HHS Public Access

Author manuscript

Sociol Methodol. Author manuscript; available in PMC 2017 August 01.

Published in final edited form as:

Sociol Methodol. 2016 August ; 46(1): 84–120. doi:10.1177/0081175015599808.

Assessing the Effectiveness of Anchoring Vignettes in Bias Reduction for Socioeconomic Disparities in Self-Rated Health among Chinese Adults

Hongwei Xu and

Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, ISR 2459, Ann Arbor, MI 48104, xuhongw@umich.edu, Phone: (734) 615-3552, Fax: (734) 763-1428

Yu Xie

Department of Sociology, University of Michigan, Center for Social Research, Peking University

Abstract

This study investigates how reporting heterogeneity may bias socioeconomic and demographic disparities in self-rated general health, a widely used health indicator, and how such bias can be adjusted by using new anchoring vignettes designed in the 2012 wave of the China Family Panel Studies (CFPS). We find systematic variation by socio-demographic characteristics in thresholds used by respondents in rating their general health status. Such threshold shifts are often non-parallel in that the effect of a certain group characteristic on the shift is stronger at one level than another. We find that the resulting bias of measuring group differentials in self-rated health can be too substantial to be ignored. We demonstrate that the CFPS anchoring vignettes prove to be an effective survey instrument in obtaining bias-adjusted estimates of health disparities not only for the CFPS sample, but also for an independent sample from the China Health and Retirement Longitudinal Study. Effective adjustment for reporting heterogeneity may require vignette administration only to a small subsample (20–30% of the full sample). Using a single vignette can be as effective as using more in terms of anchoring, but the results are sensitive to the choice of vignette design.

Keywords

anchoring vignettes; reporting heterogeneity; self-rated health; socioeconomic status; China

1. INTRODUCTION

Due to its robust predictive power for mortality (Benjamins et al. 2004; House et al. 2000; Idler and Benyamini 1997), its strong association with morbidity and physical functioning (Goldberg et al. 2001; Singh-Manoux et al. 2006), and the simplicity and low cost associated with its collection, self-rated health has been used as a general health indicator in numerous social surveys (Chen, Yang, and Liu 2010; Tandon, Zhuang, and Chatterji 2006; Wen et al. 2010; Zimmer and Kwong 2004). For the U.S. and European countries, there is a large

literature that consistently documents a positive association between socioeconomic status (SES) and self-rated health (Huisman, Kunst and Mackenbach 2003; Kakwani, Wagstaff and van Doorslaer 1997; Knesebeck et al. 2003; Mirowsky and Ross 2008; Ross and Wu 1995; Willson, Shuey and Elder 2007). What is puzzling, however, is that several studies in many developing countries, including China, Thailand, and the Philippines, have found either no significant positive association between SES and self-rated health or an inverse relationship (Luo and Wen 2002; Pei and Rodriguez 2006; Whyte and Sun 2010; Zimmer and Amornsirisomboon 2001; Zimmer et al. 2000).

Rather than providing evidence of cross-country differences in SES-based health inequalities, these findings may instead reflect reporting heterogeneity – that is, respondents of varying SES backgrounds may adopt systematically different frames of reference in rating their overall health. For example, the peer comparison theory predicts that high-SES respondents are likely to compare themselves to their peers and hence adopt a higher standard for what is considered “excellent” health; whereas low-SES respondents may apply a lower standard, resulting in an inflated level of self-rated health relative to that of high-SES respondents, despite the latter group’s advantage in true health status (Dowd and Todd 2011; Schnittker 2005). This peer comparison behavior yields an underestimated SES gradient. Alternatively, the health optimism/pessimism theory predicts that high-SES respondents, believing their affluence confers well-being, will systematically boost their self-ratings of health (Ferraro 1980), whereas low-SES respondents are more pessimistic about their health in the face of limited resources (Ferraro 1993). In other words, the health optimism/pessimism theory predicts an overestimated SES gradient.

The methodology of anchoring vignettes – brief descriptions of hypothetical people or situations that survey respondents are asked to evaluate on the same scale as they use to assess their own situations – has been proposed to address the problem of cross-group reporting heterogeneity. This approach allows a comparison of the respondents’ self-assessments to the assessments they assign to the hypothetical others on the same questions. Vignettes fix the categorical levels of interest so that variation in responses is adjusted by heterogeneity in thresholds, or cut-points, in respondents’ evaluation of health.

Several studies have reevaluated inter-group health inequalities using the vignettes methodology for self-rated health. However, most of these studies focused on cross-country comparisons (Jürges 2007; Murray et al. 2003; Salomon et al. 2004), and the few that looked at response bias in health inequalities by SES focused mainly on American and European elderly populations, largely due to the availability of vignettes data from the Health and Retirement Study (HRS) and its sibling surveys in Europe (Bago d’Uva, O’Donnell and van Doorslaer 2008; Dowd and Todd 2011; Grol-Prokopczyk, Freese and Hauser 2011). This focus limits the generalizability of results given that the elderly may tend to self-assess their health differently than do younger populations (Schnittker 2005), and that Westerners may respond differently to the vignettes methodology than would respondents in developing countries.

In particular, using anchoring vignettes to assess health may be less effective in China or other non-Western societies for two reasons. First, survey responses in East Asian regions

are characterized by a strong tendency to agree (high acquiescence) and a weak tendency to disagree (low disacquiescence) with any item, regardless of content, and by a strong preference for middle over polar response categories on ratings scales (Harzing 2006). These reporting behaviors reduce the amount of information available for differentiating true health status. Second, vignettes require that respondents evaluate the health of a hypothetical person based on a text description – a cognitive burden that may prove taxing to respondents in developing societies where the average educational attainment is relatively low (Bago d'Uva et al. 2008). These cultural and societal differences may have played a role in the results of a study by Bago d'Uva et al. (2008), which found that anchoring vignettes helped little with bias reduction in terms of measures of health disparities by SES in regional samples in China, India, and Indonesia.

Many health surveys have followed the WHO World Health Survey (WHO-WHS) in collecting multiple domain-specific health ratings (e.g., mobility, pain, cognition, vision, sleep, self-care, and affect) instead of a single general health rating, with each domain using nearly identical anchoring vignettes. This is not a feasible approach in general-purpose household surveys, where, due to time constraints, only a single question about self-rated general health is typically asked. However, to the best of our knowledge, no other surveys have designed vignettes to anchor responses to such a question about self-rated general health.

Given these limitations and gaps in the previous research, this study seeks to address the following questions: (1) Does reporting heterogeneity bias the measurement of health disparities by SES among Chinese adults? (2) Are anchoring vignettes effective in correcting such bias? (3) Can vignette adjustment estimated from a population-based sample be generalized to other surveys? This study reports the application of vignettes we designed to anchor self-rated general health in the China Family Panel Studies. We evaluate the effectiveness of our anchoring vignettes in obtaining more accurate estimates of health disparities by SES in a national sample of Chinese adults and thereby help to reconcile previous findings of an inverse association between SES and health. In addition, we evaluate the cost-effectiveness of the vignettes methodology when several vignettes are administered to a subsample and when only one vignette is administered. We further assess the validity of extrapolating vignette adjustments estimated from our sample to another national sample of middle-aged and older Chinese adults from the China Health and Retirement Longitudinal Study in an effort to demonstrate the broad utility of the anchoring vignettes methodology.

2. VIGNETTES METHODOLOGY

2.1. Inter-Group Reporting Heterogeneity

In considering the vignettes method, it is important to distinguish between adjusting for individual-level and adjusting for group-level reporting heterogeneity. The ubiquitous population heterogeneity in social science research dictates that individual reporting heterogeneity cannot be naively discarded as a mere nuisance or measurement error by assuming reporting behaviors are essentially the same within a subpopulation (Xie 2013). Unfortunately, it is impossible to estimate individual reporting heterogeneity without administering multiple vignettes in full range of the latent construct (latent true health in this

study) to each respondent, whose corresponding vignette assessments also provide enough support for estimating the full-scale individual cut-points from low to high. Not only would such a practice constitute an expensive data collection option in a multi-purpose survey, it would also be extremely challenging to design multiple vignettes that cover the full range of the latent construct and to ensure each respondent's assessments in accordance with the intended vignette ranking. To our knowledge, only one study has estimated individual reporting heterogeneity by pooling 15 vignettes across three different domains and assuming a common response scale across these domains (Kapteyn, Smith and Soest 2007). Similar efforts with fewer vignettes have not been successful (Bago d'Uva et al. 2011a; Bago d'Uva et al. 2011b). Being unable to estimate individual-level reporting heterogeneity, in this paper we essentially follow a conventional practice in empirical social research by focusing on group-level differences (Xie 2013).

Inter-group reporting heterogeneity may assume two patterns on latent response scales: parallel or non-parallel cut-point shift (see Figure 1). For the former, thresholds, or cut-points, shift up or down in parallel for each of the comparison groups, providing evidence that the covariates affect all cut-points equally, and supporting the hypothesis that different groups may simply assume higher or lower thresholds in self-evaluating their health. In the case of non-parallel shift, inter-group differences are seen in unaligned upward or downward cut-point shifts varying with covariates.

2.2. Parametric Model

Identifying parallel or non-parallel cut-point shift among groups or individuals cannot be done using a conventional ordered probit (or logit) model of self-rated health because it requires data such as objective health measures for the latent scale. Anchoring vignettes provide such auxiliary data without the high cost associated with collecting biomarker data for objective health measures. For our analyses, we estimate hierarchical ordered probit (HOPIT) models that draw on anchoring vignettes to purge reporting heterogeneity and attain inter-person comparable self-rated health (King et al. 2004; Tandon et al. 2003). A HOPIT model consists of two parts: a vignette component and a health component.

