Reporting Heterogeneity and Health Disparities across Gender and Education Levels: Evidence from Four Countries

Teresa Molina*

March 2015

Abstract

I use anchoring vignettes from Indonesia, the U.S., England, and China to study the extent to which differences in self-reported health across genders 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 after equalizing thresholds across the two groups.

Keywords: self-reported health, anchoring vignettes, gender health disparities, SES health disparities.

JEL Classification Codes: J14, J16, I10.

*University of Southern California and USC-INET, tsmolina@usc.edu; I am grateful to John Strauss, Arie Kapteyn, Geert Ridder, and Hashem Pesaran for their invaluable guidance throughout various stages of this project. Many thanks to USC seminar participants for helpful comments. I acknowledge funding from the USC Dornsife INET graduate student fellowship.

1 Introduction

One quite persistent puzzle has emerged from studies of self-reported health. From early adolescence to late middle age, women have significantly worse self-reported health than men, despite the fact that they have lower mortality rates (Case and Paxson, 2005; Macintyre et al., 1999; Nathanson, 1975; Strauss et al., 1993; Verbrugge, 1989). In this paper, I use anchoring vignettes to quantify the extent to which differences in reporting behavior may drive these differences across gender and additionally, differences across education levels. I draw on four different datasets from four different countries: the Indonesian Family Life Survey (IFLS), the United States Health and Retirement Study (HRS), the English Longitudinal Study of Aging (ELSA), and the China Health and Retirement Longitudinal Study (CHARLS). All four of these surveys ask respondents to rate their own health difficulties from 1 to 5 (where 1 represents the least and 5 the most severe problems) in six domains: mobility, pain, cognition, sleep, affect, and breathing. In addition, for each domain, all four 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 groups (both 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 the gender gaps are reduced after accounting for the use of different thresholds, though less drastically in Indonesia and the United States, where half of the domains still reveal significant gender differences after adjustment. In England, and China, however, adjusting for thresholds completely eliminates the gender gap in the majority of domains for which significant gender differences exist in the raw data. This elimination (or reduction) of significant gender differences after adjusting for response thresholds offers a potential explanation for the gender puzzle described above: women may report worse health but have better objective indicators than men because the two genders use different response thresholds when evaluating a person's health. This is not the only possible explanation for the gender paradox

or the first time this particular hypothesis has been proposed.[1] However, until now, empirical evidence for differential reporting behavior across genders has been mixed at best (Macintyre et al., 1999; Verbrugge, 1989). This paper offers direct evidence for the use of different response thresholds across men and women, which 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. In fact, across all four datasets, I find persistent education differences that do not diminish (and in most cases widen) after adjusting for the use of different thresholds. While this may be unsurprising given that numerous studies have documented the positive relationship between education and both subjective and objective measures of health,[2] the universality of this pattern is quite remarkable given the different cultural contexts, income levels, and distribution of covariates across the four countries. This adds further support to the large literature on the education health gradient, 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 paper contributes to the literature on anchoring vignettes by expanding their use to within-country gender and education differences in four different countries. Most of the early anchoring vignettes papers focus 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 focused on within-country differences, particularly in self-reported health (Bago d'Uva et al., 2008; Dowd and Todd, 2011; Mu, 2014). None of these papers, however, focus on differences across gender

[1] Mortality selection is one potential reason for the gender paradox, but Strauss et al. (1993) find that adjusting for it reduces but does not eliminate the gender gap in self-reported health. Case and Paxson (2005) find 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) for a review and Vogl (2012) for a review specifically for developing countries.

or education levels. Any discussion of these differences is usually limited to an examination of coefficients in a pooled HOPIT model, which only allows gender and education to have 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. This allows for gender and education to change how other covariates affect health and reporting behavior. Kapteyn et al. (2007), Kapteyn et al. (2010), and Mu (2014) all run the HOPIT model separately for different countries or different regions, but this paper is the first to conduct this exercise for gender and education levels. This paper 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. Finally, my use of data from multiple countries allows me to make conclusions that are not specific to just one setting.

The next section outlines how anchoring vignettes help solve the problems that arise due to individuals' use of different response thresholds. Section 3 outlines the econometric model, and Section 4 describes the four datasets used in this analysis. I outline the estimation methods in Section 5, discuss my results in Section 6, and conclude with Section 7.

2 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) since objective measures of health are often infeasible to measure for large populations or too narrow to capture the multidimensional nature of health. The particular type of measure studied in this paper takes the following form: a response to a question like "overall, in the last 30 days, how much pain or bodily aches did you have?", chosen from 5 options: none, mild, mod-

erate, severe, and extreme. These self-reports are simple and may be better suited to capture an individual's health as a whole, compared to objective measures that are more specific (like blood pressure or BMI) or more extreme (like mortality). Moreover, self-reported health has been repeatedly shown to have a significant relationship with mortality, robust to the inclusion of a host of demographic and socioeconomic controls.[3]

Despite this, 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, highlight that self-reported measures often suffer from reporting bias that is non-random. The authors argue that this bias may be correlated with variables like income, or more importantly, healthcare utilization – which is especially problematic if healthcare utilization is a regressor of interest. 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 capture an individual's interpretation of the response choices: 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 when making comparisons across groups or individuals. The underlying problem is that we cannot ascertain whether the differences we see are being driven by actual differences in health status or simply the use of different response scales, what King et al. (2004) refer to as "differential item functioning" (DIF), a term originally from the education testing literature.[4] Equivalently, we are also unsure if groups that appear similar actually have differences that are masked by different response scales. In short, with systematically different response scales, we 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

[3] Idler and Benyamini (1997) review 27 studies conducted in eight different countries. With remarkable consistency, these studies show that the coefficient on 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) finds that individuals who report being in "poor" health have almost double the mortality risk of those who reported being in "excellent" health. This calculation included studies which controlled for various covariates like age, socioeconomic status, and others.

[4] A test question with "differential item functioning" is one that two people of the same ability but from different groups (races or genders, for example) have different probabilities of answering correctly.

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? [5]

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 allow us to evaluate 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, adapted from King et al. (2004), shows two different respondents: A and B. In Panel A, Self1 represents A's numerical response to a subjective question like "how is your health in general?" Self2, 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 give very different evaluations of the three vignettes, indicating the use of potentially very 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 aligning B's vignette evaluations to A's and comparing Self1 and Self2 on the new scale. Comparing Panel A and Panel C show us that B is actually in better health than A but has a higher bar for defining what is "healthy."

Anchoring vignettes allow us to infer something about respondents' internal response scales that are otherwise completely unobservable to the researcher. When comparing two groups of individuals, we can use the scale in one group as a benchmark in order to make valid comparisons. The validity of these comparisons hinges on two important assumptions. First, we assume

[5] This vignette is from the cognition domain and used in all four datasets this paper. See Appendix section C for complete list of vignettes.

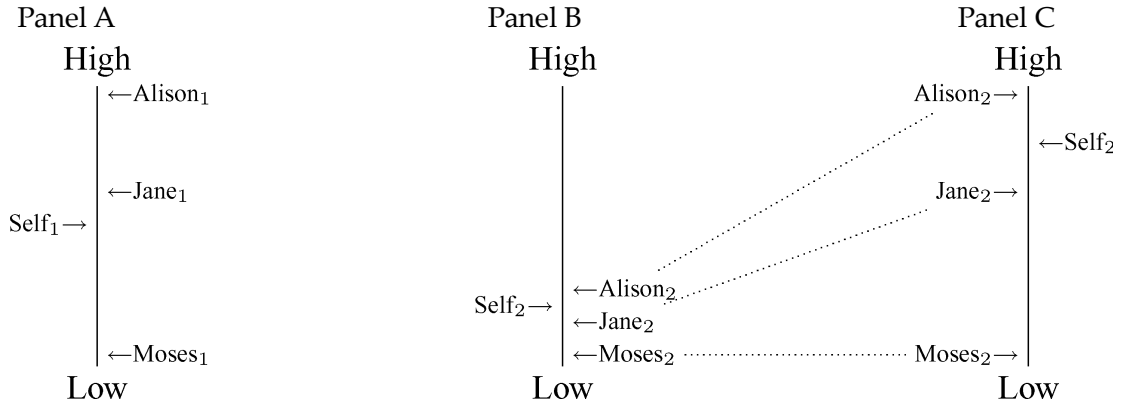


Figure 1: Comparing Subjective Scales (From King et al. (2004))

response consistency, which means that respondents use the same response scales when evaluating themselves and evaluating others. The second assumption is *vignette equivalence*, which means that the way respondents interpret the scenarios and questions are independent of their individual characteristics. In other words, respondents only differ 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) suggest 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) test for response consistency by using objective measures: vision tests to validate subjective scale questions about vision impairment (King et al., 2004) and actual counts of alcoholic drinks to validate subjective questions about the severity of drinking problems (Van Soest et al., 2011). Both find strong evidence to support response consistency. Unfortunately, tests like these are only possible when relevant objective measures, which map well to the unobserved latent variable, exist. While 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 paper make this a reasonable assumption for the self-reported health setting. The individuals described in the vignettes in this paper suffer from common ailments un-

doubtedly somewhat familiar to respondents in all countries. This familiarity, combined with the fact that health is an issue 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.

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. Since vignettes are brief, vignette equivalence may also be violated if respondents fill in any gaps they need for a complete picture of the hypothetical person by making assumptions about them. These assumptions are likely to vary by person and are problematic if correlated with individual characteristics. Fortunately, all of the vignettes used in this paper are straightforward and deal with tangible, familiar concepts. However, because of their brevity, they may be slightly open to interpretation.

In a comparison of models that relaxed various combinations of these assumptions, Van Soest et al. (2011) found that the model assuming only response consistency performed the best according to the Akaike Information Criterion. They tested DIF and response consistency, all the while maintaining vignette equivalence. In this paper, I test vignette equivalence using methods proposed by Bago d'Uva et al. (2011).

3 Econometric Model

In order 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. 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.

1. $Y_i^* = X_i\beta + \epsilon_i$; ϵ_i is $N(0, \sigma_\epsilon)$, ϵ_i independent of X_i and the other error terms in the model

2. $Y_i = j$ if $\tau_i^{j-1} < Y_i^* \leq \tau_i^j$, j=1,....5

3. $\tau_i^0 = -\infty$, $\tau_i^5 = \infty$, $\tau_i^1 = \gamma^1 X_i + u_i$, $\tau_i^j = \tau_i^{j-1} + e^{\gamma^j X_i}$, j = 2, 3, 4

   u_i is $N(0, \sigma_u^2)$ and is independent of X_i and the other error terms in the model.

What sets this 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 X characteristics to have different response scale thresholds. The individual-specific τ_i^j's are the essence of DIF.

Given data on self-reported health and individual characteristics only, it is impossible to identify β and γ^1 separately (but γ^j for $j > 1$ is identified through the non-linearity of the exponential function). For this, we use the three vignette questions asked of 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 paper, $l = 1, 2, 3$.

4. $Y_{li}^* = \theta_l + \epsilon_{li}$; ϵ_{li} is $N(0, \sigma_v)$, ϵ_{li} independent of X_i and the other error terms in the model

5. $Y_{li} = j$ if $\tau_i^{j-1} < Y_{li}^* \leq \tau_i^j$, j=1,....5

The non-negative exponential function in threshold equation (3) ensures that $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$. Its non-linearity ends up playing a key role in identifying the γ_j coefficients for $j > 1$. The main results in the paper use the exponential function to define the gaps between different thresholds, as in equation 3, but I also test the sensitivity of these results by replacing the exponential in equation 3 with a square, as follows.

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

Comparing Table A3 to A4, the results remain remarkably consistent across alternate functional forms. This is true for all domains and all four datasets. In Appendix section C, I also explore the possibility of using a linear specification for the threshold equations.

The model's first crucial assumption, response consistency, means that the τ_i's in equation 3 are used for both the self-reports (equations 1 and 2) and the vignette responses (equations 4 and 5). Since vignette responses Y_{li}^* only depend on individual characteristics through their influence on the thresholds τ_i, it is possible to identify γ and θ vectors from equations 4 and 5. Here, θ_l is a vignette fixed effect that, together with an unobserved individual error ϵ_{li}, completely determine 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 only affect evaluations of vignettes through their effect on thresholds. This leads naturally to a test of vignette equivalence which involves including respondent characteristics X_i in the vignette equation 4. I discuss this vignette equivalence check in section 5.4.

4 Data

4.1 Indonesian Family Life Survey (IFLS)

I use the 2007 wave of the IFLS, an ongoing longitudinal household survey of individuals in 13 out of the 27 Indonesian provinces, representative of 80% of the Indonesian population (Strauss et al., 2009). This paper utilizes 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 following health status questions.

1. *Mobility:* Overall in the last 30 days, how much of a problem did you have with moving around?

2. *Pain:* Overall in the last 30 days, how much of bodily aches or pains did you have?

3. *Cognition:* Overall in the last 30 days, how much difficulty did you have remembering things?

4. *Sleep:* In the last 30 days, 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. *Affect:* Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?

6. *Breathing:* In the last 30 days, how much of a problem did you have because of shortness of breath?

Respondents were instructed to respond with a number from 1 to 5, where 1=None, 2=Mild, 3=Moderate, 4=Severe, and 5=Extreme/Cannot Do.

Crucially, the IFLS included three anchoring vignettes per health domain in addition to the above self reports. While all vignette households were asked all of the questions listed above, due to time constraints each vignette household was only assigned to respond to anchoring vignettes for two randomly chosen domains out of the six, leaving between 1100-1300 individuals per domain. During the interview, the interviewers read aloud a vignette like the one described in Section 2 (see Appendix section D for a list all of the vignettes). They then repeated the domain-relevant question from the list of questions above (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 datasets, I use gender, age, and education levels. Specifically, I create a dummy variable for males, a dummy for high school graduates, and a dummy for those who completed primary but not high school. Discretizing education levels allows me to conduct my simulation analyses by educational sub-groups.

4.2 Health and Retirement Study (HRS) [6]

Since 1992, the HRS has interviewed a representative sample of Americans older than 50, re-interviewing the original sample and adding new cohorts every 2 years. In 2007, an "off-year" in between two main interview years, the Disability Vignette Study (DVS) was sent out as a mail survey to a subsample, of which 81.7% (over 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. Since 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.

4.3 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 re-interviewed 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 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 datasets, which randomized the genders of vignette individuals in varying ways, the ELSA only had one version of the questionnaire, which had the same names (and

[6]The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

thus genders) assigned to the same questions for all respondents. The vignette genders alternated throughout the questionnaire, with half of the vignette individuals assigned female names and the other 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.

4.4 China Health and Retirement Longitudinal Study (CHARLS) [7]

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 above (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 sub-sample of households. Like in the IFLS, each vignette household is randomly assigned to 2 out of the 6 domains, resulting in around 1100 to 1300 respondents per domain. The genders of the hypothetical individuals are also randomized at the household level.

As control variables, I use age, gender, and years of schooling. 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.

Table 1: Summary Statistics

	(1)	(2)	(3)	(4)
	IFLS	HRS	ELSA	CHARLS
Age	52.00	63.76	65.80	59.57
	(9.618)	(9.046)	(10.30)	(10.13)
1(Male)	0.535	0.453	0.461	0.467
	(0.499)	(0.498)	(0.499)	(0.499)
1(High Education Group)[1]	0.218	0.281	0.358	0.334
	(0.413)	(0.449)	(0.480)	(0.472)
1(Medium Education Group)[2]	0.436	0.570	0.225	0.220
	(0.496)	(0.495)	(0.418)	(0.415)
Height	155.2	167.7	165.6	158.1
	(8.476)	(9.793)	(9.752)	(8.101)
BMI	23.12	29.12	27.96	23.30
	(4.600)	(5.789)	(4.851)	(3.552)
1(BMI < 18.5)	0.118	0.00294	0.00919	0.0670
	(0.323)	(0.0542)	(0.0955)	(0.250)
1(BMI > 30)	0.0630	0.381	0.289	0.0409
	(0.243)	(0.486)	(0.453)	(0.198)
Waist Circumference	81.35	100.1	95.91	84.34
	(10.54)	(15.95)	(13.42)	(10.82)
1(Low-Risk Waist Circumf)	0.822	0.184	0.373	0.654
< 102 cm for men, < 88 cm for women	(0.383)	(0.387)	(0.484)	(0.476)
1(High Blood Pressure)	0.437	0.315	0.379	0.332
Systolic > 140 or Diastolic > 90	(0.496)	(0.465)	(0.485)	(0.471)
Mobility Self-Report	1.430	1.742	1.644	1.311
	(0.848)	(0.910)	(0.944)	(0.814)
Pain Self-Report	1.815	2.366	2.288	1.900
	(1.027)	(0.871)	(0.932)	(1.132)
Cognition Self-Report	1.687	1.834	1.801	1.756
	(0.989)	(0.776)	(0.815)	(1.037)
Affect Self-Report	1.678	2.309	2.278	1.743
	(1.034)	(0.922)	(1.044)	(1.069)
Sleep Self-Report	1.473	1.777	1.583	1.500
	(0.896)	(0.876)	(0.836)	(0.904)
Breathing Self-Report	1.282	1.450	1.408	1.348
	(0.727)	(0.772)	(0.782)	(0.797)
Average Pairwise Correlation	0.39	0.42	0.34	0.34
Year of Vignette Survey	2007	2007	2006-2007	2011
Observations	3058	4158	2192	3647

Notes:

- All data are weighted using individual cross-sectional sampling weights (without adjustment for non-response or attrition), provided by each dataset 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.

