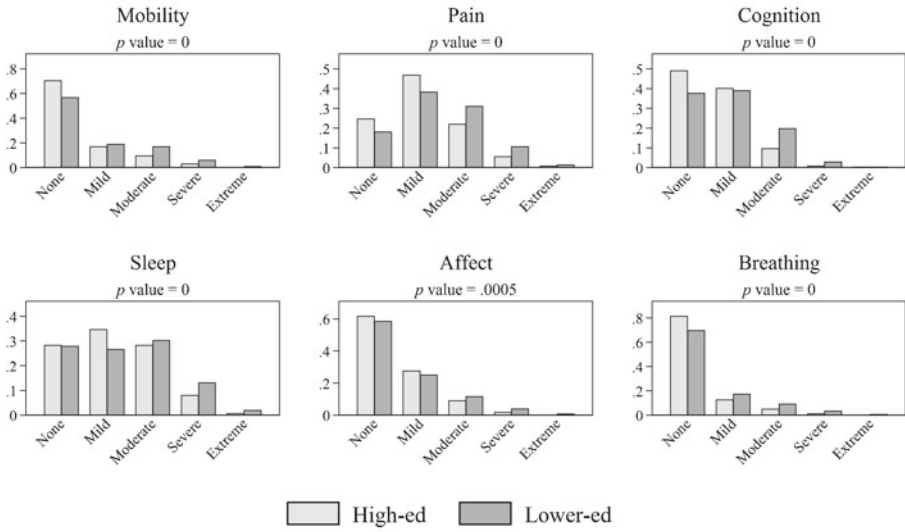


Fig. 3 Distribution of self-reports by education

ELSA



CHARLS

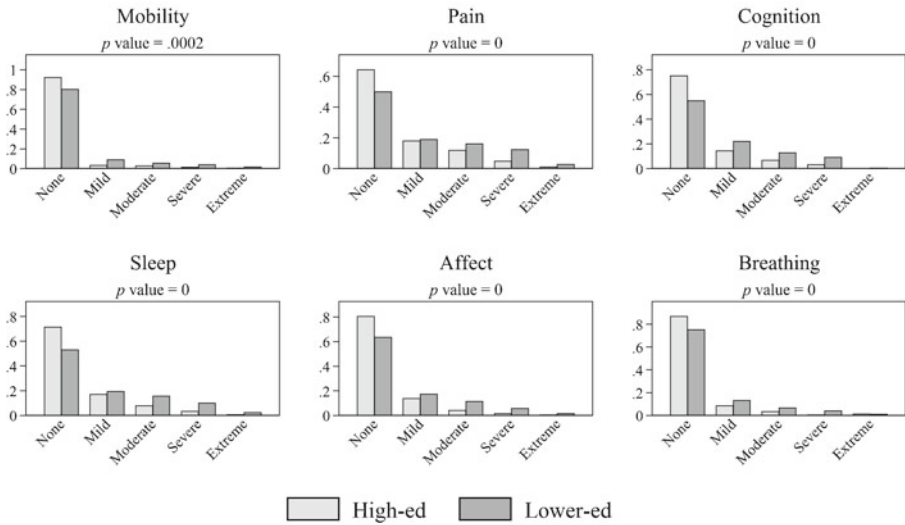


Fig. 3 (continued)