Since a vignette is a description of a hypothetical person's health status presented to all respondents in the same way, we should expect no systematic variation (apart from random error) in the ratings of the vignette by different respondents, except that they may apply different cut-points, if they perceive the vignette in the same way and on the same unidimensional scale – known as the *vignette equivalence assumption* (King et al. 2004). In other words, respondents' characteristics influence their assessments of health condition of a vignette only through affecting cut-points.

Formally, let $y_{i,j}^{v*}$ denote the continuous latent true health of each vignette as perceived by respondent *i*, and it can be modeled as a linear combination of an intercept α_j and random measurement error $\epsilon_{i,j}^{v}$:

$$y_{i,j}^{v*} = \alpha_j + \epsilon_{i,j}^{v}, \epsilon_{i,j}^{v} \sim N(0,1) \quad (1)$$

with the normalization $\alpha_1 = 0$ for identification. Respondent *i* translates the continuous latent health of vignette *j* into one of K ordered response categories, in this case, poor (=1), fair (=2), good (=3), very good (=4), and excellent (=5), through a mapping mechanism:

$$y_{i,j}^{v}=k, \text{if} \tau_i^{v,k-1} \leq y_{i,j}^{v*}<\tau_i^{v,k}, k=1,\ldots,5 \quad (2)$$

where $\tau_i^{v,k}$ denotes the cut-point for respondent *i* to rate the latent true health status of the vignettes as in one of the *K* categories; and $\tau_i^{v,0}<\tau_i^{v,1}<\tau_i^{v,2}<\ldots<\tau_i^{v,5}, \tau_i^{v,0}=-\infty$, and $\tau_i^{v,5}=\infty$. Unlike a conventional ordered probit model that assumes no reporting heterogeneity, and hence homogeneous cut-points, we allow the cut-points to vary as a linear function of covariates X_i, plus individual heterogeneity $u_i^{v,k}$:

$$\tau_i^{v,k}=\gamma_0^{v,k}+X_i\gamma^{v,k}+u_i^{v,k}, k=1\ldots,4 \quad (3)$$

where $\gamma_0^{v,k}$ are the intercepts in the respective cut-points for the vignettes and hence X_i does not include a constant. As mentioned earlier, identification of $u_i^{v,k}$ requires rich data from multiple vignettes that capture the full range of latent health, which are not available to us. We therefore follow the prevailing practice in the literature by restricting our attention to identifying group-specific cut-points. Reporting homogeneity results from imposing $\boldsymbol{\gamma}^{v,k} = 0$. Parallel cut-point shift arises when $\boldsymbol{\gamma}^{v,k} = \boldsymbol{\gamma}^{v}$ for k = 1, …, 4; that is, the impact of a covariate on shifting the cut-point location is the same for all the cut-points. Conversely, $\boldsymbol{\gamma}^{v,k} \neq \boldsymbol{\gamma}^{v}$ is the situation of non-parallel shift.

The self-rated health component takes a similar form as that of the vignette component. Let y_i^{s*} denote the continuous latent true health variable for respondent *i*. We will model it as a linear combination of the SES variables and other control variables, denoted together by X_i, and an independent normal error term $\boldsymbol{\varepsilon}_i$:

$$y_i^{s*}=\beta_0+X_i\beta+\epsilon_i, \epsilon_i \sim N(0,\sigma^2) \quad (4)$$

where β_0 is the intercept. The measurement model divides y_i^{s*} into *K* ordinal response categories of self-rated health y_i^{s} through a similar mapping mechanism as Equation (2):

$$y_i^{s}=k, \text{if} \tau_i^{s,k-1} \leq y_i^{s*}<\tau_i^{s,k}, k=1,\ldots,5 \quad (5)$$

where $\tau_i^{s,k}$ denotes the cut-point for respondent *i* to report his/her health status as in one of the *K* categories; and $\tau_i^{s,0}<\tau_i^{s,1}<\tau_i^{s,2}<\cdots<\tau_i^{s,5}$, $\tau_i^{s,0}=-\infty$, and $\tau_i^{s,5}=\infty$. Again, we allow the cut-points for self-rated health to vary as a linear function of observed covariates Z_i, plus individual heterogeneity $u_i^{s,k}$:

$$\tau_{\mathrm{i}}^{\mathrm{s,k}}=\gamma_{0}^{\mathrm{s,k}}+\mathrm{Z_i}\gamma^{\mathrm{s,k}}+u_{i}^{s,k},\mathrm{k}=1,\ldots,4 \quad (6)$$

where γ_0^{k} are the intercepts in the respective cut-points, and $\mathrm{Z_i}$ can include the same covariates as $\mathrm{X_i}$. We again choose not to identify individual reporting heterogeneity here for practical data limitation. These equations define the second component of a HOPIT model. However, without the auxiliary information provided by the vignettes, the above model is under-identified in that we cannot simultaneously estimate β (the effects of SES and other covariates on self-rated health), γ^{s} (the effects of SES and other covariates on cut-points in response styles), and σ^2. Model identification is achieved by assuming *response consistency* (King et al. 2004), meaning that respondents rate their own health in the same way as they assess all the hypothetical scenarios represented by the vignettes. Formally, the response consistency assumption amounts to setting:

$$\tau_{\mathrm{i}}^{\mathrm{s,k}}=\tau_{\mathrm{i}}^{\mathrm{v,k}} \quad (7)$$

In other words, the vignette component and the self-rated health component are linked through shared cut-points in survey reporting. The group-specific cut-points are estimated from the vignettes data, provided that the response consistency assumption holds, thereby purging out reporting heterogeneity in estimating group differences in self-rated health.

Let $\mathrm{P_{i,k}^{s}}$ denote the probability of respondent *i* reporting his/her own health as in category *k*, and $\mathrm{P_{i,j,k}^{v}}$ denote the probability of the same respondent rating vignette *j* as in category *k*. The log-likelihood function of the HOPIT model in this case (two vignettes and self-rated health, all on five response categories of health status) is defined as the sum of two components, respondent's self-rated health and his/her ratings of the vignettes:

$$\mathrm{lnL}=\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{k=1}}^{5}\mathrm{I}(\mathrm{y_i^s}+\mathrm{k})\mathrm{lnP_{i,k}^{s}}+\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{j=1}}^{2}\sum_{\mathrm{k=1}}^{5}\mathrm{I}(\mathrm{y_{i,j}^{v}}=\mathrm{k})\mathrm{lnP_{i,j,k}^{v}} \quad (8)$$

where $\mathrm{I}(\mathrm{y_j^s}=\mathrm{k})$ and $\mathrm{I}(\mathrm{y_{i,j}^v}=\mathrm{k})$ are two indicator functions that equal 1 if respectively $\mathrm{y_i^s}=\mathrm{k}$ or $\mathrm{y_{i,j}^v}=\mathrm{k}$; and equal 0 otherwise. Parameter estimates can be attained by maximizing this joint log-likelihood.

2.3. Cost-Effectiveness of the Vignettes Method

The degree of cost-effectiveness of the vignettes methodology – which is based on reducing response bias without having to collect objective health measures – hinges on implementation choices. First, cost-effectiveness can be enhanced by administering multiple vignettes to only a small subsample of respondents, from which the anchored group-level response scaling patterns can be generalized to the entire study sample (King et al. 2004), although individual-level adjustment remains intractable. For example, in the Survey of Health, Ageing and Retirement in Europe (SHARE), vignettes data were collected only in

about 10–16% of the full samples, and in the WHO Multi-Country Survey Study on Health and Responsiveness 2000–2001 (WHO-MCSS), about 25–50% of the full samples were randomly administered the vignette instrument (Bago d'Uva et al. 2008). Since most analyses have used data from the small subsamples with the multiple vignettes (often five or more), it is unclear to what extent vignette adjustment seen in these subsamples can be applied to the rest of the sample, although in principle administering vignettes to the full sample only improves statistical efficiency.

Second, it may be possible to enhance the method's cost-effectiveness while maintaining its capacity to identify and correct for group-level reporting heterogeneity by using one rather than multiple vignettes, as long as there is enough within-group variation for all response categories. Using more than one vignette may add only to statistical efficiency while increasing survey development and implementation costs, as well as respondent burden. The literature does not address this issue – and little empirical evidence supports the effectiveness of bias reduction with only one vignette –but we note reductions in the number of vignettes used per self-assessment in some recent surveys. In SHARE, for example, the number of vignette questions for each health domain was reduced from three in the first wave to one in the second wave (Peracchi and Rossetti 2013; Voňková and Hullegie 2011).

We investigate the cost-effectiveness for these two variations on the vignettes method and report the findings after the main results. Assuming comparable group-level reporting behaviors across population-based samples, we further use vignette results from one survey to correct for reporting heterogeneity in another contemporaneous survey. Such cross-sample adjustment would mean a substantial reduction in data collection cost since borrowing existing vignette results from an external source would cost much less than collecting new data.

3. EMPIRICAL DATA AND MEASURES

3.1. China Family Panel Studies

The primary data source for this study comes from the China Family Panel Studies (CFPS), a nationally representative longitudinal survey of Chinese communities, families, and individuals. The studies focus on the well-being of the Chinese population, with a wealth of information on economic activities, education outcomes, family dynamics and relationships, and health. The CFPS tracks all members of the sampled families in the 2010 baseline through biennial follow-up surveys. The first of these, in 2012, used both in-person interviews and proxy-reports administered via computer-assisted personal interviewing (CAPI) or computer-assisted telephone interviewing (CATI) to collect follow-up data.

The 2010 nation-wide CFPS baseline survey successfully interviewed 14,960 households from 635 communities, including 33,600 adults and 8,990 children, located in 25 designated provinces. The approximate response rate was 81.3% at the household level and 84.1% at the individual level, with the majority of the non-response due to non-contact. CFPS's stratified multi-stage sampling strategy ensures that the sample represents 95% of the total population in China in 2010 (Xie 2012; Xie and Hu 2014). The first full-scale follow-up survey was conducted in 2012 with more than 80.6% of the baseline respondents re-

interviewed. This study relies on the data from the 2012 follow-up survey, for which we designed our own anchoring vignettes for self-rated health that were administered to all the adult respondents.