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

2. 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

4.5 Summary Statistics

Table 1 lists summary statistics for all four datasets, including only individuals who responded to all questions 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. The first and the fourth columns report summary statistics for the whole IFLS and CHARLS samples of vignette respondents with non-missing covariates. Both of these sample sizes are much larger than the sample sizes in each individual domain, however, since individuals only responded to two domains each. The second and third columns summarize the HRS and ELSA samples, respectively, and these are roughly the same as the domain-specific sample sizes since everyone was asked to respond to all domains.

Although t-tests are not reported here, there are large and significant differences across all four countries that arise from differences in survey parameters, covariate distributions within each country, or a combination of both. For instance, the HRS and ELSA samples are older on average, which could be due in part to the higher life expectancies in these two countries but is likely driven primarily by the higher age threshold for inclusion in these datasets: 50 compared to 40 in the IFLS and 45 in the CHARLS. Rather than drop all IFLS and CHARLS respondents younger than 50, I choose to include everyone and control for age in order to retain as many observations as possible.[8]

The longer life expectancy of females relative to males is reflected in the fact that less than half 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 artefact of the survey design but potentially also generated by demographic differences across countries.

[7] 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.

[8] In the specification presented in this paper, I include three age dummies, but the results are robust to the use of quadratic age controls instead.

The education statistics must be interpreted with caution because the "high education," "medium education," and "low education" category definitions differ across the samples. These categories were defined in order to maintain sufficient mass in the highest education category that was similar across all datasets (ranging from 22% to 36%). This is roughly equivalent to using the 75th percentile as a high education cutoff. Keeping this in mind, it is clear that there are large differences in the levels of educational attainment across countries. Over 80% of the American sample are high school graduates, while this figure is less than a 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, which is why I use junior high school completion as a the cutoff for high education: around 30% of the CHARLS sample completed middle school education. 36% of the ELSA sample received their A-levels or higher, which is a similar yet slightly more stringent qualification than high school graduation in the United States.

I also report summary statistics on various physical measurements collected in all surveys: height, weight, waist circumference, and blood pressure. Although I do not include these variables in my model,[9] they are informative about the differences in objective health measures across the various populations. Americans and English are significantly taller, more likely to have high-risk waist circumference, and have higher BMI's than the other two samples. Indonesians are the most likely to have high blood pressure.

Table 1 also lists the self-report means for each health domain, and the average of all pair-wise correlations between self-reports for different domains. The correlations are positive but weak for all four datasets. For IFLS and CHARLS respondents, all self-report means fall between 1 and 2 (the lowest and "best" two possible answers). 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

[9]Doing so does not alter my main conclusions. However, I do not include them in the specification discussed in this paper because I am not interested in seeing how specific objective measures influence self-reports, but rather how other demographic characteristics influence latent health and response thresholds.

on the same level as 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 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. In all interviews and questionnaires, the self reports were asked before the anchoring vignette questions, therefore making it unlikely that the vignette questions introduced bias into the self-reports.[11] Here, I number the vignettes in order of increasing intended severity based on the IFLS sample and questionnaire.[12] In all samples, the average perceptions of severity are generally in accord with the intended relative levels. With the exception of the sleep domain for the HRS, ELSA, and CHARLS samples (which is one of the least straightforward of all vignette domains), the first vignette is on average rated healthier than the second, which in turn is rated healthier than the third.[13]

As shown in Figures A1 and A2, there are substantial within-country differences in self-reported health across gender and education. For all datasets, there are at least three domains which show significantly different distributions for men and women and at least four domains for which highly educated and less educated individuals have significantly different distributions. I investigate these differences using the HOPIT model discussed in Section 3, which I estimate using the methods described in the following section.

[10]Molina (2014) demonstrates that response thresholds play a large role in explaining the drastic cross-country differences between these four countries. Although the HRS and ELSA samples seem less healthy in initial comparisons, they are in fact significantly healthier than both the IFLS and CHARLS respondents once thresholds are equalized across countries.

[11]While this is standard practice, recent concerns about response consistency have led to the suggestion that switching the order may prime respondents to use the same response scales in both the self-reports and the vignette responses. (Bago d'Uva et al., 2011)

[12]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 confirms that the relative severity perceived by respondents is consistent with the ordering of vignettes in the interview.

[13]The differences in the average rating of the second and third sleep vignettes are very small for these three samples, but the data confirms that indeed, the second vignette is considered more severe than the third. It should be noted that my arbitrarily chosen ordering is irrelevant to the estimation of the model, as I will actually estimate parameters that capture the relative severity. In the ELSA and CHARLS, vignette 3 is considered significantly more severe than vignette 2, and in the HRS, this relationship is marginally significant.

Table 2: Vignette Responses

IFLS	Mobility	Pain	Cognition	Sleep	Affect	Breathing
Vignette 1	2.352	2.525	2.536	2.712	2.508	2.794
	(1.047)	(1.006)	(1.000)	(1.018)	(0.966)	(1.064)
Vignette 2	2.843	2.726	2.884	3.058	3.025	3.330
	(1.065)	(0.971)	(1.050)	(1.042)	(1.002)	(1.056)
Vignette 3	3.520	3.457	3.175	3.396	3.703	3.758
	(1.081)	(1.076)	(1.093)	(1.094)	(1.175)	(1.142)
Observations	1003	1027	1018	1122	944	996
HRS	Mobility	Pain	Cognition	Sleep	Affect	Breathing
Vignette 1	2.461	1.902	1.948	3.030	2.567	3.092
	(0.722)	(0.652)	(0.735)	(0.721)	(0.693)	(0.769)
Vignette 2	3.708	3.187	2.796	3.852	3.357	3.973
	(0.817)	(0.739)	(0.769)	(0.837)	(0.762)	(0.804)
Vignette 3	3.834	3.790	3.776	3.858	4.532	4.382
	(0.802)	(0.775)	(0.759)	(0.780)	(0.761)	(0.767)
Observations	4118	4123	4127	4126	4113	4119
ELSA	Mobility	Pain	Cognition	Sleep	Affect	Breathing
Vignette 1	2.485	1.967	2.098	2.994	2.627	3.197
	(0.770)	(0.569)	(0.680)	(0.718)	(0.709)	(0.789)
Vignette 2	3.616	3.035	2.888	3.649	3.274	3.865
	(0.878)	(0.733)	(0.745)	(0.890)	(0.777)	(0.816)
Vignette 3	3.860	3.902	3.690	3.582	4.318	4.434
	(0.796)	(0.785)	(0.834)	(0.778)	(0.840)	(0.808)
Observations	2115	2145	2121	2148	2088	2085
CHARLS	Mobility	Pain	Cognition	Sleep	Affect	Breathing
Vignette 1	1.784	2.081	1.879	2.351	2.121	2.715
	(0.929)	(0.802)	(0.889)	(0.933)	(0.873)	(1.090)
Vignette 2	2.425	2.096	2.496	3.162	2.729	3.453
	(1.077)	(0.813)	(0.936)	(1.159)	(0.952)	(1.056)
Vignette 3	3.524	3.250	2.654	3.029	3.789	3.918
	(1.004)	(0.961)	(1.067)	(0.992)	(1.105)	(1.098)
Observations	1073	1049	1141	1163	1120	1088

Notes:

All data are weighted using individual cross-sectional sampling weights (without adjustment for non-response or attrition), provided by each dataset 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.

5 Estimation Strategy

5.1 Estimating the Model

I use maximum likelihood to estimate the model described in Section 3. I normalize $\sigma_\epsilon^2 = 1$ and estimate σ_u^2, as these are not separately identified. I also normalize $\theta_3 = 0$. Due to the independence of ϵ_i and v_i, the individual likelihood contribution, conditional on u_i, is simply the product of four cumulative normal probabilities (one for the latent health equation and one for each of the three vignettes). I calculate the unconditional likelihood contribution of each individual using simulated methods, drawing 50 u_i's from a normal distribution and taking the average of the individual likelihood contribution over the u_i draws.[14] The likelihood function can be found in Appendix section A.

I estimate the model separately for each health domain, as common response scales across health domains is a strong assumption (Kapteyn et al., 2007). My specification includes the following in the vector X_i: three age dummies (for those aged 56 to 65, 66 to 75, and older than 75), a male dummy, a dummy for high education, and a dummy for medium education, which essentially breaks down the sample into three groups, where the omitted category is the low education group. I also include interactions between the age dummies and all other indicators (sex, high education, and medium education).

In order to illustrate the difference between the HOPIT and a normal ordered probit model, I first run a simple ordered probit on the IFLS data, using the same explanatory variables listed above. I then estimate the HOPIT model using the same sample and explanatory variables.[15]

[14] In practice, results were not sensitive to the number of draws used. I ran the analysis using 10, 20, 40, 50, 80, and 100 draws, and obtained very similar results in all attempts.

[15] For detailed econometric 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) use the exact same data as I use here, while Mu (2014) uses the pilot wave of the CHARLS. I use a slightly different specification from these papers.

5.2 Simulating Distributions

The main part of my empirical analysis utilizes all four datasets and focuses on the significant within-country differences across genders and education levels that were highlighted in Section 4. While the pooled specification described in Section 5.1 includes dummies for gender and education, this only allows these characteristics to affect the intercepts in equation 1 (the self-report latent variable equation) and equation 3 (the threshold determination equations). In order to allow for the slope coefficients and error variances to differ across groups, I run the analysis separately for men and women, and then separately for high education individuals and all others.

Using my estimates from the separately estimated models, I simulate the distribution of self-reports for the separate groups in several ways. For instance, I simulate the distribution of self-reported mobility separately for males using their own thresholds, females using their own thresholds, and then males using female thresholds. By comparing the distribution of male self reports predicted from using their own thresholds to the distribution of male self reports predicted using female thresholds, I will be able to determine whether using different thresholds drastically changes the predicted distributions. By comparing the simulated distribution of male self reports using female thresholds to female self reports predicted using their own thresholds, I will be able to determine whether, after adjustment, males and females are still different.

5.3 Standard Errors for Predicted Probabilities

In previous literature that has conducted these simulations, most analysis and interpretation has been conducted by simply comparing the distributions calculated using own-group threshold 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 calculate standard errors analytically.

As mentioned in Section 5.2, I am interested simulating distributions in two ways – without and with adjustment for DIF. 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 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 essentially asks how different the two groups would be if they used the same reporting thresholds.

The standard errors I calculate (equation A11 and A12 in Appendix section B) take into account correlations in covariates across individuals in a married couple. Appendix section B goes into greater detail about the derivations of all the formulas used.

5.4 Testing Vignette Equivalence

Vignette equivalence is an important assumption underlying this model, which is not always tested in existing applications of this methodology. I test for vignette equivalence using the methods outlined by Bago d'Uva et al. (2011). This test is based on the idea that vignette equivalence rules out systematic differences in respondents' understanding or interpretation of the vignettes. In other words, covariates can be excluded from the equation for the latent variable for vignette health, $Y_{li}^* = \theta_l + \epsilon_{li}$. In order to test this necessary condition for vignette equivalence, Bago d'Uva et al. (2011) suggest including covariates in all but one of the vignette equations. This allows for systematic variation in vignette responses that are not captured by the different response thresholds. In other words, I replace the original vignette equations (equation 4) with the following:

4a. $Y_{1i}^* = \theta_1 + \epsilon_{1i}$

4b. $Y_{li}^* = \theta_l + \alpha_l X_i + \epsilon_{li}, l \neq 1$

Under the null of vignette equivalence, $\alpha_l = 0$ for $l = 2, 3$. The results for these checks, as well as

the previously described estimation methods, are discussed in the following section.

6 Results

In this section, I first discuss the results from the pooled HOPIT estimation of the IFLS. I then move on to the simulation results by gender and by education, which I discuss for all four datasets.

6.1 HOPIT Estimation of the IFLS

Table 3 reports the coefficients from the main self-report equation (β in equation 1) using both an ordered probit and the HOPIT model, for each of the six health domains. The threshold equations for the cognition domain are discussed in this section, and the threshold equations for the remainder of the domains are available upon request. Since a "1" represents the healthiest response choice, negative coefficients mean the regressors are associated with better health.

More educated people appear to be healthier in the HOPIT model, across all domains except affect and breathing. Interestingly, the coefficient on the high school graduate dummy is often smaller and sometimes even indistinguishable from zero in the ordered probit model but negative and significant in the HOPIT model, suggesting that ignoring the possibility of DIF underestimates the positive relationship between educational attainment and health. The threshold equations for the cognition domain in Appendix Table A3 shed light on this hypothesis. In the first threshold equation, the coefficients on both education dummies are negative and significant, which means that more educated people have lower τ_i's. In other words, they set a higher bar for what they deem as having "no difficulty," in both their own self-reports and for the hypothetical vignettes. Failing to account for thresholds makes it seem like high school graduates are no different from non graduates, even though there are significant differences in both the true latent variable as well as the reporting behavior across groups. Although the high school graduate coefficient is positive and significant in the next threshold equation, coefficients in higher threshold equations are harder to interpret as they represent the effects of the covariates on the relative distance between one

threshold and the next. Furthermore, higher thresholds are less important because the majority of individuals in the full sample fall in the "healthiest" category.

Across all health domains except mobility and breathing, gender is significantly related to self-reported heath at the 5% level in the ordered probit, with males seemingly healthier. Moreover, males also appear significantly healthier in the HOPIT models for pain, sleep, and affect, suggesting that response thresholds do not explain much of the gender gap in these domains. In cognition, however, it appears as though the gender gap can be explained by threshold differences since the gender dummy is no longer significant in the HOPIT specification.

Across all domains except affect and breathing, older people appear to be in worse health in both the ordered probit and HOPIT models (the omitted category in these regressions is the youngest age category, 55 and younger). However, overall, there is little evidence that age changes the effect of gender and education on health, as most interactions are insignificant. In the next section, I discuss the results from simulating distributions using separately-estimated models for males and females and then high-education and lower-education groups.

6.2 Simulations

While the tables from the previous section are instructive about the direction and statistical significance of the correlations between various characteristics and self-reported health, they do not offer a quantitative answer to a crucial question: how important is DIF in accounting for differences across sub-groups? In order to answer this question, I use the HOPIT model to simulate the distribution of self-reports in various ways. I conduct two types of subgroup analysis: males vs. females and high education vs. low/medium education, which I will refer to as the "lower education" group for the remainder of the paper.