References

- Angelini, V., Cavapozzi, D., & Paccagnella, O. (2011). Dynamics of reporting work disability in Europe. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*, 621–638.
- Bago d'Uva, T., Lindeboom, M., O'Donnell, O., & Van Doorslaer, E. (2011). Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources*, *46*, 875–906.
- Bago d'Uva, T., O'Donnell, O., & van Doorslaer, E. (2008a). Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology*, *37*, 1375–1383.
- Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008b). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, *17*, 351–375.
- Benyamini, Y., Leventhal, E. A., & Leventhal, H. (2000). Gender differences in processing information for making self-assessments of health. *Psychosomatic Medicine*, *62*, 354–364.
- Case, A., & Paxson, C. (2005). Sex differences in morbidity and mortality. *Demography*, *42*, 189–214.
- Courtenay, W. H. (2000). Constructions of masculinity and their influence on men's well-being: A theory of gender and health. *Social Science & Medicine*, *50*, 1385–1401.
- Crimmins, E. M., Kim, J. K., & Solé-Auró, A. (2010). Gender differences in health: Results from SHARE, ELSA and HRS. *European Journal of Public Health*. doi:10.1093/eurpub/ckq022
- Cutler, D. M., & Lleras-Muney, A. (2006). *Education and health: Evaluating theories and evidence* (NBER Working Paper No. 12352). Cambridge, MA: National Bureau of Economic Research.
- DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of General Internal Medicine*, *21*, 267–275.
- Dow, W. H., Gertler, P., Schoeni, R. F., Strauss, J., & Thomas, D. (1997). *Health care prices, health and labor outcomes: Experimental evidence* (Labor and Population Program Working Paper Series 97-01 No. DRU-1588-NIA). Santa Monica, CA: RAND.
- Dowd, J. B., & Todd, M. (2011). Does self-reported health bias the measurement of health inequalities in us adults? Evidence using anchoring vignettes from the health and retirement study. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *66*, 478–489.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., . . . Baicker, K. (2012). The Oregon Health Insurance Experiment: Evidence from the first year. *Quarterly Journal of Economics*, *127*, 1057–1106.
- Gertler, P., & Gruber, J. (2002). Insuring consumption against illness. *American Economic Review*, *92*, 51–70.
- Grol-Prokopczyk, H. (2014). Age and sex effects in anchoring vignette studies: Methodological and empirical contributions. *Survey Research Methods*, *8*(1), 1–17.
- Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior*, *52*, 246–261.
- Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., & Ispány, M. (2015). Promises and pitfalls of anchoring vignettes in health survey research. *Demography*, *52*, 1703–1728.
- Grossman, M. (2006). Education and nonmarket outcomes. *Handbook of the Economics of Education*, *1*, 577–633.
- Health and Retirement Study (HRS). (2014). Public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (Grant No. NIA U01AG009740). Ann Arbor: University of Michigan.
- Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, *38*, 21–37.
- Kapteyn, A., Smith, J. P., & van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, *1*, 461–473.
- Kapteyn, A., Smith, J. P., & van Soest, A. (2010). Life satisfaction. In E. Diener, D. Kahneman, & J. Helliwell (Eds.), *International differences in well-being* (pp. 70–104). New York, NY: Oxford University Press.
- King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*, 191–207.
- Lei, X., Smith, J. P., Sun, X., & Zhao, Y. (2013). *Gender differences in cognition in China and reasons for change over time: Evidence from CHARLS* (IZA Discussion Paper No. 7536). Bonn, Germany: Institute for the Study of Labor.
- Maccini, S., & Yang, D. (2009). Under the weather: Health, schooling, and economic consequences of early-life rainfall. *American Economic Review*, *99*, 1006–1026.

- Macintyre, S., Ford, G., & Hunt, K. (1999). Do women over-report morbidity? Men's and women's responses to structured prompting on a standard question on long standing illness. *Social Science & Medicine*, *48*, 89–98.
- Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, A., & Marquis, M. S. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review*, *77*, 251–277.
- Marmot, M., Oldfield, Z., Clemens, S., Blake, M., Phelps, A., Nazroo, J., . . . Banks, J. (2014). *English Longitudinal Study of Ageing: Waves 0-6, 1998–2013* [UK Data Archive]. Retrieved from <http://dx.doi.org/10.5255/UKDA-SN-5050-8>
- Molina, T. (2014). *Adjusting for heterogeneous response thresholds in cross-country comparisons of mid-aged and elderly self-reported health*. Unpublished manuscript, Department of Economics, University of Southern California, Los Angeles, CA.
- Mu, R. (2014). Regional disparities in self-reported health: Evidence from Chinese older adults. *Health Economics*, *23*, 529–549.
- Nathanson, C. A. (1975). Illness and the feminine role: A theoretical review. *Social Science & Medicine (1967)*, *9*(2), 57–62.
- Strauss, J., Gertler, P. J., Rahman, O., & Fox, K. (1993). Gender and life-cycle differentials in the patterns and determinants of adult health. *Journal of Human Resources*, *28*, 791–837.
- Strauss, J., Witoelar, F., Sikoki, B., & Wattie, A. M. (2009). *The Fourth Wave of the Indonesian Family Life Survey (IFLS4): Overview and field report* (Labor and Population Working Paper No. WR-675/1-NIA/NICHD). Santa Monica, CA: RAND.
- van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*, 575–595.
- Verbrugge, L. M. (1989). The twain meet: Empirical explanations of sex differences in health and mortality. *Journal of Health and Social Behavior*, *30*, 282–304.
- Vogl, T. S. (2014). Education and health in developing economies. In A. J. Culyer (Ed.), *Encyclopedia of health economics* (1st ed., pp. 246–249). Boston, MA: Newnes.
- Zhao, Y., Hu, Y., Smith, J. P., Strauss, J., & Yang, G. (2012). Cohort profile: The China Health and Retirement Longitudinal Study (CHARLS). *International Journal of Epidemiology*, *43*, 61–68.
- Zhao, Y., Strauss, J., Yang, G., Giles, J., Hu, P., Hu, Y., . . . Wang, Y. (2013). *China Health and Retirement Longitudinal Study – 2011–2012 National baseline users guide*. Beijing, China: National School of Development, Peking University.