Self-rated health and vignettes—The dependent variable in this study is self-rated health, collected by asking respondents to rate their overall health status at the time of interview by selecting one of five categories: poor, fair, good, very good, or excellent. Every respondent who rated his/her own health was then administered the following two vignettes in random order, on the same response scale, about the health status of a hypothetical person with a common Chinese male or female name matched to the respondent's sex. The health vignettes were designed to reflect two substantially different health statuses, thereby providing greater power to differentiate the varying cut-points applied by respondents to assessing their own health status.

Vignette (1): Sun Jun (male) / Li Mei (female) has no problem with walking, running, or moving his/her limbs. He/she jogs 5 km twice a week. He/she does not remember the last time when he/she felt sore, which was not within the past year. He/she never feels sore after physical labor or exercise. How would you rate his/her health status?

Vignette (2): Zhao Gang (male) / Wang Li (female) has no problem walking 200 meters. He/she feels tired, however, after walking 1 km or climbing several flights of stairs. He/she has no problem with daily activities such as bringing home vegetables from market. He/she has a headache once every month, but gets better after taking medicine. Even while feeling the headache, he/she can still do daily work. How would you rate his/her health status?

SES indicators—Education is measured in years of schooling. Economic resources are measured by employment status and family income per capita. We chose not to use individual income because many Chinese households, especially in rural areas, act as single economic entities. Political capital is measured by one's own as well as other family members' cadre and/or party membership.

In addition to these conventional SES indicators, we also include a measure of cognitive functioning, known as episodic memory, as part of general fluid intelligence. Cognitive ability not only affects respondents' mental comprehension and assessment of the hypothetical vignette scenario but also correlates with a variety of educational and labor market outcomes (Herrnstein and Murray 1996; Marks 2013). Like respondents in other Chinese household surveys (e.g., the China Health and Retirement Longitudinal Study and the China Health and Nutrition Study), the CFPS-2012 respondents were read a randomly selected list of 10 simple nouns, then immediately asked to recall as many of those words as possible in any order. After 31 questions concerning subjective wellbeing, the respondents were again asked to recall as many of the original words as possible. Following the literature, (Hu et al. 2012; McArdle, Smith and Willis 2011), we calculated the final score of episodic memory by averaging the number of successes between the immediate and delayed word recalls.

We control for socio-demographic variables, including age, gender, marital status, rural-urban residence, *Hukou* (household registration) status, and region of origin. Age is centered at mean and divided by 10 to facilitate the interpretation of the parameter. We also add an age-squared term in regression models to capture potential nonlinearity in age trajectory of health. All the other control variables are discrete in nature and entered into regression models as dummies.

We focus on adults aged 16–70 years old in 2012 (N = 30,774), excluding about 4% of this sample who had missing data on self-rated health or at least one of the two vignettes, and about 15% of the remaining sample who gave ratings inconsistent with the designed rank ordering of the two vignettes, and thereby were in violation of the vignette equivalence assumption underlying the methodology (King et al. 2004). As a group, this 15% of respondents had significantly lower SES (e.g., lower educational attainment, worse memory, and lower income) and reported poorer health compared to those whose ratings of the vignettes were consistent with the survey design (results not shown). Therefore, our results may underestimate the true SES disparities in health. After excluding these respondents, the sample size was 25,141, and was further reduced to 23,207 after list-wise deletion of cases with missing data on covariates.

3.2. China Health and Retirement Longitudinal Study

To demonstrate the utility and external validity of our newly designed health vignettes for the CFPS-2012, we apply the estimated cut-point shifts to estimate health disparities in the China Health and Retirement Longitudinal Study (CHARLS), a nationally representative longitudinal survey of adults aged 45 and older and their spouses if available. The CHARLS national baseline survey was launched in 2011 and interviewed 17,708 respondents with a response rate of 80.5% (Zhao et al. 2014). Unlike the CFPS, the CHARLS did not interview all the members in a sampled household. For comparison, we constructed most of the same SES indicators and control variables in the CHARLS as those in the CFPS, except for cadre status and/or party membership. Only a random subsample of 8,712 CHARLS respondents were asked to rate their general health using the same response categories as those in the CFPS-2012, and 7,129 of them were 45–70 years old, within the age range of our CFPS analytical sample. After excluding respondents with missing data on covariates, we have 5,928 cases from the CHARLS sample.

4. MAIN RESULTS

4.1. Descriptive Statistics

Table 1 presents frequency distributions of self-rated health and vignette ratings in the CFPS and CHARLS samples. In the CFPS sample, the responses to self-assessment were more or less evenly distributed with about one third of the respondents considering themselves in fair or poor health, another third in good health, and the rest in very good or excellent health. As expected given the vignette design, the majority of respondents rated the person in the first vignette as in very good or excellent health and the person in the second vignette as in poor health. In the CHARLS sample, the entire distribution of self-rated health was shifted downward as the respondents were on average much older than those in the CFPS. Less than

10% of the sample considered themselves in very good or excellent health, whereas nearly three quarters considered themselves in fair or poor health.

Table 2 summarizes the descriptive statistics of the independent variables in the CFPS and CHARLS samples. Our CFPS analytical sample is evenly split between men and women, with an average age of about 43 years. Over 80% of the respondents were married, which is consistent with the nearly universal marriage pattern in China. The average for years of schooling was 7.6, and on average, respondents recalled about four of the ten words in the episodic memory test. Nearly two thirds of the sample was employed, with an average annual family income per capita of 14,490 RMB (about $2,415 US), more than six times above the new poverty line in rural China (2,300 RMB, see Zhang et al. 2012). About 7.7% of the respondents were members of the Communist Party of China (CPC) and/or cadres of various government agencies and public institutes, and 13.8% had at least one family member who was a CPC member or cadre. In terms of residential and migration status, just over half of the sample consisted of rural non-migrants or rural-to-rural migrants (hereafter referred to collectively as rural residents); 18.7% migrated from rural to urban areas; less than 5% were urban-to-rural migrants; and about 25% were urban non-migrants or urban-to-urban migrants (hereafter referred to as urban residents).

Our CHARLS analytical sample is unsurprisingly older than the CFPS sample but still roughly gender-balanced. Being older, the CHARLS sample had on average higher rates of marriage and widowhood, lower educational attainment, worse memory, a slightly higher employment rate but lower income, and lower percentages of cadres or CPC members relative to the CFPS sample. The CHARLS sample had a similar distribution of rural-urban residence and *hukou* status to that in the CFPS sample, but a less even distribution across regions.

4.2. Reporting Heterogeneity

We assess parallel versus non-parallel cut-point shift by estimating two nested models and performing Wald tests against parallel shift. The results are reported in Table 3. Bear in mind that, generally speaking, lower (downward shift) and higher (upward shift) cut-points would deflate and inflate group differentials in health, respectively, without vignette adjustment. Assuming parallel shift, cut-points would decline with increases in respondent age ($\beta = -0.019$), and the rate of decline would increase with age given the significant negative coefficient of the age-squared term. In other words, older respondents applied significantly lower cut-points in rating and therefore were more likely to report better health for a given level of true latent health compared to younger respondents. Men applied significantly higher cut-points ($\beta = 0.103$) and hence tended to underrate the same level of true health compared to women. Compared to being married or cohabiting, being single was associated with lower cut-points. Better educated respondents had higher cut-points ($\beta = 0.009$), whereas those with better episodic memory had lower cut-points ($\beta = -0.022$). The relationship between family income and cut-point shift was non-linear in that those in the third quartile tended to have significantly higher cut-points compared to the poorest, although the richest also had a significantly higher cut-point between good and very good health. Being a cadre or CPC member was related to downward shifted cut-points ($\beta =$

−0.083), although other family members' political status did not matter. Compared to rural residents, rural-to-urban migrants had higher and urban residents had lower cut-points.

The Wald tests provide statistical evidence in favor of non-parallel cut-point shift for most of the aforementioned covariates except family income. In other words, different social groups may not simply have higher or lower thresholds for health evaluation; instead, they exhibit greater reporting heterogeneity at some levels of health than others. For example, a higher level of education was associated with an upward cut-point shift at the higher end of health but a downward shift at the lower end when the assumption of parallel shift was relaxed. To gain a better understanding of this complex pattern, Figure 2 plots predicted cut-points for five different levels of education and varying migration status, respectively, holding everything else constant. It is clear that better educated respondents tended to apply lower cut-points when considering what constitutes poor health. The cut-point between poor and fair was roughly −2.43 for college graduates as opposed to −2.27 for those without any schooling. However, the gradient reversed at the high end of health rating: for college graduates and the unschooled, respectively, the cut-point between good and very good was approximately −0.2 and −0.64, and between very good and excellent was 0.77 and 0.42. As a result, for a given level of true health, better educated respondents would be much less likely than respondents with no schooling to report very good or excellent health. With respect to migration status, the pattern is less clear, but two findings stand out. First, urban natives had the lowest cut-points at the low end of the health distribution, and thus were more likely to report poor or fair health than the other subgroups when they were indeed in poor health. Second, urban natives did not retain a similar high standard at the high end of health rating. Instead, it was the rural-to-urban migrants who held the highest threshold for what constitutes excellent health.

4.3. Bias Reduction

To evaluate the performance of vignettes methodology in remedying reporting heterogeneity, we compare group differences in self-rated health as estimated from three models, a standard ordered probit model, a HOPIT model assuming parallel cut-point shift, and a HOPIT model assuming non-parallel shift. Because of different scaling in these models,[1] we fixed the scale of the HOPIT models by dividing the estimated coefficients by the estimated variance terms, which is equivalent to imposing the same variance as in the ordered probit model (Jones et al. 2007). Table 4 presents the comparable coefficient estimates after rescaling and suggests several related patterns.