Before the simulations, I first re-estimate the model separately for each subgroup. The models estimated for the IFLS in the previous section only included gender and education dummies (and their age-dummy interactions), allowing the intercepts to vary across these subgroups, but imposed the equality of the other slope coefficients as well as error variances across groups. I then use these

Table 3: Ordered Probit and HOPIT Estimation of Health in the IFLS

	(1) Mobility		(2) Pain		(3) Cognition		(4) Sleep		(5) Affect		(6) Breathing	
	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT
1(55 < Age <= 65)	0.160	0.164	0.406**	0.474**	0.165	0.106	0.116	0.137	0.0865	0.198	0.235	0.0304
	(0.185)	(0.198)	(0.175)	(0.187)	(0.170)	(0.180)	(0.171)	(0.179)	(0.211)	(0.223)	(0.225)	(0.242)
1(65 < Age <= 75)	0.618**	0.688**	0.679***	0.750***	0.638***	0.561**	0.544**	0.401*	0.0704	0.208	0.219	0.250
	(0.265)	(0.281)	(0.208)	(0.219)	(0.219)	(0.231)	(0.220)	(0.229)	(0.273)	(0.291)	(0.327)	(0.348)
1(Age > 75)	1.061***	1.116***	-0.0258	-0.157	0.433	0.168	0.352	0.414	-0.573	-0.838	0.715	0.392
	(0.320)	(0.340)	(0.390)	(0.438)	(0.443)	(0.482)	(0.320)	(0.333)	(0.489)	(0.516)	(0.509)	(0.542)
1(Male)	-0.110	-0.0915	-0.229***	-0.198**	-0.178**	-0.136	-0.245***	-0.308***	-0.337***	-0.319***	0.0481	0.0636
	(0.102)	(0.109)	(0.0883)	(0.0954)	(0.0892)	(0.0962)	(0.0861)	(0.0907)	(0.101)	(0.108)	(0.118)	(0.124)
1(High Education)	-0.409***	-0.544***	-0.226*	-0.330**	-0.117	-0.314**	-0.233**	-0.366***	-0.0287	-0.175	0.0963	-0.137
	(0.145)	(0.155)	(0.120)	(0.130)	(0.114)	(0.123)	(0.114)	(0.120)	(0.133)	(0.143)	(0.154)	(0.164)
1(Medium Education)	-0.0894	-0.128	-0.0847	-0.0796	-0.226**	-0.338***	-0.0261	-0.0779	0.0142	0.0299	0.0252	-0.0888
	(0.112)	(0.119)	(0.102)	(0.109)	(0.105)	(0.114)	(0.0990)	(0.104)	(0.115)	(0.122)	(0.139)	(0.145)
1(Male)	0.0122	0.00169	-0.207	-0.265	-0.239	-0.279	-0.0987	-0.105	0.0328	0.0175	-0.179	-0.0231
x 1(55 < Age <= 65)	(0.203)	(0.218)	(0.178)	(0.191)	(0.182)	(0.195)	(0.182)	(0.192)	(0.226)	(0.240)	(0.228)	(0.241)
1(High Education)	0.203	0.394	-0.240	-0.227	0.154	0.194	0.126	0.116	-0.261	-0.346	-0.242	-0.000878
x 1(55 < Age <= 65)	(0.301)	(0.322)	(0.252)	(0.273)	(0.254)	(0.275)	(0.256)	(0.271)	(0.355)	(0.379)	(0.348)	(0.366)
1(Medium Education)	0.172	0.195	-0.0728	-0.0586	0.334	0.464**	-0.0128	-0.0128	0.183	0.0626	0.334	0.499*
x 1(55 < Age <= 65)	(0.220)	(0.237)	(0.203)	(0.217)	(0.204)	(0.218)	(0.205)	(0.216)	(0.241)	(0.255)	(0.259)	(0.275)
1(Male)	-0.0813	-0.134	0.0132	0.0935	-0.113	-0.132	-0.0651	0.144	0.225	0.273	-0.241	-0.425
x 1(65 < Age <= 75)	(0.286)	(0.306)	(0.229)	(0.243)	(0.260)	(0.276)	(0.255)	(0.267)	(0.300)	(0.319)	(0.344)	(0.365)
1(High Education)	-0.281	-0.202	-0.467	-0.771**	-0.268	-0.357	-0.161	-0.000744	0.309	0.104	0.604	0.786*
x 1(65 < Age <= 75)	(0.415)	(0.440)	(0.316)	(0.345)	(0.365)	(0.402)	(0.362)	(0.379)	(0.415)	(0.443)	(0.449)	(0.478)
1(Medium Education)	-0.286	-0.577*	-0.247	-0.396	0.0808	0.106	-0.376	-0.492*	-0.351	-0.568*	0.504	0.607
x 1(65 < Age <= 75)	(0.304)	(0.326)	(0.245)	(0.260)	(0.281)	(0.298)	(0.283)	(0.296)	(0.321)	(0.342)	(0.353)	(0.375)
1(Male)	-1.120**	-1.389***	0.226	0.561	-0.402	-0.253	0.164	-0.300	0.617	0.551	-0.719	-28.95***
x 1(Age > 75)	(0.485)	(0.524)	(0.425)	(0.473)	(0.472)	(0.508)	(0.417)	(0.444)	(0.540)	(0.585)	(0.794)	(0.968)
1(High Education)	0	0	-4.287	-4.475	-4.356	-7.054	0.285	-3.076	-3.783	-5.666	-3.752	5.605
x 1(Age > 75)	(.)	(.)	(94.51)	(388.3)	(119.6)	(1361.1)	(1.071)	(53.57)	(135.1)	(617.5)	(204.9)	(.)
1(Medium Education)	-0.992*	-1.048*	0.139	0.0230	0.935**	0.970*	-0.185	0.0387	0.460	0.687	-4.192	-7.619
x 1(Age > 75)	(0.526)	(0.569)	(0.429)	(0.474)	(0.470)	(0.506)	(0.432)	(0.455)	(0.528)	(0.575)	(99.44)	(.)
Constant		-1.714***		-1.281***		-1.126***		-1.055***		-1.682***		-2.086***
		(0.112)		(0.0981)		(0.0976)		(0.0878)		(0.112)		(0.134)
Cutoff 1(probit)/	0.521***	-0.743***	-0.147*	-0.651***	0.0459	-0.434***	0.0990	-0.390***	0.458***	-0.660***	1.139***	-0.465***
theta 1 (HOPIT)	(0.0941)	(0.0481)	(0.0871)	(0.0350)	(0.0886)	(0.0298)	(0.0825)	(0.0237)	(0.0946)	(0.0422)	(0.115)	(0.0364)
Cutoff 2(probit)/	1.066***	-0.436***	0.621***	-0.504***	0.737***	-0.209***	0.549***	-0.185***	0.973***	-0.386***	1.515***	-0.218***
theta 2 (HOPIT)	(0.0980)	(0.0346)	(0.0881)	(0.0311)	(0.0906)	(0.0250)	(0.0834)	(0.0189)	(0.0979)	(0.0298)	(0.119)	(0.0223)
Cutoff 3 (probit)/	1.662***	0.506***	1.240***	0.495***	1.346***	0.485***	1.126***	0.367***	1.448***	0.392***	2.196***	0.316***
sigma v (HOPIT)	(0.109)	(0.0305)	(0.0943)	(0.0223)	(0.0980)	(0.0234)	(0.0880)	(0.0172)	(0.106)	(0.0238)	(0.139)	(0.0234)
Cutoff 4 (probit)/	2.377***	0.369***	2.138***	0.448***	2.274***	0.438***	2.088***	0.399***	2.244***	0.366***	2.780***	0.371***
sigma u (HOPIT)	(0.152)	(0.0240)	(0.129)	(0.0217)	(0.142)	(0.0225)	(0.122)	(0.0192)	(0.150)	(0.0234)	(0.193)	(0.0273)
Observations	1003	1003	1027	1027	1018	1018	1122	1122	944	944	996	996

Notes: t-statistics in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table 4: Simulated Proportion Falling in Healthiest Category, by Gender

IFLS	(1)	(2)	(3)	(4)	(5)	(6)
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
1 Male sample raw data	76.16%	54.95%	65.25%	66.96%	78.53%	85.28%
2 Male sample using Male thresholds	75.11%	55.01%	62.11%	65.97%	77.24%	84.10%
3 Male sample using Female thresholds	70.93%	54.55%	58.64%	62.28%	77.26%	81.16%
4 Female sample using Female thresholds	69.78%	44.92%	53.11%	54.15%	67.44%	84.50%
5 Female sample raw data	71.70%	44.67%	54.15%	55.08%	68.04%	85.60%

HRS						
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
1 Male sample raw data	51.07%	14.96%	38.43%	21.32%	50.81%	68.82%
2 Male sample using Male thresholds	53.02%	17.65%	40.63%	26.21%	52.83%	70.97%
3 Male sample using Female thresholds	59.16%	18.75%	34.08%	26.03%	44.11%	67.56%
4 Female sample using Female thresholds	52.44%	15.47%	36.21%	21.37%	43.53%	71.33%
5 Female sample raw data	50.99%	12.29%	35.10%	17.25%	42.06%	69.09%

ELSA						
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
1 Male sample raw data	64.64%	24.19%	43.39%	34.07%	65.59%	76.19%
2 Male sample using Male thresholds	66.01%	25.59%	46.08%	36.82%	67.36%	78.24%
3 Male sample using Female thresholds	64.38%	19.82%	38.55%	35.48%	58.37%	69.70%
4 Female sample using Female thresholds	60.96%	18.26%	42.33%	24.26%	56.12%	74.13%
5 Female sample raw data	59.56%	17.36%	40.59%	22.93%	54.63%	71.90%

CHARLS						
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
1 Male sample raw data	86.27%	56.60%	66.50%	67.09%	75.05%	81.81%
2 Male sample using Male thresholds	86.64%	57.14%	63.18%	67.70%	74.88%	82.04%
3 Male sample using Female thresholds	84.73%	50.44%	59.29%	59.10%	68.29%	84.97%
4 Female sample using Female thresholds	82.50%	48.78%	52.07%	51.97%	65.89%	77.84%
5 Female sample raw data	82.47%	49.86%	54.43%	53.17%	66.19%	76.30%

Notes:
- Individual cross-sectional sampling weights (without adjustment for non-response or attrition) are used
- Proportions calculated using a HOPIT specification with the following explanatory variables: 3 age dummies, 1(High Ed), 1(Medium Ed), and all age-education interactions

estimated parameters to simulate distributions in various ways, which are summarized in Table 4 and Table 5. Table 4 reports the results of various simulations that compare males to females. Each panel summarizes the results from a different dataset, 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").

The first row of each table simply reports the proportion of 1's in the self-reports raw data for men, while the last row reports the proportion among women. These reflect the same numbers represented graphically in Figure A1. The second row 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. The middle row 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 essentially predicts what the male distribution would look like if they had the same thresholds as women.

In the IFLS data, the middle row narrows the gap between males and females in all domains except affect. In the HRS, the gap is narrowed for cognition, sleep, affect, and breathing, but widened in mobility and pain. The ELSA data also show substantial reductions in differences across all domains. In the CHARLS, the gender gap is close to eliminated in the pain domain and is narrowed in several others. In all of the datasets, the significance of the reductions or increases that take place are often unclear.

On the other hand, Table 5, which summarizes the results of this same analysis conducted instead to compare high-education to lower-education individuals, shows a more universal pattern across countries. Across the overwhelming majority of domains and datasets, 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, all domains for the 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 healthiest category increases when predicted using the same thresholds as lower-education individuals. As discussed in section 6.1, this is 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 absolutely no difficulty with a particular health problem. This results in an understatement of differences across education levels.

6.3 Standard Errors for Simulated Probabilities

The preceding discussion about the importance of response thresholds has been based on simply comparing one simulated proportion to another, without considering standard errors. 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. Despite this, existing literature has conducted this type of analysis without calculating standard errors for the simulated proportions. 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 low 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, masks even larger differences.

However, certain types of analysis, like 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

Table 5: Simulated Proportion Falling in Healthiest Category, by Education Level

IFLS	(1)	(2)	(3)	(4)	(5)	(6)
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
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	81.48%	57.79%	58.93%	67.51%	74.40%	84.25%
3 High-Ed sample using Lower-Ed thresholds	89.33%	69.26%	69.03%	68.81%	82.27%	89.33%
4 Lower-Ed sample using Lower-Ed thresholds	70.54%	48.18%	57.67%	58.68%	72.32%	84.34%
5 Lower-Ed sample raw data	71.52%	48.58%	59.68%	59.46%	73.44%	85.11%

HRS						
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
1 High-Ed sample raw data	64.25%	18.18%	45.75%	23.70%	54.13%	79.89%
2 High-Ed sample using High-Ed thresholds	64.86%	21.53%	46.44%	27.57%	55.23%	80.55%
3 High-Ed sample using Lower-Ed thresholds	70.17%	29.71%	63.13%	43.60%	60.84%	90.22%
4 Lower-Ed sample using Lower-Ed thresholds	47.98%	14.26%	34.58%	22.73%	44.82%	67.63%
5 Lower-Ed sample raw data	45.82%	11.70%	33.04%	17.33%	42.89%	64.70%

ELSA						
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
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	70.94%	25.55%	50.70%	30.08%	63.10%	82.29%
3 High-Ed sample using Lower-Ed thresholds	67.69%	29.93%	62.09%	46.43%	78.29%	84.72%
4 Lower-Ed sample using Lower-Ed thresholds	58.08%	19.27%	39.83%	30.51%	60.80%	72.24%
5 Lower-Ed sample raw data	57.03%	18.15%	37.79%	27.99%	58.62%	69.67%

CHARLS						
	Mobility	Pain	Cognition	Sleep	Affect	Breathing
1 High-Ed sample raw data	91.92%	61.82%	72.66%	70.71%	81.09%	86.01%
2 High-Ed sample using High-Ed thresholds	91.61%	62.52%	67.96%	70.40%	80.52%	88.01%
3 High-Ed sample using Lower-Ed thresholds	88.20%	63.18%	62.72%	72.27%	85.10%	93.38%
4 Lower-Ed sample using Lower-Ed thresholds	80.61%	47.28%	50.93%	52.95%	64.29%	75.52%
5 Lower-Ed sample raw data	80.10%	48.30%	52.75%	53.37%	64.24%	75.33%

Notes:

-"Lower-Ed" pools both the medium and low education categories.

- Individual cross-sectional sampling weights (without adjustment for non-response or attrition) are used

- Proportions calculated using a HOPIT specification with the following explanatory variables: three age dummies, 1(Male), and all age-gender interactions

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. Are the groups significantly different from each other after the threshold adjustment? This second issue is also relevant to some education comparisons, where differences appeared trivial to begin with and widened after the DIF adjustment: in short, standard errors are necessary in order to determine whether the groups are significantly different before and after adjustment.

In order to assess the statistical significance of the differences between sub-groups, before and after accounting for thresholds, I calculate standard errors for two differences: first, 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); second, 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. The formulas for the estimated variances, which take into account correlated covariates within married couples) are in Appendix section B (equation A11 for the gender differences and A12 for the education differences).

In Tables 6 and 7, I report 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 dataset, 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. For instance, the second rows of Table 6 and Table 7 (the pain domain of the IFLS) report two significant t-statistics for both the gender differences and the education differences, implying that the significant differences that existed before adjustment remained even after the DIF-adjustment of thresholds. On the other hand, in the cognition domain of the IFLS, the gender difference starts out significant but becomes insignificant after adjustment (in Table 6), while the opposite is true of the education

Table 6: Standard Errors and t-statistics for Simulated Gender Differences

IFLS	(1)	(2)	(3)	(4)	(5)	(6)
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	**t-statistic**	Gender Difference	Standard Error	**t-statistic**
Mobility	0.0532	0.0326	**1.6328**	0.0114	0.0423	**0.2702**
Pain	0.1009	0.0304	**3.3228**	0.0962	0.0373	**2.5809**
Cognition	0.0901	0.0313	**2.8747**	0.0553	0.0360	**1.5361**
Sleep	0.1182	0.0299	**3.9489**	0.0813	0.0313	**2.6012**
Affect	0.0979	0.0336	**2.9149**	0.0982	0.0443	**2.2185**
Breathing	-0.0040	0.0299	**-0.1332**	-0.0335	0.0355	**-0.9431**

HRS						
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	t-statistic	Gender Difference	Standard Error	t-statistic
Mobility	0.0058	0.0195	**0.2956**	0.0673	0.0302	**2.2287**
Pain	0.0218	0.0118	**1.8507**	0.0328	0.0159	**2.0686**
Cognition	0.0442	0.0178	**2.4831**	-0.0213	0.0336	**-0.6349**
Sleep	0.0484	0.0141	**3.4432**	0.0466	0.0204	**2.2846**
Affect	0.0929	0.0183	**5.0751**	0.0058	0.0376	**0.1528**
Breathing	-0.0037	0.0205	**-0.1783**	-0.0377	0.0410	**-0.9209**

ELSA						
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	**t-statistic**	Gender Difference	Standard Error	**t-statistic**
Mobility	0.0505	0.0210	**2.4016**	0.0342	0.0313	**1.0947**
Pain	0.0733	0.0172	**4.2600**	0.0156	0.0220	**0.7109**
Cognition	0.0375	0.0209	**1.7979**	-0.0378	0.0435	**-0.8692**
Sleep	0.1256	0.0189	**6.6610**	0.1122	0.0247	**4.5330**
Affect	0.1124	0.0202	**5.5600**	0.0225	0.0479	**0.4688**
Breathing	0.0412	0.0186	**2.2163**	-0.0442	0.0447	**-0.9887**

CHARLS						
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	**t-statistic**	Gender Difference	Standard Error	**t-statistic**
Mobility	0.0415	0.0433	**0.9582**	0.0223	0.0487	**0.4591**
Pain	0.0835	0.0357	**2.3382**	0.0166	0.0360	**0.4607**
Cognition	0.1111	0.0484	**2.2959**	0.0722	0.0483	**1.4944**
Sleep	0.1573	0.0431	**3.6529**	0.0713	0.0423	**1.6877**
Affect	0.0899	0.0420	**2.1387**	0.0240	0.0516	**0.4651**
Breathing	0.0420	0.0420	**1.0003**	0.0714	0.0513	**1.3909**

Table 7: Standard Errors and t-statistics for Simulated Education Differences

IFLS	(1)	(2)	(3)	(4)	(5)	(6)
	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	**t-statistic**	Education Difference	Standard Error	**t-statistic**
Mobility	0.1094	0.0519	**2.1081**	0.1879	0.0592	**3.1753**
Pain	0.0962	0.0433	**2.2190**	0.2108	0.0549	**3.8402**
Cognition	0.0125	0.0421	**0.2980**	0.1136	0.0471	**2.4133**
Sleep	0.0883	0.0417	**2.1173**	0.1013	0.0446	**2.2692**
Affect	0.0208	0.0510	**0.4081**	0.0996	0.0596	**1.6716**
Breathing	-0.0009	0.0458	**-0.0195**	0.0499	0.0501	**0.9965**

HRS						
	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	t-statistic	Education Difference	Standard Error	t-statistic
Mobility	0.1688	0.0240	**7.0444**	0.2219	0.0367	**6.0434**
Pain	0.0727	0.0150	**4.8450**	0.1545	0.0263	**5.8707**
Cognition	0.1186	0.0214	**5.5520**	0.2855	0.0479	**5.9636**
Sleep	0.0484	0.0171	**2.8278**	0.2087	0.0296	**7.0611**
Affect	0.1041	0.0221	**4.7165**	0.1602	0.0460	**3.4842**
Breathing	0.1292	0.0249	**5.1989**	0.2259	0.0455	**4.9676**

ELSA						
	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	**t-statistic**	Education Difference	Standard Error	**t-statistic**
Mobility	0.1286	0.0214	**6.0014**	0.0962	0.0330	**2.9133**
Pain	0.0628	0.0180	**3.4887**	0.1066	0.0273	**3.9062**
Cognition	0.1087	0.0213	**5.1001**	0.2225	0.0466	**4.7772**
Sleep	-0.0042	0.0198	**-0.2149**	0.1593	0.0264	**6.0434**
Affect	0.0230	0.0210	**1.0928**	0.1749	0.0454	**3.8519**
Breathing	0.1004	0.0192	**5.2360**	0.1248	0.0460	**2.7123**

CHARLS						
	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	**t-statistic**	Education Difference	Standard Error	**t-statistic**
Mobility	0.1099	0.0535	**2.0541**	0.0759	0.0587	**1.2933**
Pain	0.1524	0.0397	**3.8423**	0.1590	0.0424	**3.7526**
Cognition	0.1703	0.0531	**3.2080**	0.1179	0.0514	**2.2922**
Sleep	0.1745	0.0499	**3.4976**	0.1931	0.0552	**3.5009**
Affect	0.1623	0.0517	**3.1409**	0.2081	0.0651	**3.1944**
Breathing	0.1248	0.0559	**2.2328**	0.1786	0.0660	**2.7053**

difference (in Table 7).