First, anchoring vignettes did affect the estimates of health disparities by socioeconomic and demographic groups as demonstrated by the changes in coefficients between the ordered probit and HOPIT models for every covariate that induced cut-point shift (as shown in Table 3).

Second, the magnitude of some of these changes was substantial. For example, the coefficient for years of education nearly tripled from 0.004 to 0.011 after vignette

[1]The scale in the standard ordered probit model is normalized to 1 (i.e., the error term is assumed to follow a standard normal distribution), while it is estimated in HOPIT models (i.e., σ^2 in Equation (4)).

adjustment (assuming non-parallel cut-point shift), whereas the coefficient for the episodic memory dropped by half from 0.036 to about 0.017. More strikingly, certain coefficients that were not significant in the ordered probit model became significant in the HOPIT models. For example, none of the coefficients for family income quartiles was significant in the standard ordered probit model. But estimates from both HOPIT models (parallel and non-parallel shift) indicated that respondents in the top two quartiles of family income reported significantly better health than those in the bottom quartile. This finding is consistent with the conventional wisdom of positive SES gradients in health as well as the positive association between family income and cut-point shift reported in Table 3. It is also noteworthy that the size of the coefficient associated with family income nearly doubled, from about 0.03~0.04 to 0.06~0.07, after vignette adjustment. For other covariates such as divorce and widowhood, one's own cadre, and CPC membership, significant differences disappeared after vignette adjustment.

Third, the assumption of parallel or non-parallel cut-point shift exerted limited impact on estimating the self-rated health component as evidenced by the very small size/sign changes in the coefficients between the two specifications. Nevertheless, the model specification assuming non-parallel shift revealed a more complex pattern of reporting heterogeneity with respect to many covariates, as suggested by the significant Wald tests. In practice, we recommend a step-wise and iterative model-building strategy by exploring parallel cut-point shift first and then allowing all the cut-points to vary freely. There are advantages to applying different modelling specifications. If a statistical test between the two nested models leads us to prefer the parallel shift model, we have a parsimonious model and can easily interpret the results. If the statistical test rejects the parallel shift model, we may still constrain certain cut-points to be constant or parallel shift while retaining others to be non-parallel to obtain a more parsimonious model yet with sufficient explanatory power.

To further gauge the amount of reporting bias reduction achieved by using vignettes, we carried out a simple counterfactual exercise as employed in prior research (Bago d'Uva, O'Donnell and van Doorslaer 2008). Specifically, we first fixed the latent health status for a reference person,[2] and then predicted the probability of reporting very good or excellent health with varying cut-points as would be adopted by people with different characteristics while holding everything else constant. We computed the ratio of probabilities (relative probability) with any two different sets of cut-points to measure the relative magnitude of the reporting effect. To preserve space, we focus on the effects of education and migration here. Figure 3 plots the relative probabilities of reporting very good and excellent health when using the cut-points of different levels of education and migration status, respectively. The denominator in case of education, held constant, is the predicted probability of reporting very good or excellent health when using the cut-points of no schooling, while the numerators are calculated in the same way but with cut-points shifting from primary schooling to college. Again, the effect of reporting heterogeneity was quite large. The relative probability of reporting very good health dropped from 0.85 to 0.62 and for

[2]The reference person is a married man of the sample average age, with 9 years of schooling (junior high school) and an episodic memory test score of 4 (rounded up the sample mean), employed as a rural non-migrant, and living in the poorest family income quartile.

reporting excellent health dropped from 0.77 to 0.48 as the associated cut-points shifted from those of primary school to those of college education. This means that, given the same latent health for any respondent, the probability of giving an excellent health self-rating with the cut-points of college education imposed would be less than half the probability if applying the cut-points of no schooling (the denominator of the relative probability). For migration status, the denominator refers to rural natives. The gradient is more evident with regard to the rating of excellent health than that of very good health. The relative probability of reporting excellent health decreased from 0.97 to 0.82 when we replaced the corresponding cut-points of urban natives with those of rural-to-urban migrants.

In light of China's longstanding rural-urban divide and likely rural-urban difference in health-related reporting patterns, we analyzed the rural and urban CFPS-2012 respondents separately. The main results are reported in Table 5. Several findings are worth of highlighting. First, vignette adjustment was more effective in uncovering the education gradient in health in the rural subsample, as the unadjusted coefficient of schooling was not statistically significant. Second, variation in self-rated health by episodic memory was mainly driven by reporting heterogeneity among urban residents, since the coefficient of episodic memory became insignificant after vignette adjustment. Third, the income gradient in self-rated health seems to be mainly a rural phenomenon.

5. COST-EFFECTIVE ANALYSIS

5.1. Administering Vignettes to Subsample

Identification of group-level reporting heterogeneity rests on the assumption of significant within-group similarity, or group-specific reporting patterns, apart from additional within-group individual variation. This assumption implies that group-specific cut-points estimated from a random subsample (or even an external sample from the same population) can be applied to the full sample. Of course, estimating cut-points from the full sample is better because it has more statistical efficiency. However estimating group-specific cut-points with a subsample has significant practical implications, as it substantially reduces survey costs and respondent burden. We therefore proceed to perform cross-validation to assess the degree to which vignettes administered to a small subsample can assist in bias reduction for the entire sample. Our cross-validation procedure hinges on a unique feature of the data used in this study. That is, unlike in other large-scale social surveys, CFPS vignettes data were collected on the same respondents who were administered self-assessments, producing a large sample that ensures enough statistical power for cross-validation. Specifically, we randomly partition the full sample into a relatively small subsample as training data and the remaining larger subsample as validation data, since prior studies indicate that cost-effectiveness is achieved by administering vignettes to a subsample that is 10% to 50% the size of the overall sample. We experiment with a series of partitions, including 10%, 20%, 30%, 40%, and 50%, for each of which we repeat the random partition 500 times. After each partition, we fit a HOPIT model to the training data and compute out-of-sample predictive cut-points and latent health status. We then fit another HOPIT model to the larger subsample validation data, and compute in-sample predictive cut-points and latent health status. Closeness between the two sets of predictive values, measured by mean-square error,

indicates external validity and thereby the cost-effectiveness of extrapolating vignette adjustment obtained from a small subsample to the full sample.

How big does the subsample of those administered vignettes have to be in order to make reasonably good extrapolation of adjustment for reporting heterogeneity? Our cross-validation analyses suggest that a surprisingly small sample would be sufficient. Figure 4 plots the distributions of mean-squared errors between the out-of-sample predictions of latent health for the validation subsample based on the model fitted to the training subsample and the in-sample predictions based on the model fitted to the validation subsample. When using both vignettes available in the CFPS, the mean-squared errors take the form of exponential decay as the size of training data increases. The decay rate is greater at the lower end — the largest decline in mean-squared errors occurs as the proportion of the full sample used as training data increases from 10% to 20%. The trend of decline flattens out beyond 30%. As shown in Figure 5, the same pattern holds for the mean-squared errors between the out-of-sample predictions of cut-points for the validation subsample based on the model fitted to the training subsample and the in-sample predictions based on the model fitted to the validation subsample.

5.2. Number of Vignettes

Would one vignette be sufficient to anchor reporting behaviors? If so, does it make a difference which single vignette is used? In principle, one vignette is sufficient for identifying group-level differences, provided it yields sufficient variation in the vignette ratings, or full support, which enables estimation of the full range of cut-points. Adding more vignettes would then improve the estimation efficiency. As shown in Table 1, however, the assumption of full support is not satisfied in the CFPS data because neither the first nor the second vignette yielded responses in all categories – that is, the rating of poor for the first vignette and excellent for the second vignette received zero responses. This means we have no statistical power to identify the cut-point at the low end if using the first vignette only, or that at the high end if using the second vignette only. Therefore, we expect that using both vignettes complementally is the best solution in this particular scenario.

To demonstrate this, we repeat the HOPIT model estimation by using one vignette at a time to ascertain whether it can attain similar bias reduction to using two vignettes. We not only compare coefficient estimates, but also examine whether different vignettes lead to similar adjusted self-rated health (Voňková and Hullegie 2011). Since data using two vignettes are collected for anchoring health in the CFPS data, we should expect similar adjusted self-rated health when using either one of the vignettes or both, provided that both vignettes are equally effective in anchoring response patterns. We then compute pair-wise correlation coefficients among the three sets of vignette-anchored self-rated health data (two sets using one vignette only, and the third set using both vignettes). A correlation coefficient close to 1 indicates a similar adjustment when using different sets. We also repeat the above cross-validation procedure using one vignette only to determine whether it is valid to extrapolate a subsample anchoring to the full sample by using a single vignette.

First, we compare coefficient estimates of the associations of covariates with self-rated health anchored by using different vignettes. Figure 6 plots the point estimates and the

associated 95% confidence intervals. It is notable that the point estimates when using the second (worse health) vignette are generally bigger in terms of absolute values than the point estimates when using the first (better health) vignette, while the coefficient sizes when using both vignettes fall in between, reflecting a result of smoothing. In most cases, the 95% confidence intervals are overlapped for the same covariate, indicating insufficient statistical power to distinguish estimates using different vignettes. However, substantive variation does occur with certain important SES indicators. For example, the 95% confidence interval for education covers 0 when using the first vignette only, but not so when using either the second only or both vignettes. Similar patterns can be observed for episodic memory, top income quartile, and family members' cadre or party membership. To the extent that we expect significant SES disparities in health, it is likely that the second (worse health) vignette is relatively more effective than the first vignette in anchoring reporting behaviors.

Second, as shown in Table 6, correlation coefficients range between 0.97 and 0.99 for predicted latent health and between 0.87 and 0.92 for predicted ordinal health ratings when different vignettes are used. These large positive correlation coefficients lend further support to the robustness of vignette adjustment in our CFPS design, which is not achieved in other surveys (Voňková and Hullegie 2011).

Can we make valid inferences about the reporting behaviors in the full sample by administering a single vignette to only a subsample? Our cross-validation analyses reveal a positive answer. As shown in Figure 4, the same pattern of exponential decay in mean-squared errors for predicted latent health when using both vignettes holds for using either one of the two vignettes. The mean-squared errors experience a substantial decline when the proportion of the full sample used as training data increases from 10% to 20%, and the decline trend levels off beyond 30%. Similar results are retained for mean-squared errors related to cut-points as plotted in Figure 5. It is worth noting that the mean-squared errors for the cut-points are greater at the lower end (poor vs. fair and fair vs. good) when using the first vignette, but greater at the higher end (good vs. very good and very good vs. excellent) when using the second vignette. This is not surprising given the first vignette's description of relatively good health, which should provide greater differential power toward the higher end, and the second vignette's description of relatively worse health, which should engender better anchors at the lower end.

5.3. External Validation in CHARLS

A working assumption that motivated the CFPS vignette design and the current study is that inter-group reporting heterogeneity is more or less stable. We now generalize the estimated cut-point shifts from the CFPS sample to the CHARLS sample by applying group-specific (as defined by observed characteristics) cut-points estimated from the CFPS vignettes to the CHARLS respondents. We again compare estimated disparities in self-rated health before and after vignette adjustment, assuming parallel and non-parallel cut-point shifts, respectively (see the right panel in Table 4). The effect of vignette adjustment is evident in several coefficient estimates. First, the coefficient of education increased by 70% from 0.01 in the ordered probit model to 0.017 in the parallel shift model, but dropped back to 0.009 in the non-parallel shift model, reflecting the complicated bi-directional reporting

heterogeneity by education. As shown in Figure 2, cut-points shift by education was non-parallel not only in magnitude but also in direction. It seems that well-educated people tend to invoke health optimism and consider themselves in fair health – downward shift of the cut-point between poor and fair – when they are indeed in poor health. But when they are indeed in very good or excellent health, they tend to employ peer comparison and higher standards – upward shift of the cut-points between good and very good and between very good and excellent – and thus underrate their true health status. Despite their similar coefficient estimates, the non-parallel vignette adjustment uncovers meaningful nuances underlying the educational gradient in health that cannot be noted in the conventional model.

Second, the coefficient of episodic memory was halved, while the coefficient of gender was doubled after adjusting for cut-point shift, either parallel or non-parallel. The income gradient in self-rated health also increased albeit to a smaller extent due to the relatively weak impact of income on reporting behaviors. Similarly, the coefficient of employment status remained relatively stable because of no significant cut-point shift by employment (see Table 3).

Third, being a cadre or CPC member was associated with significantly better self-rated health when reporting heterogeneity was ignored. After correcting for the negative cut-point shift by cadre and CPC membership, again either parallel or non-parallel, this association was no longer statistically significant. Similarly, the significant difference between rural and urban natives in the ordered probit model was considerably reduced (parallel shift model) or even disappeared (non-parallel shift model) after adjusting for urban natives' lower thresholds in rating their health status. In contrast, largely driven by a higher threshold between poor and fair health, the divorced or widowed respondents turned out to have better self-rated health compared to those who were married.

6. DISCUSSION

Using vignettes to anchor survey responses is not a new idea. The history of vignette methodology can date back to at least the 1970s, when sociologists employed vignettes to measure social status (Nosanchuk 1972) and racial attitudes (Farley et al. 1978). The statistical methods for analyzing vignette data were developed more than ten years ago (King et al. 2004; Tandon et al. 2003), and abundant research efforts have been devoted to understanding and refining the assumptions related to model identification (Angelini, Cavapozzi and Paccagnella 2011; Bago d'Uva et al. 2011; Kapteyn et al. 2011; Paccagnella 2011; Peracchi and Rossetti 2013). However, the challenge of how to design feasible vignettes and implement them effectively in large-scale general-population surveys remains largely unsettled. In their study using the SHARE data across a dozen European countries, Voňková and Hullegie (2011) found that the effectiveness of the vignette method varied significantly by both the choice of health domain (particularly problematic for cognition and breathing but less so for mobility) and the choice of vignette. This finding is worrisome because the same health domains and the corresponding vignettes have been adopted in, among others, WHO-WHS, WHO-MCSS, HRS, and CHARLS without any modification other than literal language translation, and thus the same sensitivity issues may have been widespread if not exacerbated. Further, it is usually impractical to collect data on multiple

health domains in general-purpose surveys where financial resource and interview time for measuring health are highly constrained and self-rated general health is likely a primary alternative. It is against this background that we have conducted this study as a contribution to survey methodology, through designing simple anchoring vignettes for self-rated general health in a general-purpose survey (the CFPS) and providing an evaluation of their usefulness in correcting for reporting heterogeneity bias as well as their external applicability to a different survey (i.e., CHARLS).

Capitalizing on the vignettes data from the nationally representative CFPS-2012 sample, we reach two significant conclusions in this study. First, reporting heterogeneity plays a significant role in biasing the measurement of health disparities among Chinese adults. In fact, our empirical findings suggest that reporting heterogeneity appears to be a predominant rather exceptional phenomenon in self-rated health because most of the socioeconomic and demographic characteristics examined here induce cut-points shifts, either parallel or non-parallel. And second, anchoring vignettes appear to be a cost-effective method of ameliorating the effects of reporting bias in surveys of self-rated health.

We quantify the consequential effect of reporting bias in self-rated health, revealing in vignette-anchored regression results that coefficients could be under- or over-estimated by twice as much as those without adjustment (e.g., education and episodic memory), depending on whether the cut-points are shifted upward or downward. Moreover, the significance levels changed for other covariates (e.g., political capital and residential and migration status) after adjustment.

We also quantify the magnitude of reporting heterogeneity through an experiment in which we fix the level of latent health status for a reference person but allow cut-points to vary within a single domain such as education. We found that the probability of reporting excellent health when applying the cut-points of college education is less than half of that when applying the cut-points of no schooling. This result is in marked contrast to previous research that reported less than a 10% difference (Bago d'Uva et al. 2008). Although we examine different measures of self-rated health[3] than do Bago d'Uva et al. (2008) and have the advantage of greater statistical power conferred by the large sample size of the CFPS, the interpretations of our findings are nonetheless unambiguous: the effects of reporting heterogeneity are substantial and anchoring vignettes can significantly reduce reporting bias.

Our analyses reveal three significant features of vignettes methodology. First, adjustment for reporting heterogeneity in the full sample can be achieved by extrapolating anchoring points from a relatively small subsample. In the CFSP data, administering vignettes to about 20% to 30% of the full sample was as effective as adding more cases. Second, using a single vignette can provide some anchoring that is comparable to using more vignettes. However, in a sample such as the CFSP that has a large age range, and hence great health differentials, a vignette that describes a relatively poor health scenario may lend more discriminant power to the lower end of the health spectrum, where the most striking gap occurs, compared to a

[3]We measure overall health status here whereas Bago d'Uva et al. (2008) divided general health into six domains, including mobility, cognition, pain, self-care, usual activities, and affect.

vignette that describes a relatively good health scenario. Third, collecting vignette data in a large-scale nationally representative sample can benefit other surveys of the same population, since we have demonstrated that the cut-point shifts estimated from the CFPS-2012 vignette data can be effectively applied to correct for reporting heterogeneity bias in the CHARLS-2011 sample.

Taken together, our findings have important implications for future research and public health policy. Given that measures of self-rated health have strong predictive power for objective health status and low data collection costs, they are likely to remain in use for research on health disparities in developing countries like China. On the other hand, the rapid social changes and the associated rising socioeconomic inequalities and social stratification in transition societies will increasingly complicate the pattern of health disparities. Reporting heterogeneity in health surveys may become more substantial as people of different social groups continue to diverge in their choice of reference group and the criteria they apply to gauge good versus poor health. If adjustment techniques to account for such heterogeneity, such as anchoring vignettes, become common practice, our research will yield better estimates of health disparities and provide higher quality information for policy makers.

7. LIMITATIONS AND FUTURE RESEARCH

Our study has several limitations that will benefit from future research. First, the vignette equivalence assumption may not hold in reality. For example, high-SES respondents may value mental health as much as physical health, whereas low-SES respondents may not. Also, given the complex multidimensional nature of health, vignette descriptions are likely to be incomplete and respondents may call upon their own experience to impute the missing information (van Soest et al. 2011). Similarly, the response consistency assumption may be violated when respondents report their own situation with a certain strategic consideration that is absent from vignette assessment (Bago d'Uva et al. 2011). A prominent example is that respondents from welfare-state countries tend to apply lower thresholds when assessing their own disability status than when evaluating the vignettes because of the economic incentive to exaggerate personal health problems for disability benefits eligibility (Gupta, Kristensen, and Pozzoli 2010). Although it is hard to contemplate such strategic behavior in China given that social welfare and health insurance benefits are largely contingent on social institutions (e.g. the household registration system) and collective entities (e.g. work units) rather than an individual's self-rating, we should still consider the possibility of the invalid response consistency assumption for other reasons.

Rigorous tests of these assumptions require extra data such as valid and reliable objective health measures, which are often available only in ad hoc studies. The present study is merely a first step toward a better understanding of the effects of reporting heterogeneity and the utility of anchoring vignettes in survey data on the socioeconomic and demographic disparities in self-rated health. Nevertheless, we find that even with short vignettes that do not attempt to incorporate particular aspects of health or age-specific health conditions, this method is useful in detecting reporting heterogeneity by SES and demographic characteristics and enabling appropriate anchoring to identify true health disparities. We also

find that vignette data collected in one population-based sample can be used to anchor reporting behaviors in a different sample of the same population or subpopulation. These methodological findings suggest that researchers may do well in designing their own anchoring vignettes, however simple they may seem, to fit a specific context or population, instead of merely borrowing standard ones that are context-blind. Future research is needed to improve the vignette design while retaining its simplicity and cost-effectiveness with respect to survey operation and anchoring performance, especially with general-purpose surveys in which resources are highly limited.