The gender results reported in Table 6 reveal an important role for reporting behavior in explaining the gender gap in two out of the four datasets. In the ELSA, five domains show significant differences before adjustment, but only one (sleep) remains significant after using the same thresholds to simulate the probabilities. In the CHARLS data, four domains start out with differences significant at the 5% level, but none remain significant after adjusting for DIF. For these two datasets, it is clear that reporting differences are driving the majority of the significant gender differences that show up in naive comparisons.

On the other hand, in the IFLS, out of the four domains which show significant differences prior to adjustment, the differences in pain, sleep, and affect remain significant even after adjustment, although all of the differences are narrowed. In the HRS, significant differences in mobility, pain, and sleep remain even after adjusting for thresholds. Interestingly, the significant difference in the mobility domain arises only after adjusting for thresholds, suggesting that DIF does distort naive comparisons, although instead by masking existing differences instead of generating spurious ones.

Affect and breathing show no significant gender differences after adjustment in any of the four datasets. Sleep appears to be the domain in which gender differences are most prevalent and pronounced even after adjustment.

Table 7 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 only exist 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 domains. The CHARLS shows significant differences in all six domains before adjustment, but for mobility, this differences narrows and becomes insignificant after adjusting for DIF. Despite this, across all datasets including the CHARLS, education differences are generally quite large and persistent. For pain, cognition,

and sleep, all datasets show significant differences across education levels after accounting for reporting heterogeneity.

6.4 Testing Vignette Equivalence

In this section, I test vignette equivalence using the IFLS data and explore an adjustment to the model to account for potential violations. If the major concern with the vignette equivalence assumption is that individual characteristics may affect the interpretation and understanding of vignettes, a natural solution and also a rigorous test would be to include individual characteristics in the vignette health equation (equation 4), which in the basic model only includes a constant and an error. Following Bago d'Uva et al. (2011), I run the HOPIT model again but replace the original vignette equations with equations 4a and 4b. Table 8 displays the results from the original HOPIT model (which assumed vignette equivalence) and compares this to the model which tests for vignette equivalence by including covariates in the vignette equations. I report the coefficients from equation 1 in the basic HOPIT, equation 1 in the HOPIT testing for vignette equivalence, then equations 4a and 4b in the vignette equivalence tests. The first 4 columns show the results from the pain domain in the IFLS and the last 4 show the results from the cognition domain (results for all other domains and datasets available upon request).

To test whether the covariates belong in the vignette equations, I run a likelihood ratio test. For the cognition domain, the likelihood ratio test cannot reject the null of vignette equivalence. Moreover, both the AIC and the BIC prefer the simpler model. For the pain domain, I reject the null that all coefficients in the vignette equivalence equations are equal to zero, which translates to a rejection of the vignette equivalence assumption. However, in order to judge the severity of the consequences of this violation, I compare the coefficients in the first two columns of each dataset, equation 1 in the basic model and equation 1 in the enhanced model. None of the coefficients are significantly different from each other, which suggests that including the covariates in the vignette equations leaves all major interpretations unchanged. In fact, even for this domain, both the AIC and the BIC prefer the simpler model. In short, although vignette equivalence may not

Table 8: Testing Vignette Equivalence

	Pain				Cognition			
	Assuming V.E.	Testing V.E.	Vignette Equation 1	Vignette Equation 2	Assuming V.E.	Testing V.E.	Vignette Equation 1	Vignette Equation 2
1(55 < Age <= 65)	0.474**	0.449**	-0.0553	0.00319	0.106	0.195	0.164	0.0853
	(0.187)	(0.203)	(0.129)	(0.126)	(0.180)	(0.193)	(0.114)	(0.112)
1(65 < Age <= 75)	0.750***	0.656***	-0.135	-0.0952	0.561**	0.554**	0.0177	-0.0330
	(0.219)	(0.239)	(0.156)	(0.151)	(0.231)	(0.248)	(0.154)	(0.151)
1(Age > 75)	-0.157	-0.201	0.0827	-0.158	0.168	0.342	0.0706	0.512
	(0.438)	(0.476)	(0.293)	(0.293)	(0.482)	(0.528)	(0.334)	(0.315)
1(Male)	-0.198**	-0.240**	-0.0491	-0.0614	-0.136	-0.130	0.0272	-0.0187
	(0.0954)	(0.104)	(0.0628)	(0.0615)	(0.0962)	(0.103)	(0.0576)	(0.0563)
1(High Education)	-0.330**	-0.542***	-0.320***	-0.203**	-0.314**	-0.286**	0.0629	0.000902
	(0.130)	(0.144)	(0.0870)	(0.0840)	(0.123)	(0.132)	(0.0745)	(0.0728)
1(Medium Education)	-0.0796	-0.146	-0.0702	-0.109	-0.338***	-0.304**	0.0371	0.0586
	(0.109)	(0.119)	(0.0719)	(0.0709)	(0.114)	(0.121)	(0.0687)	(0.0670)
1(Male)	-0.265	-0.229	0.0961	0.000477	-0.279	-0.314	-0.107	0.0143
x 1(55 < Age <= 65)	(0.191)	(0.210)	(0.132)	(0.128)	(0.195)	(0.209)	(0.121)	(0.118)
1(High Education)	-0.227	-0.266	-0.0785	-0.00127	0.194	-0.153	-0.562***	-0.288*
x 1(55 < Age <= 65)	(0.273)	(0.303)	(0.187)	(0.179)	(0.275)	(0.299)	(0.178)	(0.167)
1(Medium Education)	-0.0586	-0.167	-0.162	-0.135	0.464**	0.428*	-0.0548	-0.0438
x 1(55 < Age <= 65)	(0.217)	(0.237)	(0.150)	(0.146)	(0.218)	(0.232)	(0.134)	(0.131)
1(Male)	0.0935	0.255	0.289*	0.163	-0.132	-0.0674	0.0772	0.0940
x 1(65 < Age <= 75)	(0.243)	(0.266)	(0.172)	(0.167)	(0.276)	(0.299)	(0.184)	(0.180)
1(High Education)	-0.771**	-0.531	0.201	0.421*	-0.357	-0.664	-0.295	-0.470*
x 1(65 < Age <= 75)	(0.345)	(0.378)	(0.234)	(0.228)	(0.402)	(0.442)	(0.262)	(0.261)
1(Medium Education)	-0.396	-0.448	-0.156	-0.0159	0.106	-0.0561	-0.236	-0.206
x 1(65 < Age <= 75)	(0.260)	(0.284)	(0.181)	(0.178)	(0.298)	(0.322)	(0.198)	(0.192)
1(Male)	0.561	0.790	0.408	0.197	-0.253	-0.437	-0.114	-0.442
x 1(Age > 75)	(0.473)	(0.517)	(0.320)	(0.321)	(0.508)	(0.555)	(0.354)	(0.339)
1(High Education)	-4.475	-7.128	0.483	-5.808	-7.054	-10.88	-3.770	-4.025
x 1(Age > 75)	(388.3)	(344.2)	(361.7)	(296.3)	(1361.1)	(1090.4)	(53.26)	(53.26)
1(Medium Education)	0.0230	-0.110	-0.331	-0.0281	0.970*	0.677	-0.229	-0.691**
x 1(Age > 75)	(0.474)	(0.516)	(0.318)	(0.320)	(0.506)	(0.553)	(0.357)	(0.345)
Constant	-1.281***	-1.178***	-0.514***	-0.375***	-1.126***	-1.139***	-0.466***	-0.215***
	(0.0981)	(0.103)	(0.0629)	(0.0607)	(0.0976)	(0.103)	(0.0608)	(0.0569)
Observations	1027	1027			1018	1018		
AIC	10696.62	10707.4			10672.8	10701.35		
BIC	11096.3	11260.05			11086.55	11262.86		
LR test Chi-squared stat	51.22				31.46			
LR test df	30				30			
LR test p-value	0.0126				0.393			

Notes: t-statistics in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

hold in this particular scenario, adjusting the model to allow for violations does little to change the conclusions. In fact, using the IFLS data, the likelihood ratio test rejects the null for four out of the six domains, but in none of these domains are the β coefficients from the simple and enhanced model significantly different from each other. This result is not unique to the IFLS. In the other three datasets, although vignette equivalence is more consistently rejected across domains (as in Bago d'Uva et al. (2011) using the ELSA data), the vast majority of coefficients are statistically indistinguishable across specifications.

7 Conclusion

Anchoring vignettes are a vital tool that can be used to account for reporting bias in subjective scale measures. I demonstrate that accounting for DIF is crucial to understanding what really determines self-reported health. Ignoring DIF underestimates the differences in health across education levels in Indonesia, the United States, England, and China, even though the schooling distributions are drastically different across these countries.

The evidence on gender differences is more mixed but suggests a larger role for reporting thresholds. Although significant differences between males and females remain in three out of the six domains for the IFLS and HRS even after adjusting for thresholds, in England and China, accounting for thresholds completely eliminates significant differences between males and females in all but one domain (sleep in the ELSA). 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 paper, these findings suggest that the gender of the *object* of evaluation, 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 paper.[16] What I can conclude from this analysis is that, irrespective of the reasons for their use of different

[16] Although some studies are able to include vignette gender as a variable in the vignette latent variable equation, I do not have this information for all four datasets.

thresholds, males and females in the ELSA and CHARLS would report much more similar levels of health if they used the same thresholds.

Although the role of thresholds in explaining the within-country differences is certainly country and domain-specific, anchoring vignettes offer informative insight into these comparisons and highlight some surprisingly universal findings.

A Likelihood Function

In order to express the log-likelihood function, I define the indicator function $D_{ijj_1j_2j_3} = 1(Y_i = j, Y_{1i} = j_1, Y_{2i} = j_2, Y_{3i} = j_3)$. Then,

$$L(\beta, \gamma, \theta, \sigma_v, \sigma_u) \;=\; \prod_{i=1}^{N} \prod_{j_3=1}^{5} \prod_{j_2=1}^{5} \prod_{j_1=1}^{5} Pr(Y_i = j, Y_{1i} = j_1, Y_{2i} = j_2, Y_{3i} = j_3)^{D_{ijj_1j_2j_3}}.$$

I calculate the unconditional likelihood contribution $\prod_{j_3=1}^{5} \prod_{j_2=1}^{5} \prod_{j_1=1}^{5} Pr(Y_i = j, Y_{1i} = j_1, Y_{2i} = j_2, Y_{3i} = j_3)^{D_{ijj_1j_2j_3}}$ by taking the average of the following conditional likelihood contribution over 50 simulated u_i's (from a standard normal distribution) for each individual.

$$\begin{aligned}
& Pr(Y_i = j, Y_{1i} = j_1, Y_{2i} = j_2, Y_{3i} = j_3 | u_i) \\
= \quad & \left[\Phi(\tau_i^j(u_i) - \beta X_i) - \Phi(\tau_i^{j-1}(u_i) - \beta X_i)\right] \left[\Phi(\frac{\tau_i^{j_1}(u_i) - \theta_1}{\sigma_v}) - \Phi(\frac{\tau_i^{j_1-1}(u_i) - \theta_1}{\sigma_v})\right] \\
& \left[\Phi(\frac{\tau_i^{j_2}(u_i) - \theta_2}{\sigma_v}) - \Phi(\frac{\tau_i^{j_2-1}(u_i) - \theta_2}{\sigma_v})\right] \left[\Phi(\frac{\tau_i^{j_3}(u_i)}{\sigma_v}) - \Phi(\frac{\tau_i^{j_3-1}(u_i)}{\sigma_v})\right]. \qquad \text{(A1)}
\end{aligned}$$

For j, j_1, j_2, $j_3 > 2$, this becomes

$$\begin{aligned}
= \quad & \left[\Phi((\gamma^1 - \beta)X_i + \sum_{n=2}^{j} e^{\gamma_n X_i} + \sigma_u u_i) - \Phi((\gamma^1 - \beta)X_i + \sum_{n=2}^{j-1} e^{\gamma_n X_i} + \sigma_u u_i)\right] \\
& \left[\Phi(\frac{\gamma^1 X_i - \theta_1 + \sum_{n=2}^{j_1} e^{\gamma_n X_i} + \sigma_u u_i}{\sigma_v}) - \Phi(\frac{\gamma^1 X_i - \theta_1 + \sum_{n=2}^{j_1-1} e^{\gamma_n X_i} + \sigma_u u_i}{\sigma_v})\right] \\
& \left[\Phi(\frac{\gamma^1 X_i - \theta_2 + \sum_{n=2}^{j_2} e^{\gamma_n X_i} + \sigma_u u_i}{\sigma_v}) - \Phi(\frac{\gamma^1 X_i - \theta_2 + \sum_{n=2}^{j_2-1} e^{\gamma_n X_i} + \sigma_u u_i}{\sigma_v})\right] \\
& \left[\Phi(\frac{\gamma^1 X_i + \sum_{n=2}^{j_3} e^{\gamma_n X_i} + \sigma_u u_i}{\sigma_v}) - \Phi(\frac{\gamma^1 X_i + \sum_{n=2}^{j_3-1} e^{\gamma_n X_i} + \sigma_u u_i}{\sigma_v})\right].
\end{aligned}$$

This follows directly from equation A1 and the formulas for the τ_i's in equation 3 in Section 3 . The individual likelihood contributions for j, j_1, j_2, $j_3 \leq 2$ can be obtained in the same way.

B Standard Error Derivations

B.1 General Case

I begin with the general case and in the next sub-section specialize to the setting relevant to this paper. I define $\hat{f}$, an estimate of a population proportion, as

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N} h(X_i, \hat{\theta}),$$

where $h(X, \theta)$ is a continuous and differentiable function. In my application, $0 \leq h(X, \theta) \leq 1$. The parameter vector θ is estimated in a preliminary step and is $\sqrt{N}$ consistent with

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$$

I define $\tilde{f}$ as the sample fraction calculated using the true parameter:

$$\tilde{f} = \frac{1}{N}\sum_{i=1}^{N} h(X_i, \theta_0).$$

The population fraction is

$$f = E[h(X, \theta_0)]$$

where the expectation is over the joint distribution of X. If the uniform law of large numbers (ULLN) holds, or in other words, if

$$E\left[\sup_{\theta \in \Theta} h(X, \theta)\right] < \infty$$

then $\hat{f} \xrightarrow{p} f$. I decompose the difference between my estimated $\hat{f}$ and the population proportion f into two parts:

$$\sqrt{N}(\hat{f} - f) = \sqrt{N}(\hat{f} - \tilde{f}) + \sqrt{N}(\tilde{f} - f) \tag{A2}$$

I start with the first term. By the mean value theorem,

$$\hat{f} = \tilde{f} + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial h}{\partial \theta}(X_i, \overline{\theta})'(\hat{\theta} - \theta_0),$$

where $\overline{\theta}$ is a random variable strictly between $\hat{\theta}$ and θ_0.