Acknowledgements

Earlier versions of this article were presented at the methodology seminar at the Quantitative Methodology Program of Survey Research Center, University of Michigan and the 2014 conference of Chinese Sociological and Demographic Methodology. The authors thank seminar participants, conference discussant Jun Li, and session participants for useful comments.

Funding

This study was supported by the National Institutes of Health under an investigator grant to Yu Xie (R01-HD074603).

References

Angelini, Viola; Cavapozzi, Danilo; Paccagnella, Omar. Dynamics of Reporting Work Disability in Europe. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011; 174(3):621–638.

Bago d'Uva, Teresa; Lindeboom, Maarten; O'Donnell, Owen; van Doorslaer, Eddy. Education-related Inequity in Healthcare with Heterogeneous Reporting of Health. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011a; 174(3):639–664.

Bago d'Uva, Teresa; Lindeboom, Maarten; O'Donnell, Owen; van Doorslaer, Eddy. Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. Journal of Human Resources. 2011b; 46(4):875–906. [PubMed: 22184479]

Bago d'Uva, Teresa; O'Donnell, Owen; van Doorslaer, Eddy. Differential Health Reporting by Education Level and its Impact on the Measurement of Health Inequalities among Older Europeans. International Journal of Epidemiology. 2008; 37(6):1375–1383. [PubMed: 18676985]

Bago d'Uva, Teresa; Van Doorslaer, Eddy; Lindeboom, Maarten; O'Donnell, Owen. Does Reporting Heterogeneity Bias the Measurement of Health Disparities? Health Economics. 2008; 17(3):351–375. [PubMed: 17701960]

Benjamins, Maureen Reindl; Hummer, Robert A.; Eberstein, Isaac W.; Nam, Charles B. Self-reported Health and Adult Mortality Risk: An Analysis of Cause-specific Mortality. Social Science & Medicine. 2004; 59(6):1297–1306. [PubMed: 15210100]

Chen, Feinian; Yang, Yang; Liu, Guangya. Social Change and Socioeconomic Disparities in Health over the Life Course in China: A Cohort Analysis. American Sociological Review. 2010; 75(1): 126–150. [PubMed: 20379373]

Datta Gupta, Nabanita; Kristensen, Nicolai; Pozzoli, Dario. External Validation of the Use of Vignettes in Cross-country Health Studies. Economic Modelling. 2010; 27(4):854–865.

Dowd, Jennifer Beam; Todd, Megan. Does Self-reported Health Bias the Measurement of Health Inequalities in U.S. Adults? Evidence Using Anchoring Vignettes From the Health and Retirement Study. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2011; 66B(4):478–489.

Farley, Reynolds; Schuman, Howard; Bianchi, Suzanne; Colasanto, Diane; Hatchett, Shirley. "Chocolate City, Vanilla Suburbs:" Will the Trend toward Racially Separate Communities Continue? Social Science Research. 1978; 7(4):319–344.

Ferraro, Kenneth F. Self-Ratings of Health among the Old and the Old-Old. Journal of Health and Social Behavior. 1980; 21(4):377–383. [PubMed: 7204931]

Ferraro, Kenneth F. Are Black Older Adults Health-Pessimistic? Journal of Health and Social Behavior. 1993; 34(3):201–214. [PubMed: 7989665]

Goldberg P, Guéguen A, Schmaus A, Nakache J-P, Goldberg M. Longitudinal Study of Associations between Perceived Health Status and Self Reported Diseases in the French Gazel Cohort. Journal of Epidemiology & Community Health. 2001; 55(4):233–238. [PubMed: 11238577]

Grol-Prokopczyk, Hanna; Freese, Jeremy; Hauser, Robert M. Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health. Journal of Health and Social Behavior. 2011; 52(2):246–261. [PubMed: 21673148]

Harzing, Anne-Wil. Response Styles in Cross-national Survey Research: A 26-country Study. International Journal of Cross Cultural Management. 2006; 6(2):243–266.

Herrnstein, Richard J.; Murray, Charles. The Bell Curve: Intelligence and Class Structure in American Life. New York: Free Press; 1996.

House, James S.; Lepkowski, James M.; Williams, David R.; Mero, Richard P.; Lantz, Paula M.; Robert, Stephanie A.; Chen, Jieming. Excess Mortality Among Urban Residents: How Much, for Whom, and Why? American Journal of Public Health. 2000; 90(12):1898–1904. [PubMed: 11111263]

Hu, Yuqing; Lei, Xiaoyan; Smith, James P.; Zhao, Yaohui. Effects of Social Activities on Cognitive Functions: Evidence from CHARLS. In: Smith, JP.; Majmundar, M., editors. National Research Council (US) Panel on Policy Research and Data Needs to Meet the Challenge of Aging in Asia. Washington D.C.: National Academies Press; 2012. p. 279-306.

Huisman, Martijn; Kunst, Anton E.; Mackenbach, Johan P. Socioeconomic Inequalities in Morbidity among the Elderly; a European Overview. Social Science & Medicine. 2003; 57(5):861–873. [PubMed: 12850111]

Idler, Ellen L.; Benyamini, Yael. Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. Journal of Health and Social Behavior. 1997; 38(1):21–37. [PubMed: 9097506]

Jones, Andrew M.; Rice, Nigel; Bago d'Uva, Teresa; Balia, Silvia. Applied Health Economics. New York: Routledge; 2007.

Jürges, Hendrik. True Health vs Response Styles: Exploring Cross-country Differences in Self-reported Health. Health Economics. 2007; 16(2):163–178. [PubMed: 16941555]

Kakwani, Nanak; Wagstaff, Adam; van Doorslaer, Eddy. Socioeconomic Inequalities in Health: Measurement, Computation, and Statistical Inference. Journal of Econometrics. 1997; 77(1):87–103.

Kapteyn, Arie; Smith, James P.; van Soest, Arthur. Vignettes and Self-Reports of Work Disability in the United States and the Netherlands. The American Economic Review. 2007; 97(1):461–473.

Kapteyn, Arie; Smith, James P.; Van Soest, Arthur; Vonkova, Hana. RAND Working Paper Series No. WR-840. Santa Monica: 2011. Anchoring Vignettes and Response Consistency.

King, Gary; Murray, Christopher JL.; Salomon, Joshua A.; Tandon, Ajay. Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. American Political Science Review. 2004; 98(1):191–207.

Knesebeck, Olaf von dem; Lüschen, Günther; Cockerham, William C.; Siegrist, Johannes. Socioeconomic Status and Health among the Aged in the United States and Germany: A Comparative Cross-sectional Study. Social Science & Medicine. 2003; 57(9):1643–1652. [PubMed: 12948573]

Luo, Ye; Wen, Ming. Can We Afford Better Health? A Study of the Health Differentials in China. Health: An Interdisciplinary Journal for the Social Study of Health Illness and Medicine. 2002; 6(4):471–500.

Marks, Gary N. Education, Social Background and Cognitive Functioning: The Decline of the Social. New York: Routledge; 2013.

McArdle, John J.; Smith, James P.; Willis, Robert J. Cognition and Economic Outcomes in the Health and Retirement Survey. In: Wise, David A., editor. Explorations in the Economics of Aging. Chicago: University of Chicago Press; 2011. p. 209-236.

Mirowsky, John; Ross, Catherine E. Education and Self-Rated Health: Cumulative Advantage and Its Rising Importance. Research on Aging. 2008; 30(1):93–122.

Murray, Christopher JL.; Özaltin, Emre; Tandon, Ajay; Salomon, Joshua A.; Chatterji, Somnath. Empirical Evalucation of the Anchoring Vignette Approach in Health Surveys. In: Murrary, Christopher JL.; Evans, David B., editors. Health Systems Performance Assessment: Debates, Methods and Empiricism. Geneva: World Health Organization; 2003. p. 369-399.

Nosanchuk TA. The Vignette as an Experimental Approach to the Study of Social Status: An Exploratory Study. Social Science Research. 1972; 1(1):107–120.

Paccagnella, Omar. Anchoring Vignettes with Sample Selection Due to Non-response. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011; 174(3):665–687.

Pei X, Rodriguez E. Provincial Income Inequality and Self-reported Health Status in China during 1991-7. Journal of Epidemiology & Community Health. 2006; 60(12):1065–1069. [PubMed: 17108303]

Peracchi, Franco; Rossetti, Claudio. The Heterogeneous Thresholds Ordered Response Model: Identification and Inference. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2013; 176(3):703–722.

Ross, Catherine E.; Wu, Chia-ling. The Links Between Education and Health. American Sociological Review. 1995; 60(5):719–745.

Salomon, Joshua A.; Tandon, Ajay; Murrary, Christopher JL. World Health Survey Pilot Study Collaborating Group. Comparability of Self Rated Health: Cross-sectional Multi-country Survey Using Anchoring Vignettes. British Medical Journal. 2004:328.

Schnittker, Jason. When Mental Health Becomes Health: Age and the Shifting Meaning of Self-Evaluations of General Health. Milbank Quarterly. 2005; 83(3):397–423. [PubMed: 16201998]

Singh-Manoux, Archana; Martikainen, Pekka; Ferrie, Jane; Zins, Marie; Marmot, Michael; Goldberg, Marcel. What Does Self Rated Health Measure? Results from the British Whitehall II and French Gazel Cohort Studies. Journal of Epidemiology & Community Health. 2006; 60(4):364–372. [PubMed: 16537356]

Tandon, Ajay; Murray, Christopher JL.; Salomon, Joshua A.; King, Gary. Murrary CJL, Evans DB. Satistical Models for Enhancing Cross-Population Comparability. Health Systems Performance Assessment: Debates, Methods and Empiricism. 2003:727–746.