If

$$E\left[\sup_{\theta \in \Theta} \frac{\partial h}{\partial \theta}(X, \theta)\right] < \infty,$$

then another application of the ULLN and the Slutsky theorem gives

$$\sqrt{N}(\hat{f} - \tilde{f}) = E\left[\frac{\partial h}{\partial \theta}(X, \theta_0)\right]' \sqrt{N}(\hat{\theta} - \theta_0) + o_p(1),$$

so the asymptotic variance of the asymptotic normal distribution is

$$\mathrm{Var}(\hat{f} - \tilde{f}) = \frac{1}{N} E\left[\frac{\partial h}{\partial \theta}(X, \theta_0)\right]' V E\left[\frac{\partial h}{\partial \theta}(X, \theta_0)\right] \equiv \frac{\sigma^2}{N}. \tag{A3}$$

Moving on to the second term of equation A2, we have that

$$\sqrt{N}(\tilde{f} - f) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (h(X_i, \theta_0) - E(h(X, \theta_0))).$$

By the central limit theorem, this has an asymptotic normal distribution with variance

$$\mathrm{Var}(\tilde{f} - f) = \frac{1}{N}\mathrm{Var}(h(X, \theta_0))) \equiv \frac{s^2}{N}. \tag{A4}$$

Because $\sqrt{N}(\hat{f} - \tilde{f})$ and $\sqrt{N}(\tilde{f} - f)$ are independent,

$$\mathrm{Var}(\hat{f} - f) = \frac{\sigma^2}{N} + \frac{s^2}{N}. \tag{A5}$$

where $\frac{\sigma^2}{N}$ is defined by equation A3 , and $\frac{s^2}{N}$ is defined by equation A4.

B.2 Standard Errors for Proportion Differences

In this paper, rather than the standard error of an estimated proportion, I am interested in the standard error of a *difference* between estimated proportions. In fact, there are two differences of interest. The first is the difference between the estimated proportion of males and females (or high vs lower-education groups) who fall into the healthiest category, calculated using their own group's coefficients to estimate the model. I will denote these $\hat{p}_m$ and $\hat{p}_f$, respectively. The second comparison is the difference between the simulated proportion of healthy males predicted using female thresholds (which I will denote $\hat{p}_g$) and the simulated proportion of healthy females using female thresholds (the same $\hat{p}_f$ as above). 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. As the calculation of standard errors for ($\hat{p}_m$ - $\hat{p}_f$) is a special case of the more complex second comparison, I focus on the the latter: the difference between $\hat{p}_g$ and $\hat{p}_f$.

I formally define $\hat{p}_g$ as

$$\begin{aligned}
\hat{p}_g &= \frac{1}{N_m}\sum_{i\in M} Pr(X_i'\hat{\beta}_m + \epsilon_i \leq X_i'\hat{\gamma}_f + u_i) \\
&= \frac{1}{N_m}\sum_{i\in M} Pr(\epsilon_i - u_i \leq X_i'(\hat{\gamma}_f - \hat{\beta}_m)) \\
&= \frac{1}{N_m}\sum_{i\in M} \Phi(\frac{X_i'(\hat{\gamma}_f - \hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})
\end{aligned}$$

The m and f subscripts indicate the sample (male or female) used to estimate the coefficients. For simplicity, I omit the 1 superscript in γ^1 as this is the only γ vector that is relevant to this discussion. Defining $\hat{p}_f$ using these coefficient subscripts,

$$\hat{p}_f = \frac{1}{N_f}\sum_{i\in F} \Phi(\frac{X_i'(\hat{\gamma}_f - \hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}}),$$

it is clear now that I cannot simply calculate Var($\hat{p}_f$) and Var($\hat{p}_g$) separately because of the common $\hat{\gamma}_f$ and $\hat{\sigma}_{uf}$. Therefore, I consider the difference (rather than the individual proportions) as my

estimate of interest:

$$\hat{f} \;=\; \hat{p}_g - \hat{p}_f$$

Defining

$$h(X, \beta, \gamma, \sigma_u) = \Phi(\frac{X_i'(\gamma - \beta)}{\sqrt{1 + \sigma_u^2}}),$$

I have

$$\begin{aligned} \hat{f} &= \frac{1}{N_m} \sum_{i \in M} h(X_i, \hat{\beta}_m, \hat{\gamma}_f, \hat{\sigma}_{uf}) - \frac{1}{N_f} \sum_{i \in F} h(X_i, \hat{\beta}_f, \hat{\gamma}_f, \hat{\sigma}_{uf}) \\ &\equiv f(\hat{\theta}, X) \end{aligned}$$

where in the last line I define

$$\hat{\theta} = (\hat{\beta}_m \; \hat{\beta}_f \; \hat{\alpha})',$$

grouping the common parameters together and letting $\hat{\alpha} \equiv (\hat{\gamma}_f \; \hat{\sigma}_{uf})'$.

The analogous sample difference, calculated using true parameters, is

$$\begin{aligned} \tilde{f} &= \tilde{p}_g - \tilde{p}_f \\ &= \frac{1}{N_m} \sum_{i \in M} h(X_i, \beta_{m0}, \gamma_{f0}, \sigma_{uf0}) - \frac{1}{N_f} \sum_{i \in F} h(X_i, \beta_{f0}, \gamma_{f0}, \sigma_{uf0}) \end{aligned}$$

and the population difference is

$$\begin{aligned} f &= p_g - p_f \\ &= E[h(X_i, \beta_{m0}, \gamma_{f0}, \sigma_{uf0}) | X \in M] - E[h(X_i, \beta_{f0}, \gamma_{f0}, \sigma_{uf0}) | X \in F] \end{aligned}$$

Recalling that the variance of a simulated proportion consists of two terms (as shown in equation A5), I begin with calculating an estimate for the first term, $\frac{\sigma^2}{N}$.

B.2.1 Estimating $\frac{\sigma^2}{N}$

Again using the mean value theorem, I have

$$f(\hat{\theta}, X) = f(\theta_0, X) + \frac{\partial f(X, \bar{\theta})}{\partial \theta}(\hat{\theta} - \theta_0)$$

which can be decomposed into three sums that involve the the male-specific coefficients, the female-specific coefficients, and the common coefficients.

$$\begin{aligned}
\hat{f} &= \tilde{f} + \frac{\partial f(X, \bar{\theta})}{\partial \beta_m}(\hat{\beta}_m - \beta_{m0}) + \frac{\partial f(X, \bar{\theta})}{\partial \beta_f}(\hat{\beta}_f - \beta_{f0}) + \frac{\partial f(X, \bar{\theta})}{\partial \alpha}(\hat{\alpha} - \alpha_0) \\
&= \tilde{f} + \frac{1}{N_m}\sum_{i \in M} \frac{\partial h(X_i, \bar{\beta}_m, \bar{\alpha})}{\partial \beta_m}(\hat{\beta}_m - \beta_{m0}) + \frac{1}{N_f}\sum_{i \in F} \frac{\partial f(X_i, \bar{\beta}_f, \bar{\alpha})}{\partial \beta_f}(\hat{\beta}_f - \beta_{f0}) + \\
&\quad \left(\frac{1}{N_m}\sum_{i \in M} \frac{\partial f(X_i, \bar{\beta}_m, \bar{\alpha})}{\partial \alpha} - \frac{1}{N_f}\sum_{i \in F} \frac{\partial f(X_i, \bar{\beta}_f, \bar{\alpha})}{\partial \alpha} \right)(\hat{\alpha} - \alpha_0)
\end{aligned}$$

The ULLN and Slutsky theorem once again give

$$\sqrt{N}(\hat{f} - \tilde{f}) = \begin{pmatrix} E\left[\frac{\partial h}{\partial \beta_m}(X, \beta_{m0}, \alpha_0) | X \in M\right] \\ E\left[\frac{\partial h}{\partial \beta_f}(X, \beta_{f0}, \alpha_0) | X \in F\right] \\ E\left[\frac{\partial h}{\partial \alpha}(X, \beta_{m0}, \alpha_0) | X \in M\right] - E\left[\frac{\partial h}{\partial \alpha}(X, \beta_{f0}, \alpha_0) | X \in F\right] \end{pmatrix}' \sqrt{N}(\hat{\theta} - \theta_0) + o_p(1)$$

so that the variance of the asymptotic normal distribution is

$$\begin{aligned}
Var(\hat{f} - \tilde{f}) &= \begin{pmatrix} E\left[\frac{\partial h}{\partial \beta_m}(X, \beta_{m0}, \alpha_0) | X \in M\right] \\ E\left[\frac{\partial h}{\partial \beta_f}(X, \beta_{f0}, \alpha_0) | X \in F\right] \\ E\left[\frac{\partial h}{\partial \alpha}(X, \beta_{m0}, \alpha_0) | X \in M\right] - E\left[\frac{\partial h}{\partial \alpha}(X, \beta_{f0}, \alpha_0) | X \in F\right] \end{pmatrix}' \frac{V}{N} \\
&\quad \begin{pmatrix} E\left[\frac{\partial h}{\partial \beta_m}(X, \beta_{m0}, \alpha_0) | X \in M\right] \\ E\left[\frac{\partial h}{\partial \beta_f}(X, \beta_{f0}, \alpha_0) | X \in F\right] \\ E\left[\frac{\partial h}{\partial \alpha}(X, \beta_{m0}, \alpha_0) | X \in M\right] - E\left[\frac{\partial h}{\partial \alpha}(X, \beta_{f0}, \alpha_0) | X \in F\right] \end{pmatrix} = \frac{\sigma^2}{N}
\end{aligned}$$

and can be estimated by

$$\frac{\hat{\sigma}^2}{N} = \begin{pmatrix} -\frac{1}{N_m}\sum_{i\in M}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i \\ -\frac{1}{N_f}\sum_{i\in F}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i \\ \frac{1}{N_m}\sum_{i\in M}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i - \frac{1}{N_f}\sum_{i\in F}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i \\ \frac{1}{N_m}\sum_{i\in M}\frac{-\hat{\sigma}_{uf}}{(1+\hat{\sigma}_{uf}^2)^{\frac{3}{2}}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i'(\hat{\gamma}_f-\hat{\beta}_m) - \frac{1}{N_f}\sum_{i\in F}\frac{-\hat{\sigma}_{uf}}{(1+\hat{\sigma}_{uf}^2)^{\frac{3}{2}}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i'(\hat{\gamma}_f-\hat{\beta}_f) \end{pmatrix}' \frac{\hat{V}}{N} \begin{pmatrix} -\frac{1}{N_m}\sum_{i\in M}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i \\ -\frac{1}{N_f}\sum_{i\in F}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i \\ \frac{1}{N_m}\sum_{i\in M}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i - \frac{1}{N_f}\sum_{i\in F}\frac{1}{\sqrt{1+\hat{\sigma}_{uf}^2}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i \\ \frac{1}{N_m}\sum_{i\in M}\frac{-\hat{\sigma}_{uf}}{(1+\hat{\sigma}_{uf}^2)^{\frac{3}{2}}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i'(\hat{\gamma}_f-\hat{\beta}_m) - \frac{1}{N_f}\sum_{i\in F}\frac{-\hat{\sigma}_{uf}}{(1+\hat{\sigma}_{uf}^2)^{\frac{3}{2}}}\phi(\frac{X_i'(\hat{\gamma}_f-\hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}})X_i'(\hat{\gamma}_f-\hat{\beta}_f) \end{pmatrix} \quad \text{(A6)}$$

Here, because θ involves coefficients from the estimation over the male population and over the female population, the matrix V involves a combination of estimated variance-covariance matrices from both estimations. In particular, let $\frac{\hat{V}_m}{N_m}$ represent the variance-covariance matrix for the male-specific parameters of interest $(\hat{\beta}_m)$, and $\frac{\hat{V}_f}{N_f}$ represent the variance-covariance matrix for the female parameters of interest $(\hat{\beta}_f, \hat{\alpha})$. Then, the relevant variance covariance equation needed for this calculation is

$$\frac{\hat{V}}{N} = \begin{pmatrix} \frac{\hat{V}_m}{N_m} & 0 \\ 0 & \frac{\hat{V}_f}{N_f} \end{pmatrix}.$$

Running separate estimations for males and females, I assume independence of the male and female coefficients. Note that the formula for $\frac{\hat{\sigma}^2}{N}$ (equation A6) can be easily applied to calculating $\frac{\hat{\sigma}^2}{N}$ for the simpler difference, $\hat{p}_m - \hat{p}_f$. The female coefficients in the male summations are replaced by male coefficients, gender-specific γ's and σ_u's are included in the male- and female-specific vectors, and the common coefficients, $\hat{\alpha}$, are dropped. $\frac{\hat{V}_m}{N_m}$ and $\frac{\hat{V}_f}{N_f}$ are simply the variance-covariance matrices from the separately-conducted male estimation and female estimation, respectively.

B.2.2 Estimating $\frac{s^2}{N}$

The calculation of $\frac{s^2}{N}$ is straightforward if I assume the independence of the X's across the male and female populations.

$$\text{Var}(\tilde{f} - f) = \frac{1}{N_m}\text{Var}(h(X, \beta_{m0}, \alpha_0)|X \in M) + \frac{1}{N_f}\text{Var}(h(X, \beta_{f0}, \alpha_0)|X \in F) = \frac{s^2}{N}$$

It can be estimated by

$$\begin{aligned}\frac{\hat{s}^2}{N} &= \frac{1}{N_m}\left(\frac{1}{N_m}\sum_{i\in M}\left(\Phi(\frac{X_i'(\hat{\gamma}_f - \hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}}) - \frac{1}{N_m}\sum_{k\in M}\Phi(\frac{X_k'(\hat{\gamma}_f - \hat{\beta}_m)}{\sqrt{1+\hat{\sigma}_{uf}^2}})\right)^2\right) \\ &+\frac{1}{N_f}\left(\frac{1}{N_f}\sum_{i\in F}\left(\Phi(\frac{X_i'(\hat{\gamma}_f - \hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}}) - \frac{1}{N_f}\sum_{k\in F}\Phi(\frac{X_k'(\hat{\gamma}_f - \hat{\beta}_f)}{\sqrt{1+\hat{\sigma}_{uf}^2}})\right)^2\right). \qquad \text{(A7)}\end{aligned}$$

For simplicity in future notations, I define $\tilde{s}(g(X_i), N)$ as the deviation of a function $g(X_i)$ from its sample mean, from a sample of size N:

$$\tilde{s}(g(X_i), N) = g(X_i) - \frac{1}{N}\sum_{k=1}^{N} g(X_k). \qquad \text{(A8)}$$

Using this notation, I can rewrite equation A7 as

$$\frac{\hat{s}^2}{N} = \frac{1}{N_m^2}\sum_{i\in M}\tilde{s}(h(X_i, \hat{\beta}_m, \hat{\gamma}_f, \hat{\sigma}_{uf}), N_m)^2 + \frac{1}{N_f^2}\sum_{i\in F}\tilde{s}(h(X_i, \hat{\beta}_f, \hat{\gamma}_f, \hat{\sigma}_{uf}), N_f)^2$$

The calculation of $\frac{s^2}{N}$ becomes more complicated if I consider correlations between couples. In all of the surveys used, when a household is (randomly) selected, both husband and wife are included in the sample if both are present and eligible. If couples match non-randomly, this would create correlations across observations (within couples), which violates the assumption of independence across male and female covariates.[17]

[17]For the validity of the maximum likelihood estimation, I require independence across observations conditional on the included covariates. Therefore, if couples only match on age and education (which are included as my regressors),

Though the entire discussion has been framed in terms of the male-female comparison, any formulas described until now can be directly applied to the comparison between educated and non-educated groups. However, taking into account correlations within couples when comparing across high and lower-education groups requires a slightly different approach than what is needed when simply comparing across males and females. I first describe the methods used to account for correlations in the gender comparisons.