Tandon, Ajay; Zhuang, Juzhong; Chatterji, Somnath. Inclusiveness of Economic Growth in the People's Republic of China: What Do Population Health Outcomes Tell Us? Asian Development Review. 2006; 23(2):53–69.

van Soest, Arthur; Delaney, Liam; Harmon, Colm; Kapteyn, Arie; Smith, James P. Validating the Use of Anchoring Vignettes for the Correction of Response Scale Differences in Subjective Questions. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011; 174(3):575–595.

Voňková, Hana; Hullegie, Patrick. Is the Anchoring Vignette Method Sensitive to the Domain and Choice of the Vignette? Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011; 174(3):597–620.

Wen, Ming; Fan, Jessie; Jin, Lei; Wang, Guixin. Neighborhood Effects on Health among Migrants and Natives in Shanghai, China. Health & Place. 2010; 16(3):452–460. [PubMed: 20060767]

Whyte, Martin King; Sun, Zhongxin. The Impact of China's Market Reforms on the Health of Chinese Citizens: Examining Two Puzzles. China: An International Journal. 2010; 8(1):1–32.

Willson, Andrea E.; Shuey, Kim M.; Elder, Glen H. Cumulative Advantage Processes as Mechanisms of Inequality in Life Course Health. American Journal of Sociology. 2007; 112(6):1886–1924.

Xie, Yu. The User's Manual of the China Family Panel Studies (2010). Beijing: Institute of Social Science Survey, Peking University; 2012.

Xie, Yu. Population Heterogeneity and Causal Inference. Proceedings of the National Academy of Sciences. 2013

Xie, Yu; Hu, Jingwei. An Introduction to the China Family Panel Studies (CFPS). Chinese Sociological Review. 2014

Zhang, Chunni; Xu, Qi; Zhou, Xiang; Zhang, Xiaobo; Xie, Yu. Comparing Poverty Rates from CFPS, CGSS, CHIP, and CHFS. In: Xie, Y., editor. Technical Reports of the China Family Panel Studies. Beijing: Institute for Social Science Survey, Peking University; 2012.

Zhao, Yaohui; Hu, Yisong; Smith, James P.; Strauss, John; Yang, Gonghuan. Cohort Profile: The China Health and Retirement Longitudinal Study (CHARLS). International Journal of Epidemiology. 2014; 43(1):61–68. [PubMed: 23243115]

Zimmer, Zachary; Amornsirisomboon, Pattama. Socioeconomic Status and Health among Older adults in Thailand: An Examination Using Multiple Indicators. Social Science & Medicine. 2001; 52(8): 1297–1311. [PubMed: 11281411]

Zimmer, Zachary; Kwong, Julia. Socioeconomic Status and Health Among Older Adults in Rural and Urban China. Journal of Aging and Health. 2004; 16(1):44–70. [PubMed: 14979310]

Zimmer, Zachary; Natividа, Josephina; Lin, Hui-Sheng; Chayovan, Napaporn. A Cross-National Examination of the Determinants of Self-Assessed Health. Journal of Health and Social Behavior. 2000; 41(4):465–481. [PubMed: 11198569]

Biographies

Hongwei Xu is a research assistant professor at the Survey Research Center of the Institute for Social Research, University of Michigan. His substantive research areas are focused on health inequalities, child development, and residential segregation. His methodological interests include hierarchical modeling of spatial, multilevel, and longitudinal data, causal inference using observational data, and survey methodology. His recent work has appeared in *Demography*, *Population Studies*, *Sociological Methodology*, *Journal of Marriage and Family*, and *Journal of Epidemiology & Community Health*.

Yu Xie is Otis Dudley Duncan University Distinguished Professor of Sociology, Statistics, and Public Policy and a research professor at the Institute for Social Research, University of Michigan, and Visiting Chair Professor at Peking University. His main areas of interest are social stratification, demography, statistical methods, Chinese studies, and sociology of science. His recently published works include: *Marriage and Cohabitation* (University of Chicago Press, 2007) with Arland Thornton and William Axinn, *Statistical Methods for Categorical Data Analysis* (Emerald, 2008, 2nd ed.) with Daniel Powers, and *Is American Science in Decline?* (Harvard University Press, 2012) with Alexandra Killewald.

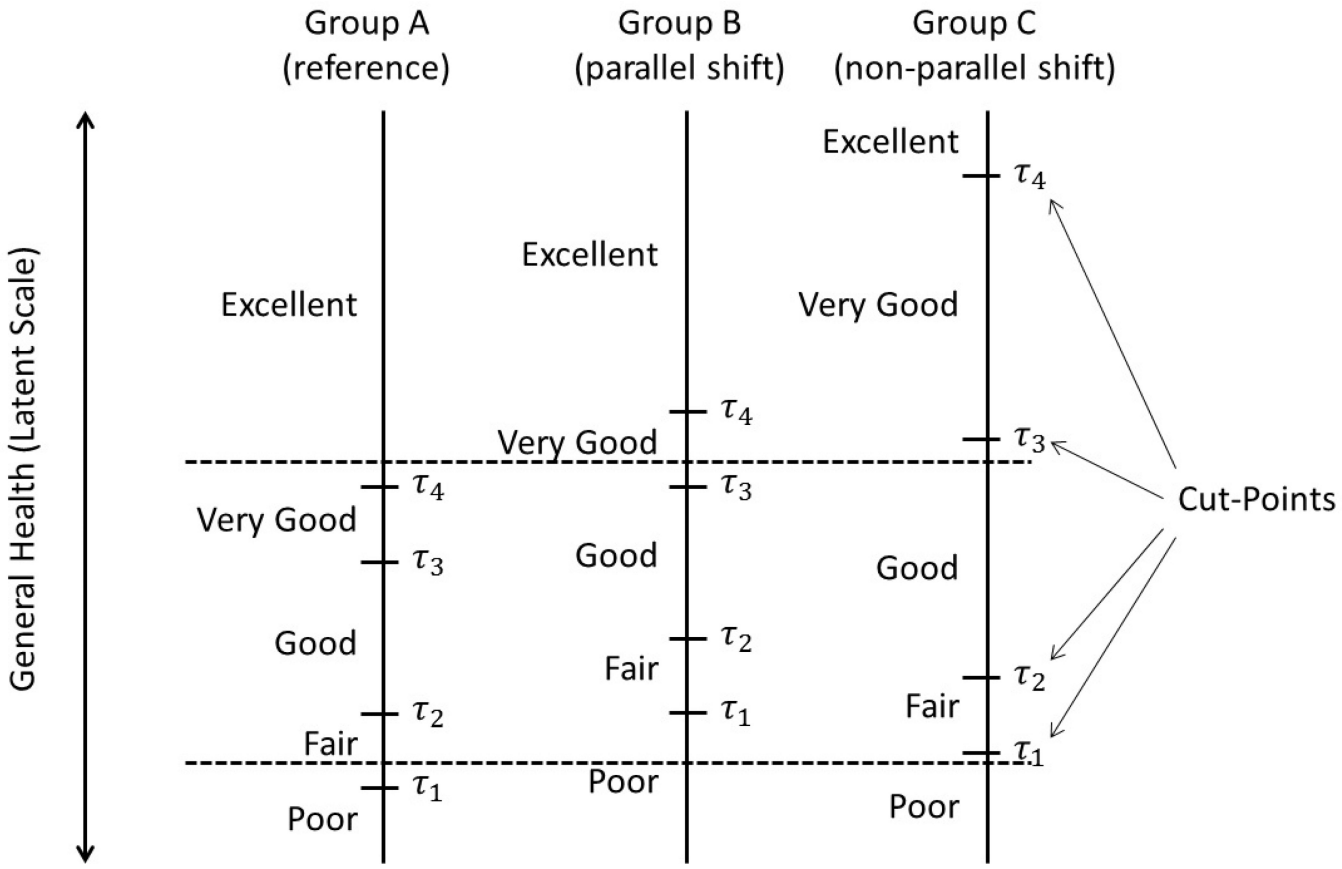

Figure 1.
Illustration of Cut-Point Shifts on the Latent Response Scales of Self-Rated Health