Let SM denote the set of single males and N_{SM} the number of individuals in this set. Similarly, let SF represent the set of single females, N_{SF} the number of single females, C the set of individuals belonging to a married couple with both individuals in the sample, and N_C the number of such couples. Within a couple j, let X_j^m represent the characteristics of the male in the couple and X_j^f the characteristics of the female in the couple. With this additional notation, I rewrite $\tilde{f}$ as follows:

$$\begin{aligned}\tilde{f} &= \frac{1}{N_{SM}}\sum_{i\in SM} h(X_i,\beta_{m0},\alpha_0)\frac{N_{SM}}{N_m} - \frac{1}{N_{SF}}\sum_{i\in SF} h(X_i,\beta_{f0},\alpha_0)\frac{N_{SF}}{N_f} + \\ &\quad \frac{1}{N_C}\sum_{j\in C}(h(X_j^m,\beta_{m0},\alpha_0)\frac{N_C}{N_m} - h(X_j^f,\beta_{f0},\alpha_0)\frac{N_C}{N_f}).\end{aligned}$$

Assuming independence across couples but not within couples, I can calculate the asymptotic variance as follows:

$$\begin{aligned}\mathrm{Var}(\tilde{f}-f) &= \frac{1}{N_{SM}}\mathrm{Var}(h(X_i,\beta_{m0},\alpha_0)\frac{N_{SM}}{N_m}) + \frac{1}{N_{SF}}\mathrm{Var}(h(X_i,\beta_{f0},\alpha_0)\frac{N_{SF}}{N_f}) \\ &\quad +\frac{1}{N_C}\mathrm{Var}(h(X_j^m,\beta_{m0},\alpha_0)\frac{N_C}{N_m} - h(X_j^f,\beta_{f0},\alpha_0)\frac{N_C}{N_f}) \\ &= \frac{1}{N_{SM}}\mathrm{Var}(h(X_i,\beta_{m0},\alpha_0)\frac{N_{SM}}{N_m}) + \frac{1}{N_{SF}}\mathrm{Var}(h(X_i,\beta_{f0},\alpha_0)\frac{N_{SF}}{N_f}) \\ &\quad +\frac{1}{N_C}\left[\mathrm{Var}(h(X_j^m,\beta_{m0},\alpha_0)\frac{N_C}{N_m}) + \mathrm{Var}(h(X_j^f,\beta_{f0},\alpha_0)\frac{N_C}{N_f})\right] \\ &\quad -2\frac{1}{N_C}\left[\mathrm{Cov}(h(X_j^m,\beta_{m0},\alpha_0)\frac{N_C}{N_m}, h(X_j^f,\beta_{f0},\alpha_0)\frac{N_C}{N_f})\right].\end{aligned}$$

this conditional independence is not violated.

The corresponding estimate is:

$$\begin{aligned}\frac{\hat{s}^2}{N}_{\text{gender}} &= \frac{1}{N_{SM}^2}\sum_{i\in SM}\tilde{s}(h(X_i,\hat{\beta}_m,\hat{\alpha})\frac{N_{SM}}{N_m},N_{SM})^2+\frac{1}{N_{SF}^2}\sum_{i\in SF}\tilde{s}(h(X_i,\hat{\beta}_f,\hat{\alpha})\frac{N_{SF}}{N_f},N_{SF})^2 \\ &+\frac{1}{N_C^2}\sum_{j\in C}\left[\tilde{s}(h(X_j^m,\hat{\beta}_m,\hat{\alpha})\frac{N_C}{N_m},N_C)^2+\tilde{s}(h(X_j^f,\hat{\beta}_f,\hat{\alpha})\frac{N_C}{N_f},N_C)^2\right] \\ &-2\frac{1}{N_C^2}\sum_{j\in C}\tilde{s}(h(X_j^m,\hat{\beta}_m,\hat{\alpha})\frac{N_C}{N_m},N_C)\tilde{s}(h(X_j^f,\hat{\beta}_f,\hat{\alpha})\frac{N_C}{N_f},N_C) \qquad \text{(A9)}\end{aligned}$$

As mentioned earlier, adjusting for correlations in the education analysis requires a slightly different approach. Here, I break the sample into 6 groups: N_{SH} single educated individuals (in set SH), N_{SL} single lower-education individuals (in set SL), N_{C10} couples (in set $C10$) where the male is educated and the female is not, N_{C11} couples (in set $C11$) where both partners are educated, N_{C00} couples (in set $C00$) where both partners are in the lower-education category, and N_{C01} couples (in set $C01$) where the female is educated but the male is not. Let N_H denote the total number of high-education individuals and N_L denote the total number of lower-education individuals. Using an h subscript to indicate the high education group and the l subscript for the lower-education group, I can therefore re-write $\tilde{f}$ as:

$$\begin{aligned}\tilde{f} &= \frac{1}{N_{SH}}\sum_{i\in SH}h(X_i,\beta_{h0},\alpha_0)\frac{N_{SH}}{N_H}-\frac{1}{N_{SL}}\sum_{i\in SL}h(X_i,\beta_{l0},\alpha_0)\frac{N_{SL}}{N_L}+ \\ &+\frac{1}{N_{C11}}\sum_{j\in C11}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H}+h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H}) \\ &+\frac{1}{N_{C10}}\sum_{j\in C10}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C10}}{N_H}-h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C10}}{N_L}) \\ &+\frac{1}{N_{C01}}\sum_{j\in C01}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C01}}{N_H}-h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C01}}{N_L}) \\ &-\frac{1}{N_{C00}}\sum_{j\in C00}(h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L}+h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L})\end{aligned}$$

The asymptotic variance is

$$
\begin{aligned}
\mathrm{Var}(\tilde{f}-f) \;=\; & \frac{1}{N_{SH}}\mathrm{Var}(h(X_i,\beta_{h0},\alpha_0)\frac{N_{SH}}{N_H}) + \frac{1}{N_{SL}}\mathrm{Var}(h(X_i,\beta_{l0},\alpha_0)\frac{N_{SL}}{N_L}) \\
& + \frac{1}{N_{C11}}\left[\mathrm{Var}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H}) + \mathrm{Var}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H})\right] \\
& + 2\frac{1}{N_{C11}}\mathrm{Cov}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H}, h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H}) \\
& + \frac{1}{N_{C10}}\left[\mathrm{Var}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C10}}{N_H}) + \mathrm{Var}(h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C10}}{N_L})\right] \\
& - 2\frac{1}{N_{C10}}\mathrm{Cov}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C10}}{N_H}, h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C10}}{N_L}) \\
& + \frac{1}{N_{C01}}\left[\mathrm{Var}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C01}}{N_H}) + \mathrm{Var}h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C01}}{N_L})\right] \\
& - 2\frac{1}{N_{C01}}\mathrm{Cov}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C01}}{N_H}, h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C01}}{N_L}) \\
& + \frac{1}{N_{C00}}\left[\mathrm{Var}((h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L}) + \mathrm{Var}(h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L})\right] \\
& + 2\frac{1}{N_{C00}}\mathrm{Cov}(h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L}, h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L}),
\end{aligned}
$$

and sample analog follows directly:

$$
\begin{aligned}
\frac{\hat{s}^2}{N}_{\text{educ}} = & \frac{1}{N_{SH}^2}\sum_{i\in SH}\tilde{s}(h(X_i,\beta_{h0},\alpha_0)\frac{N_{SH}}{N_H},N_{SH})^2+\frac{1}{N_{SL}^2}\sum_{i\in SL}\tilde{s}(h(X_i,\beta_{l0},\alpha_0)\frac{N_{SL}}{N_L},N_{SL})^2 \\
& +\frac{1}{N_{C11}^2}\sum_{i\in C11}\left[\tilde{s}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H},N_{C11})^2+\tilde{s}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H},N_{C11})^2\right] \\
& +2\frac{1}{N_{C11}^2}\sum_{j\in C11}\tilde{s}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H},N_{C11})\tilde{s}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C11}}{N_H},N_{C11}) \\
& +\frac{1}{N_{C10}^2}\sum_{i\in C10}\left[\tilde{s}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C10}}{N_H},N_{C10})^2+\tilde{s}(h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C10}}{N_L},N_{C10})^2\right] \\
& -2\frac{1}{N_{C10}^2}\sum_{j\in C10}\tilde{s}(h(X_j^m,\beta_{h0},\alpha_0)\frac{N_{C10}}{N_H},N_{C10})\tilde{s}(h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C10}}{N_L},N_{C10}) \\
& +\frac{1}{N_{C01}^2}\sum_{i\in C01}\left[\tilde{s}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C01}}{N_H},N_{C01})^2+\tilde{s}h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C01}}{N_L},N_{C01})^2\right] \\
& -2\frac{1}{N_{C01}^2}\sum_{j\in C01}\tilde{s}(h(X_j^f,\beta_{h0},\alpha_0)\frac{N_{C01}}{N_H},N_{C01})\tilde{s}(h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C01}}{N_L},N_{C01}) \\
& +\frac{1}{N_{C00}^2}\sum_{i\in C00}\left[\tilde{s}((h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L},N_{C00})^2+\tilde{s}(h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L},N_{C00})^2\right] \\
& +2\frac{1}{N_{C00}^2}\sum_{j\in C00}\tilde{s}(h(X_j^m,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L},N_{C00})\tilde{s}(h(X_j^f,\beta_{l0},\alpha_0)\frac{N_{C00}}{N_L},N_{C00}). \qquad \text{(A10)}
\end{aligned}
$$

Therefore, the estimate for the variance of $\hat{p}_g-\hat{p}_m$ when comparing across genders is

$$\hat{V}(\hat{p}_g-\hat{p}_m)=\frac{\hat{\sigma}^2}{N}+\frac{\hat{s}^2}{N}_{\text{gender}}, \qquad \text{(A11)}$$

while the estimate for the variance when comparing across education levels can be written

$$\hat{V}(\hat{p}_g-\hat{p}_m)=\frac{\hat{\sigma}^2}{N}+\frac{\hat{s}^2}{N}_{\text{educ}}, \qquad \text{(A12)}$$

where $\frac{\hat{\sigma}^2}{N}$ is defined by equation A6, $\frac{\hat{s}^2}{N}_{\text{gender}}$ by equation A9, and $\frac{\hat{s}^2}{N}_{\text{educ}}$ by equation A10.

Another way to deal with correlations between couples is to randomly select one individual from each household in order to generate a random sample of individuals. Now, with no correlations between men and women or educated and uneducated individuals, I can simply calculate $\frac{\hat{s}^2}{N}$

using equation A7, for both the gender and education comparisons. The results from this analysis are reported in Tables A1 and A2.

Even though the sample sizes have fallen dramatically, the basic story remains the same: education differences either remain constant or are exacerbated after adjusting for DIF. On the other hand, significant differences between males and females lose significance in most domains for the ELSA and CHARLS after accounting for thresholds, while there is stronger evidence for remaining gender differences in the HRS and IFLS.

Table A1: Standard Errors and t-statistics for Simulated Gender Differences using a Random Sample of Individuals

IFLS	(1)	(2)	(3)	(4)	(5)	(6)
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	**t-statistic**	Gender Difference	Standard Error	**t-statistic**
Mobility	0.0249	0.0392	**0.6338**	0.0388	0.0505	**0.7677**
Pain	0.1228	0.0374	**3.2807**	0.0962	0.0454	**2.1198**
Cognition	0.1058	0.0383	**2.7577**	0.0759	0.0433	**1.7511**
Sleep	0.1156	0.0359	**3.2205**	0.0658	0.0367	**1.7907**
Affect	0.1074	0.0409	**2.6271**	0.1207	0.0552	**2.1855**
Breathing	0.0183	0.0361	**0.5065**	-0.0061	0.0454	**-0.1334**

HRS						
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	t-statistic	Gender Difference	Standard Error	t-statistic
Mobility	0.0110	0.0218	**0.5021**	0.1045	0.0340	**3.0769**
Pain	0.0202	0.0134	**1.5066**	0.0476	0.0185	**2.5755**
Cognition	0.0453	0.0199	**2.2768**	-0.0037	0.0374	**-0.0982**
Sleep	0.0344	0.0156	**2.2034**	0.0598	0.0236	**2.5312**
Affect	0.0890	0.0202	**4.3940**	0.0307	0.0414	**0.7404**
Breathing	0.0073	0.0231	**0.3171**	-0.0340	0.0463	**-0.7341**

ELSA						
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	**t-statistic**	Gender Difference	Standard Error	**t-statistic**
Mobility	0.0568	0.0216	**2.6267**	0.0553	0.0324	**1.7068**
Pain	0.0819	0.0179	**4.5860**	0.0093	0.0222	**0.4189**
Cognition	0.0314	0.0215	**1.4591**	-0.0353	0.0438	**-0.8062**
Sleep	0.1261	0.0197	**6.3974**	0.1112	0.0253	**4.3912**
Affect	0.1142	0.0210	**5.4384**	0.0181	0.0487	**0.3712**
Breathing	0.0401	0.0192	**2.0882**	-0.0336	0.0455	**-0.7384**

CHARLS						
	Using Different Thresholds			Using Same Thresholds		
Domain	Gender Difference	Standard Error	**t-statistic**	Gender Difference	Standard Error	**t-statistic**
Mobility	0.0035	0.0579	**0.0605**	0.0132	0.0655	**0.2014**
Pain	0.0654	0.0439	**1.4882**	-0.0212	0.0436	**-0.4863**
Cognition	0.1044	0.0558	**1.8712**	0.0893	0.0579	**1.5428**
Sleep	0.1881	0.0496	**3.7880**	0.0925	0.0485	**1.9070**
Affect	0.0885	0.0561	**1.5773**	0.0496	0.0649	**0.7649**
Breathing	0.0548	0.0557	**0.9844**	0.0428	0.0673	**0.6365**

Table A2: Standard Errors and t-statistics for Simulated Education Differences using a Random Sample of Individuals

IFLS

	(1)	(2)	(3)	(4)	(5)	(6)
	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	**t-statistic**	Education Difference	Standard Error	**t-statistic**
Mobility	0.1139	0.0465	**2.4486**	0.1780	0.0560	**3.1769**
Pain	0.0526	0.0469	**1.1205**	0.1407	0.0567	**2.4808**
Cognition	0.0456	0.0454	**1.0041**	0.1254	0.0480	**2.6104**
Sleep	0.1165	0.0426	**2.7345**	0.1354	0.0450	**3.0114**
Affect	0.0303	0.0502	**0.6031**	0.1017	0.0600	**1.6958**
Breathing	-0.0027	0.0394	**-0.0689**	0.0514	0.0419	**1.2264**

HRS

	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	t-statistic	Education Difference	Standard Error	t-statistic
Mobility	0.1731	0.0256	**6.7649**	0.2510	0.0402	**6.2477**
Pain	0.0836	0.0168	**4.9619**	0.1635	0.0295	**5.5376**
Cognition	0.1083	0.0230	**4.7036**	0.3415	0.0486	**7.0231**
Sleep	0.0429	0.0181	**2.3761**	0.2246	0.0346	**6.4985**
Affect	0.1010	0.0239	**4.2210**	0.1872	0.0517	**3.6222**
Breathing	0.1349	0.0259	**5.2034**	0.2321	0.0499	**4.6502**

ELSA

	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	**t-statistic**	Education Difference	Standard Error	**t-statistic**
Mobility	0.1324	0.0221	**5.9911**	0.0949	0.0346	**2.7435**
Pain	0.0567	0.0188	**3.0145**	0.1012	0.0283	**3.5771**
Cognition	0.1064	0.0221	**4.8121**	0.2175	0.0483	**4.4999**
Sleep	-0.0013	0.0205	**-0.0631**	0.1765	0.0281	**6.2818**
Affect	0.0268	0.0219	**1.2245**	0.1804	0.0465	**3.8831**
Breathing	0.1066	0.0196	**5.4457**	0.1304	0.0477	**2.7331**

CHARLS

	Using Different Thresholds			Using Same Thresholds		
Domain	Education Difference	Standard Error	**t-statistic**	Education Difference	Standard Error	**t-statistic**
Mobility	0.1232	0.0706	**1.7450**	0.0818	0.0763	**1.0721**
Pain	0.1323	0.0485	**2.7306**	0.1498	0.0509	**2.9449**
Cognition	0.1788	0.0674	**2.6534**	0.1676	0.0688	**2.4356**
Sleep	0.1626	0.0525	**3.0989**	0.1949	0.0590	**3.3022**
Affect	0.1687	0.0639	**2.6411**	0.2490	0.0792	**3.1451**
Breathing	0.1210	0.0703	**1.7219**	0.2090	0.0844	**2.4769**

C Appendix Tables and Figures

C.1 Self-Report Distributions

Figures A1 and A2 explore within-country differences across gender and education. Figure A1 depicts the distribution of self-report responses by gender for each dataset 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 are 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, there are significantly different male and female distributions in the cognition, affect, and sleep domains going in the same direction. In the ELSA, the domains that exhibit significant gender differences are pain, sleep, and affect.

Figure A2 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 disproportionally represented in the healthiest categories.

Figure A1: Distribution of Self-Reports by Gender

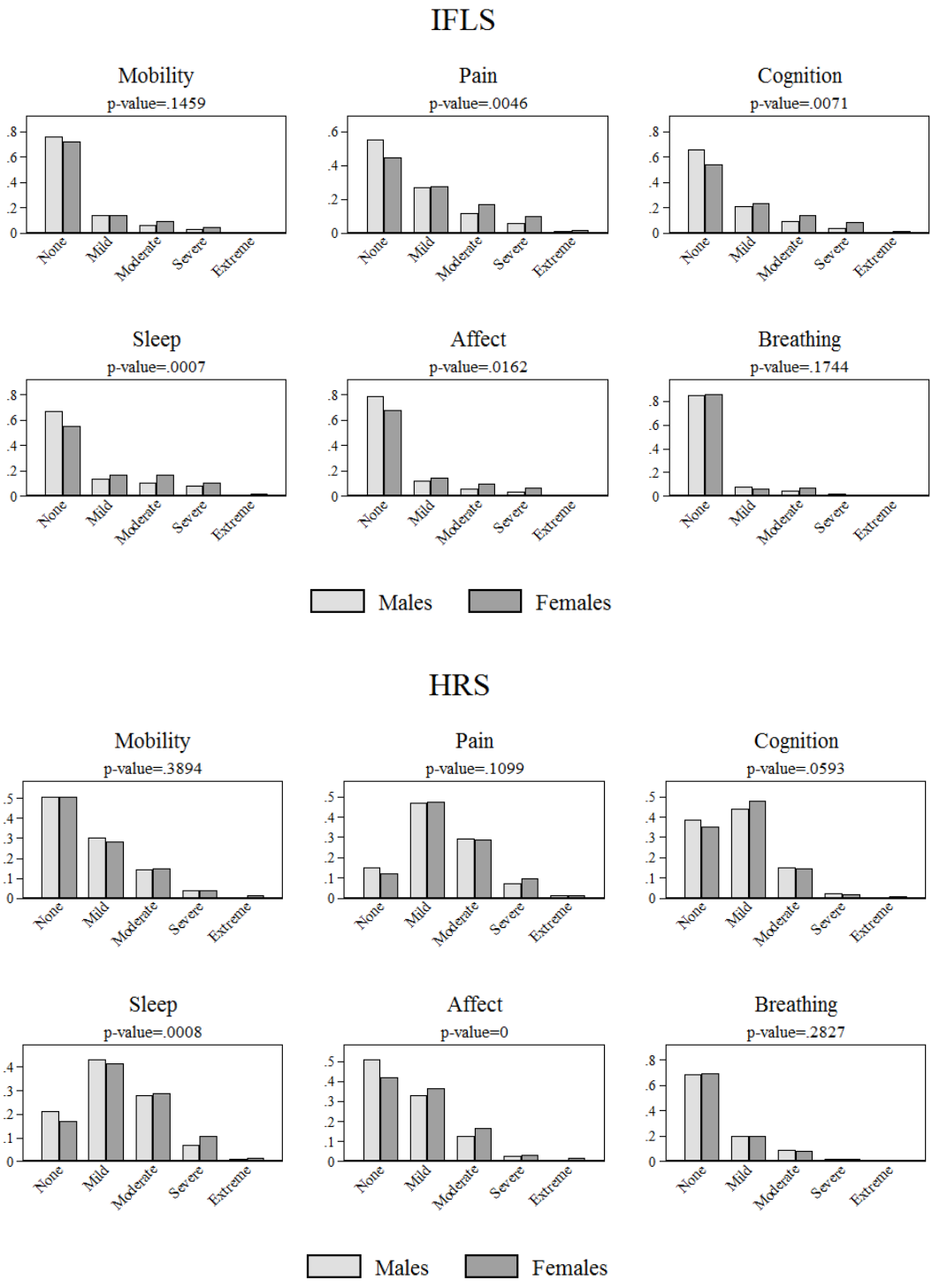


Figure A1: Distribution of Self-Reports by Gender, continued

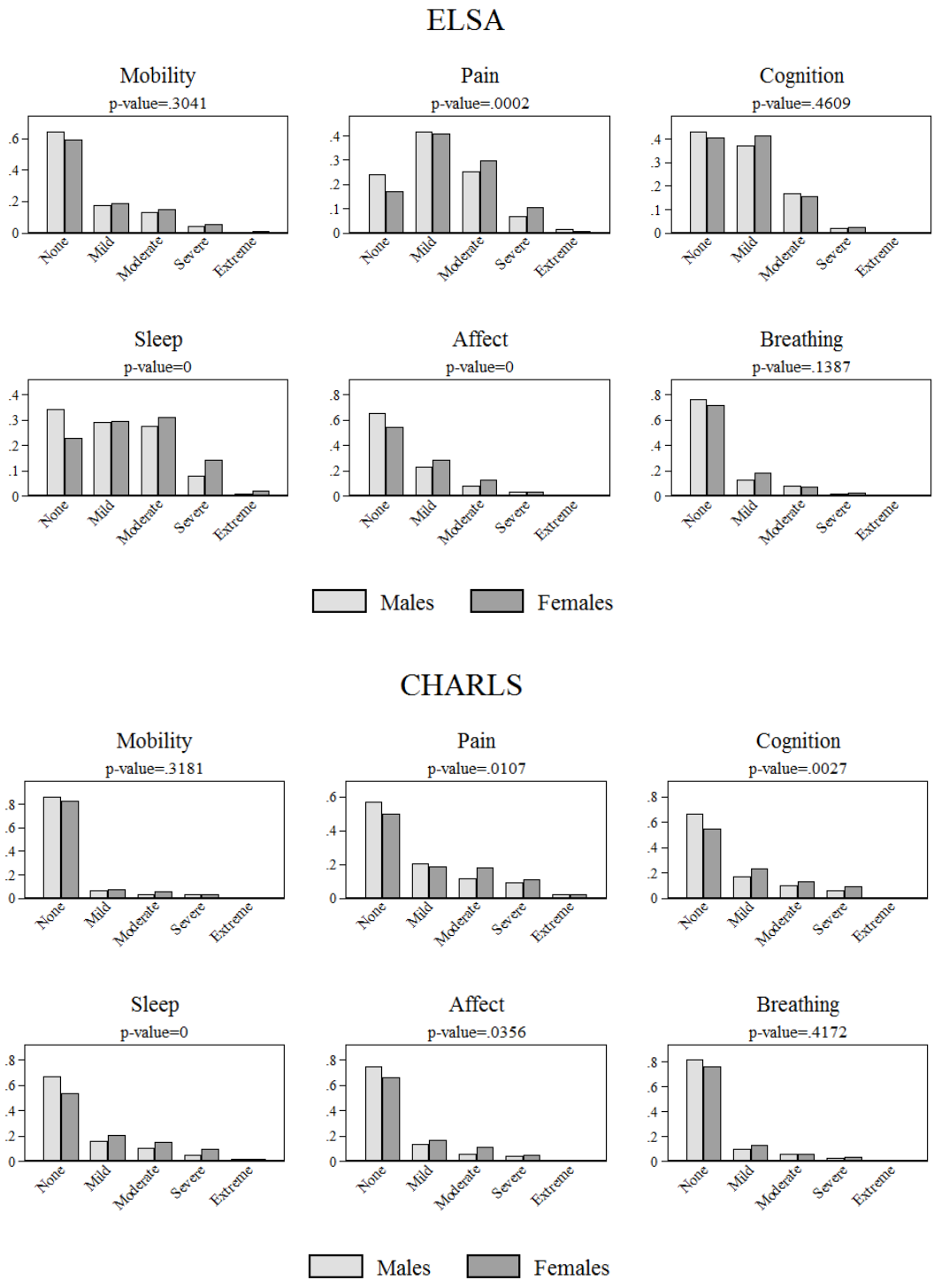


Figure A2: Distribution of Self-Reports by Education

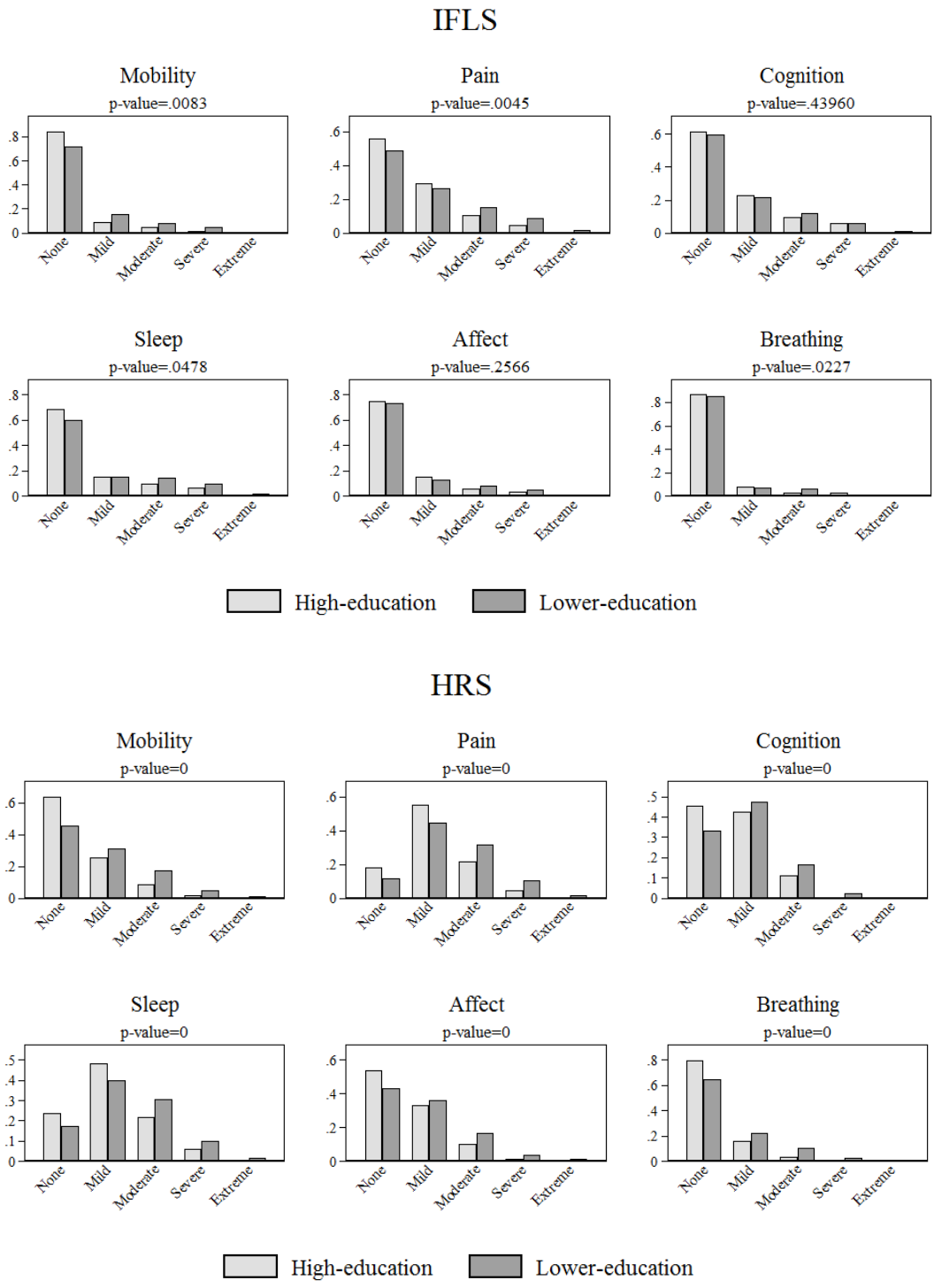
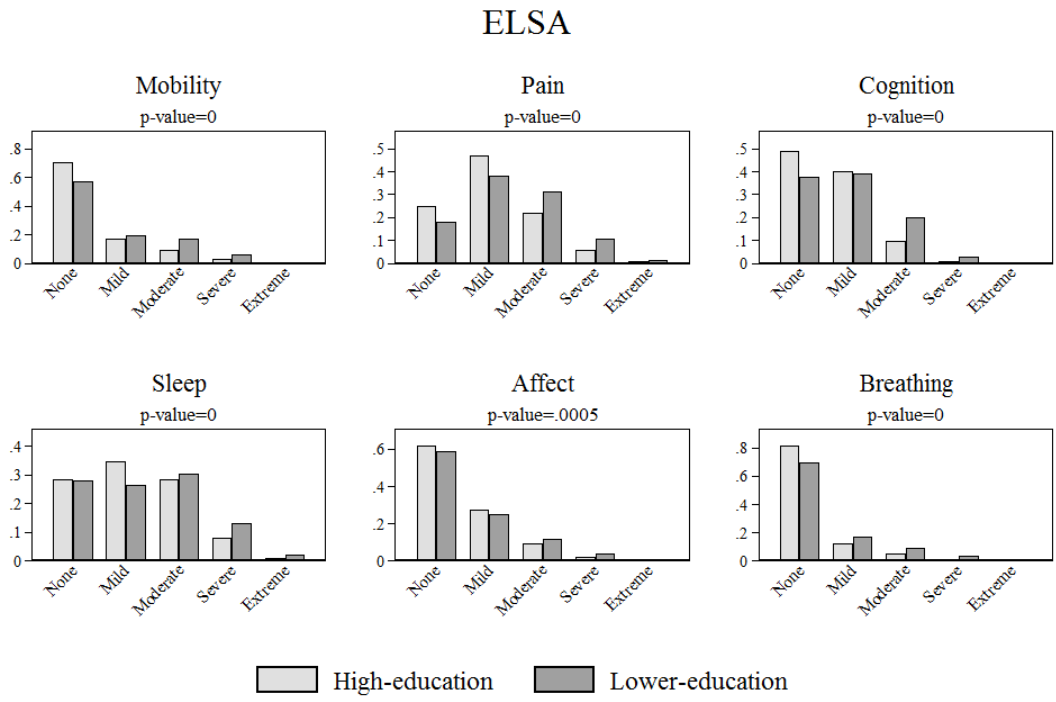
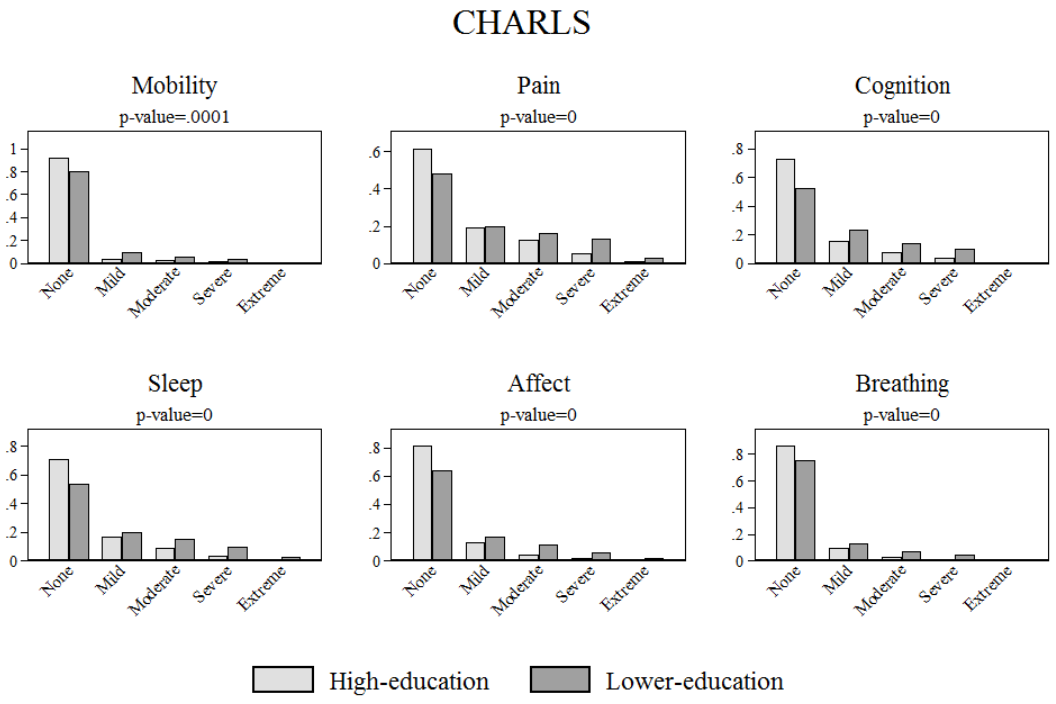


Figure A2: Distribution of Self-Reports by Education, continued





C.2 Ordered Probit and HOPIT Estimation Results

Table A3 reports the coefficients for the HOPIT estimation of the cognition domain in the IFLS, including all threshold equations.

C.3 Robustness to Alternative Functional Form

Table A4 reports the coefficients for the HOPIT estimation of the cognition domain in the IFLS, including all threshold equations. Instead of the exponential function in Equation 3, I estimate the model using a squared term:

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

The coefficients in the latent variable equation and the first threshold equation are almost identical when comparing Table A4 with Table A3, as are the signs and significance levels in the coefficients in the second to fourth threshold equations, alleviating concerns about sensitivity to functional form assumptions. This lack of sensitivity to functional form holds for all domains and datasets.

It is also possible to drop the requirement that $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$ and instead use a linear specification for the threshold equations, as below. Bago d'Uva et al. (2011) use this specification because they find that $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$ is always satisfied.

3b. $\tau_i^0 = -\infty$, $\tau_i^5 = \infty$, $\tau_i^j = \gamma^j X_i + u_i$, j = 1, 2, 3, 4

As Table A5 shows, the latent variable equation coefficients are virtually identical when this linear specification is used instead. The threshold coefficients for $j > 1$ cannot be directly compared because in the exponential and square specifications, these coefficients represent the marginal effect on the difference between two thresholds, while in the linear specification, they represent the marginal effect on the level of one specific threshold.

Table A3: Threshold Equations for Cognition Domain in the IFLS

	HOPIT	Threshold 1	ln(Threshold2 – Threshold1)	ln(Threshold3 – Threshold2)	ln(Threshold4- Threshold3)
1(55 < Age <= 65)	0.106	-0.120	0.130	0.0522	-0.0619
	(0.180)	(0.105)	(0.122)	(0.119)	(0.162)
1(65 < Age <= 75)	0.561**	0.0979	-0.192	-0.149	-0.163
	(0.231)	(0.133)	(0.186)	(0.170)	(0.212)
1(Age > 75)	0.168	-0.372	0.342	0.152	0.130
	(0.482)	(0.333)	(0.324)	(0.297)	(0.387)
1(Male)	-0.136	0.0472	0.00628	-0.0220	-0.0613
	(0.0962)	(0.0542)	(0.0650)	(0.0607)	(0.0866)
1(High Education)	-0.314**	-0.257***	0.201**	0.0871	0.125
	(0.123)	(0.0720)	(0.0841)	(0.0790)	(0.110)
1(Medium Education)	-0.338***	-0.147**	0.105	0.0992	-0.0433
	(0.114)	(0.0623)	(0.0771)	(0.0735)	(0.104)
1(Male)	-0.279	-0.0171	-0.0427	-0.124	0.0703
x 1(55 < Age <= 65)	(0.195)	(0.116)	(0.132)	(0.131)	(0.184)
1(High Education)	0.194	0.0614	-0.00101	-0.206	0.179
x 1(55 < Age <= 65)	(0.275)	(0.170)	(0.185)	(0.187)	(0.228)
1(Medium Education)	0.464**	0.133	-0.00881	-0.0102	0.143
x 1(55 < Age <= 65)	(0.218)	(0.127)	(0.148)	(0.146)	(0.211)
1(Male)	-0.132	-0.161	0.248	0.215	0.243
x 1(65 < Age <= 75)	(0.276)	(0.165)	(0.210)	(0.183)	(0.236)
1(High Education)	-0.357	-0.246	0.274	-0.0147	0.461
x 1(65 < Age <= 75)	(0.402)	(0.269)	(0.270)	(0.260)	(0.319)
1(Medium Education)	0.106	0.0144	-0.333	0.0675	0.0186
x 1(65 < Age <= 75)	(0.298)	(0.175)	(0.246)	(0.200)	(0.263)
1(Male)	-0.253	0.0287	0.0477	0.0188	0.441
x 1(Age > 75)	(0.508)	(0.365)	(0.347)	(0.350)	(0.424)
1(High Education)	-7.054	-1.310	0.747	0.0966	1.134
x 1(Age > 75)	(1361.1)	(17.34)	(7.253)	(0.547)	(5636.7)
1(Medium Education)	0.970*	0.0674	-0.0449	-0.412	-0.0237
x 1(Age > 75)	(0.506)	(0.356)	(0.334)	(0.360)	(0.436)
Constant	-1.126***	-0.995***	-0.472***	-0.544***	-0.537***
	(0.0976)	(0.0657)	(0.0736)	(0.0744)	(0.0956)
Cutoff 1(probit)/	-0.434***				
theta 1 (HOPIT)	(0.0298)				
Cutoff 2(probit)/	-0.209***				
theta 2 (HOPIT)	(0.0250)				
Cutoff 3 (probit)/	0.485***				
sigma v (HOPIT)	(0.0234)				
Cutoff 4 (probit)/	0.438***				
sigma u (HOPIT)	(0.0225)				
Observations	1018				

Notes: t-statistics in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table A4: HOPIT Estimation of Cognition Domain in the IFLS (Using Alternative Functional Form)

	HOPIT	Threshold 1	ln(Threshold2 – Threshold1)	ln(Threshold3 – Threshold2)	ln(Threshold4- Threshold3)
1(55 < Age <= 65)	0.109	-0.111	0.0497	0.0204	-0.0265
	(0.180)	(0.105)	(0.0515)	(0.0466)	(0.0618)
1(65 < Age <= 75)	0.560**	0.0947	-0.0673	-0.0583	-0.0554
	(0.231)	(0.131)	(0.0678)	(0.0618)	(0.0765)
1(Age > 75)	0.169	-0.378	0.147	0.0733	0.0257
	(0.482)	(0.327)	(0.149)	(0.125)	(0.176)
1(Male)	-0.134	0.0520	0.000705	-0.00917	-0.0210
	(0.0962)	(0.0544)	(0.0271)	(0.0236)	(0.0326)
1(High Education)	-0.311**	-0.255***	0.0822**	0.0328	0.0452
	(0.123)	(0.0716)	(0.0351)	(0.0304)	(0.0424)
1(Medium Education)	-0.343***	-0.163***	0.0473	0.0396	-0.0160
	(0.114)	(0.0629)	(0.0314)	(0.0284)	(0.0388)
1(Male)	-0.288	-0.0442	-0.00985	-0.0417	0.0287
x 1(55 < Age <= 65)	(0.195)	(0.115)	(0.0561)	(0.0495)	(0.0701)
1(High Education)	0.194	0.0606	0.00453	-0.0781	0.0767
x 1(55 < Age <= 65)	(0.275)	(0.171)	(0.0818)	(0.0681)	(0.0910)
1(Medium Education)	0.475**	0.164	-0.00994	-0.0119	0.0539
x 1(55 < Age <= 65)	(0.218)	(0.126)	(0.0617)	(0.0555)	(0.0787)
1(Male)	-0.127	-0.159	0.0866	0.0860	0.0759
x 1(65 < Age <= 75)	(0.276)	(0.165)	(0.0826)	(0.0710)	(0.0895)
1(High Education)	-0.352	-0.228	0.122	-0.00819	0.194
x 1(65 < Age <= 75)	(0.401)	(0.269)	(0.126)	(0.101)	(0.140)
1(Medium Education)	0.113	0.0380	-0.133	0.0233	0.00778
x 1(65 < Age <= 75)	(0.298)	(0.174)	(0.0892)	(0.0770)	(0.0963)
1(Male)	-0.249	0.0324	0.0237	-0.0178	0.189
x 1(Age > 75)	(0.508)	(0.363)	(0.165)	(0.136)	(0.187)
1(High Education)	-6.442	-1.217	0.471	0.0566	0.738
x 1(Age > 75)	(382.7)	(23.79)	(7.852)	(0.237)	(766.7)
1(Medium Education)	0.976*	0.0719	-0.0191	-0.161	0.0275
x 1(Age > 75)	(0.506)	(0.356)	(0.160)	(0.136)	(0.191)
Constant	-1.127***	-0.997***	0.790***	0.761***	0.764***
	(0.0976)	(0.0659)	(0.0293)	(0.0284)	(0.0364)
Cutoff 1(probit)/	-0.434***				
theta 1 (HOPIT)	(0.0297)				
Cutoff 2(probit)/	-0.209***				
theta 2 (HOPIT)	(0.0249)				
Cutoff 3 (probit)/	0.483***				
sigma v (HOPIT)	(0.0233)				
Cutoff 4 (probit)/	0.435***				
sigma u (HOPIT)	(0.0222)				
Observations	1018				

Notes: t-statistics in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table A5: HOPIT Estimation of Cognition Domain in the IFLS (Using Linear Functional Form)

	HOPIT	Threshold 1	ln(Threshold2 – Threshold1)	ln(Threshold3 – Threshold2)	ln(Threshold4-Threshold3)
1(55 < Age <= 65)	0.109	-0.111	-0.0307	0.00235	-0.0398
	(0.180)	(0.106)	(0.0890)	(0.0919)	(0.113)
1(65 < Age <= 75)	0.558**	0.0863	-0.00483	-0.0934	-0.168
	(0.231)	(0.131)	(0.117)	(0.121)	(0.144)
1(Age > 75)	0.165	-0.387	-0.133	0.0161	-0.00477
	(0.482)	(0.323)	(0.248)	(0.258)	(0.351)
1(Male)	-0.134	0.0522	0.0524	0.0382	0.0105
	(0.0962)	(0.0543)	(0.0447)	(0.0460)	(0.0592)
1(High Education)	-0.311**	-0.254***	-0.118**	-0.0672	0.00109
	(0.123)	(0.0716)	(0.0578)	(0.0593)	(0.0779)
1(Medium Education)	-0.344***	-0.163***	-0.0858	-0.0243	-0.0475
	(0.114)	(0.0630)	(0.0534)	(0.0554)	(0.0702)
1(Male)	-0.291	-0.0481	-0.0621	-0.121	-0.0758
x 1(55 < Age <= 65)	(0.195)	(0.115)	(0.0936)	(0.0969)	(0.127)
1(High Education)	0.195	0.0616	0.0765	-0.0420	0.0855
x 1(55 < Age <= 65)	(0.275)	(0.171)	(0.130)	(0.132)	(0.173)
1(Medium Education)	0.477**	0.167	0.153	0.126	0.208
x 1(55 < Age <= 65)	(0.218)	(0.126)	(0.104)	(0.109)	(0.143)
1(Male)	-0.121	-0.143	-0.0313	0.107	0.202
x 1(65 < Age <= 75)	(0.276)	(0.166)	(0.141)	(0.141)	(0.170)
1(High Education)	-0.347	-0.218	0.00283	-0.0187	0.311
x 1(65 < Age <= 75)	(0.401)	(0.271)	(0.194)	(0.196)	(0.281)
1(Medium Education)	0.111	0.0380	-0.162	-0.129	-0.115
x 1(65 < Age <= 75)	(0.297)	(0.173)	(0.152)	(0.154)	(0.182)
1(Male)	-0.246	0.0432	0.0920	0.0249	0.386
x 1(Age > 75)	(0.508)	(0.365)	(0.263)	(0.270)	(0.389)
1(High Education)	-6.422	-1.496	0.0109	0.117	1.809
x 1(Age > 75)	(222.7)	(112.7)	(0.443)	(0.485)	(370.0)
1(Medium Education)	0.983*	0.0749	0.0498	-0.220	-0.102
x 1(Age > 75)	(0.506)	(0.359)	(0.262)	(0.272)	(0.381)
Constant	-1.127***	-0.997***	-0.372***	0.208***	0.789***
	(0.0976)	(0.0659)	(0.0496)	(0.0488)	(0.0687)
Cutoff 1(probit)/	-0.434***				
theta 1 (HOPIT)	(0.0297)				
Cutoff 2(probit)/	-0.209***				
theta 2 (HOPIT)	(0.0249)				
Cutoff 3 (probit)/	0.483***				
sigma v (HOPIT)	(0.0233)				
Cutoff 4 (probit)/	0.435***				
sigma u (HOPIT)	(0.0222)				
Observations	1018				

Notes: t-statistics in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

D Anchoring Vignette Questions

These vignette questions were taken from the IFLS, but the HRS, ELSA, and CHARLS data all use the same scenarios except with different names.

D.1 Domain: Mobility

- Pak Taryono/Bu Taryini is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometer. He has no problems with day-to-day activities, such as carrying food from the market

- Pak Tumino/Bu Tumini does not exercise. He cannot climb stairs or do other physical activities because he is obese. He is able to carry the groceries and do some light household work.

- Pak Sidik/Bu Endah has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feel heavy.

D.2 Domain: Pain

- Pak Budiarto/ Bu Budiarti has a headache once a month that is relieved after taking a pill. During the headache she can carry on with her day-to-day affairs.

- Pak Sumarno/ Bu Sumarni has pain that radiates down her right arm and wrist during her day at work. This is slightly relieved in the evenings when she is no longer working on her computer.

- Pak Mulyono/ Bu Mulyanti 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.

D.3 Domain: Cognition

- Pak Taryono/ Bu Taryini 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.

- Pak Suwarso/ Bu Suwarsih is keen to learn new recipes but finds that she often makes mistakes and has to reread several times before she is able to do them properly.

- Pak Mugiono/ Bu Mugianti cannot concentrate for more than 15 minutes and has difficulty paying attention to what is being said to him. Whenever he starts a task, he never manages to finish it and often forgets what he was doing. He is able to learn the names of people he meets.

D.4 Domain: Sleep

- Pak Partono/ Bu Partini falls asleep easily at night, but two nights a week she wakes up in the middle of the night and cannot go back to sleep for the rest of the night.

- Pak Darma/ Bu Darmi wakes up almost once every hour during the night. When he wakes up in the night, it takes around 15 minutes for him to go back to sleep. In the morning he does not feel well-rested.

- Pak Parto/ Bu Parti takes about two hours every night to fall asleep. He wakes up once or twice a night feeling panicked and takes more than one hour to fall asleep again

D.5 Domain: Affect

- Pak Arman/ Bu Lina enjoys her work and social activities and is generally satisfied with her life. She gets depressed every 3 weeks for a day or two and loses interest in what she usually enjoys but is able to carry on with her day-to-day activities.

- Pak Sukarso/ Bu Sukarsih feels nervous and anxious. He worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests him. When he is alone he tends to feel useless and empty.

- Pak Rano/ Bu Rina feels depressed most of the time. She weeps frequently and feels hopeless about the future. She feels that she has become a burden on others and that she would be better dead.

D.6 Domain: Breathing

- Pak Sugiarto/ Bu Suwarsih has no problems while walking slowly. He gets out of breath easily when climbing uphill for 20 meters or a flight of stairs.

- Pak Ramlan/ Bu Badriah suffers from respiratory infections about once every year. He is short of breath 3 or 4 times a week and had to be admitted in hospital twice in the past month with a bad cough that required treatment with antibiotics.

- Pak Hamid/ Bu Karsini has been a heavy smoker for 30 years and wakes up with a cough every morning. He gets short of breath even while resting and does not leave the house anymore. He often needs to be put on oxygen.

References

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., and Van Doorslaer, E. (2011). Slipping anchor? testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources*, 46(4):875–906.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., and O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health economics*, 17(3):351–375.

Case, A. and Paxson, C. (2005). Sex differences in morbidity and mortality. *Demography*, 42(2):189–214.

Cutler, D. M. and Lleras-Muney, A. (2006). Education and health: evaluating theories and evidence. Technical report, National Bureau of Economic Research.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., and Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of general internal medicine*, 21(3):267–275.

Dow, W. H., Gertler, P., Schoeni, R. F., Strauss, J., and Thomas, D. (1997). *Health care prices, health and labor outcomes: Experimental evidence*. RAND.

Dowd, J. B. and 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. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(4):478–489.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., and Baicker, K. (2012). The oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics*.

Gertler, P. and Gruber, J. (2002). Insuring consumption against illness. *American Economic Review*, 92(1):51–70.

Idler, E. L. and Benyamini, Y. (1997). Self-rated health and mortality: a review of twenty-seven community studies. *Journal of health and social behavior*, pages 21–37.

Kapteyn, A., Smith, J., and Soest, A. V. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *The American Economic Review*, 1.

Kapteyn, A., Smith, J. P., and Van Soest, A. (2010). Life satisfaction. *International differences in well-being*, pages 70–104.

King, G., Murray, C. J. L., Salomon, J. a., and Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98(01).

Maccini, S. and Yang, D. (2009). Under the weather: Health, schooling, and economic consequences of early-life rainfall. *The American Economic Review*, pages 1006–1026.

Macintyre, S., Ford, G., and 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(1):89–98.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, a., and Marquis, M. S. (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *The American Economic Review*, 77(3):251–77.

Marmot, M., Oldfield, Z., Clemens, S., Blake, M., Phelps, A., Nazroo, J., Steptoe, A., Rogers, N., and Banks, J. (2014). *English Longitudinal Study of Ageing: Waves 0-6, 1998-2013 [computer file]*. : UK Data Archive [distributor], Colchester, Essex, 21 edition. SN: 5050 , 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.

Mu, R. (2014). Regional disparities in self-reported health: Evidence from chinese older adults. *Health economics*, 23(5):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., and Fox, K. (1993). Gender and life-cycle differentials in the patterns and determinants of adult health. *Journal of Human Resources*.

Strauss, J., Witoelar, F., Sikoki, B., and Wattie, A. (2009). The fourth wave of the indonesian family life survey (ifls4): Overview and field report. Technical report, WR-675/1-NIA/NICHD.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., and 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*.

Verbrugge, L. M. (1989). The twain meet: empirical explanations of sex differences in health and mortality. *Journal of health and social behavior*, pages 282–304.

Vogl, T. (2012). Education and Health in Developing Economies. (December 2012).

Zhao, Y., Strauss, J., Yang, G., Giles, J., Hu, P., Hu, Y., Lei, X., Park, A., Smith, J. P., and Wang, Y. (2013). China health and retirement longitudinal study–2011-2012 national baseline users guide. *Beijing: National School of Development, Peking University*.