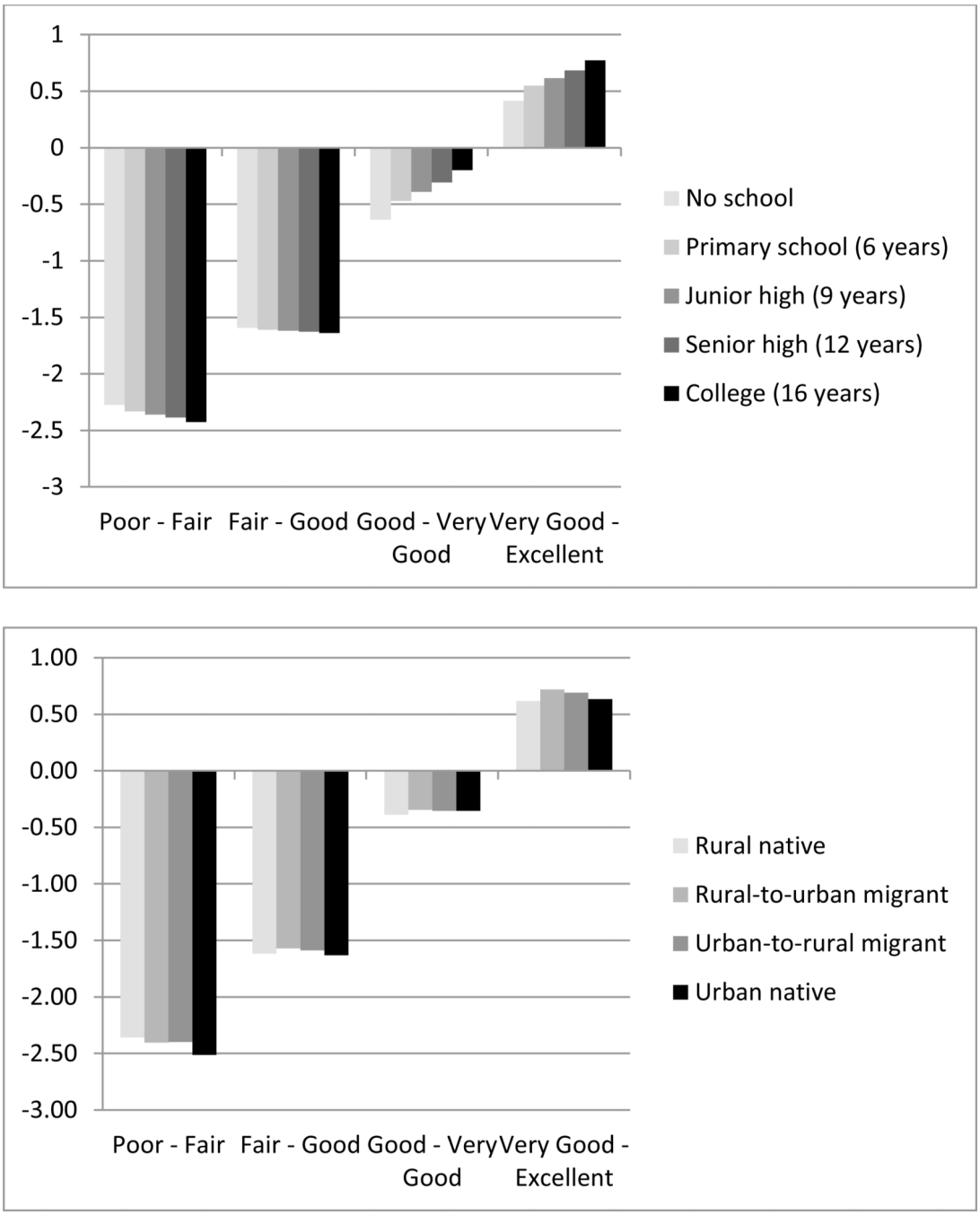

Figure 2.
Predicted cut-points by levels of education and migration status from the HOPIT model assuming non-parallel shift: CFPS-2012.

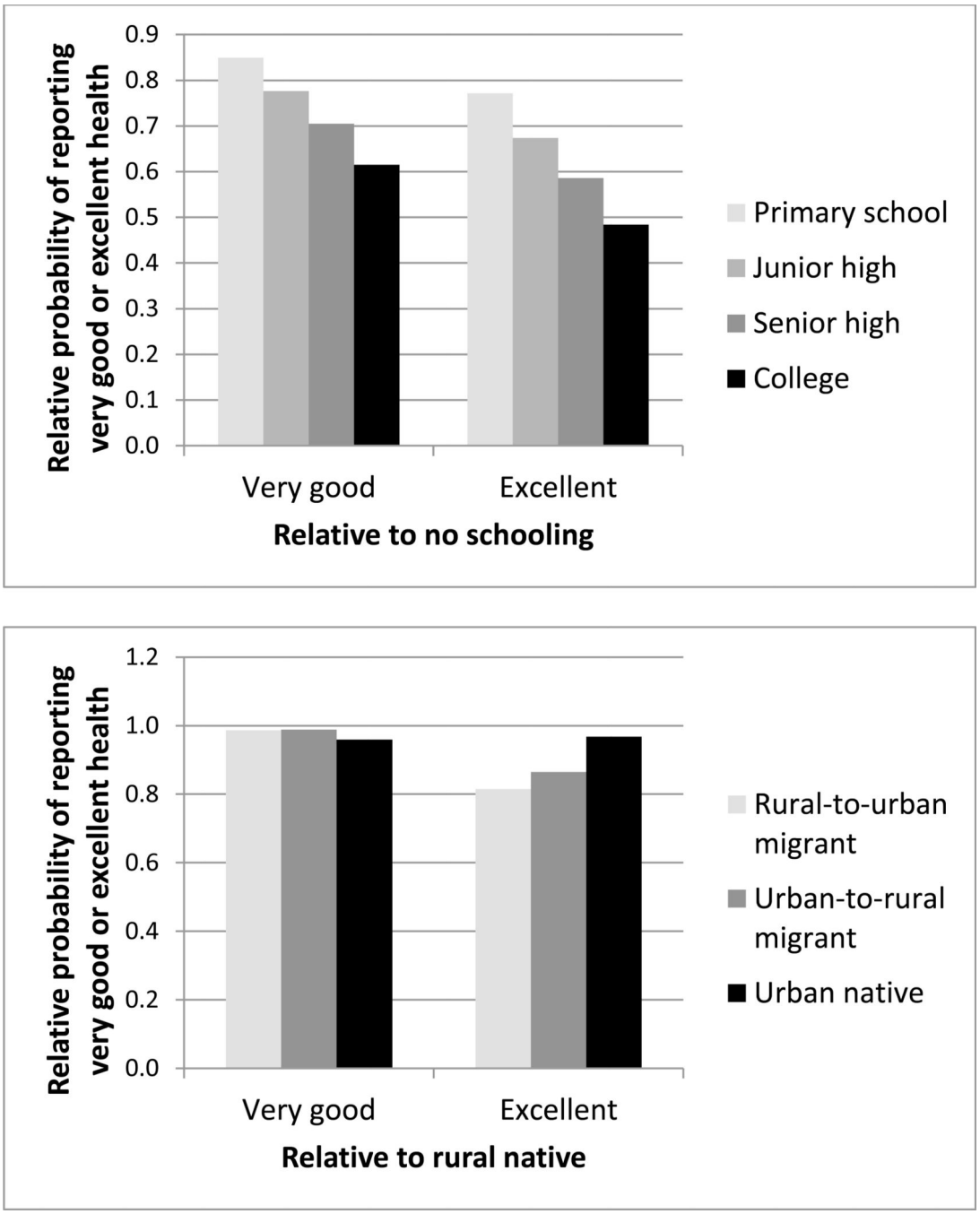

Figure 3.
Relative probabilities of reporting very good or excellent health for a reference person's health with varying cut-points by levels of education and migration status: CFPS-2012.

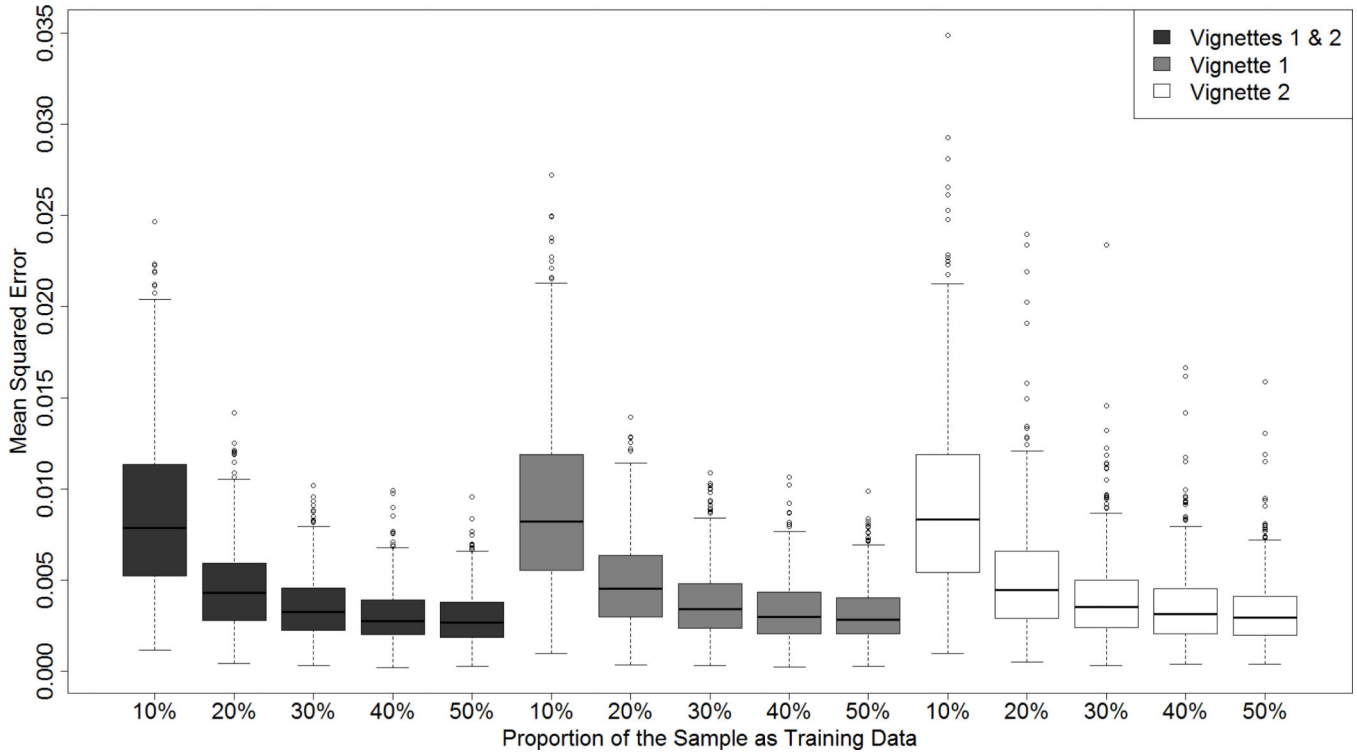

Figure 4.

Mean-squared error of predicted latent health from cross-validation of HOPIT models by randomly selecting a subset of the CFPS-2012 sample as training data.

Note: Vignette 1 describes a healthier person compared to Vignette 2.

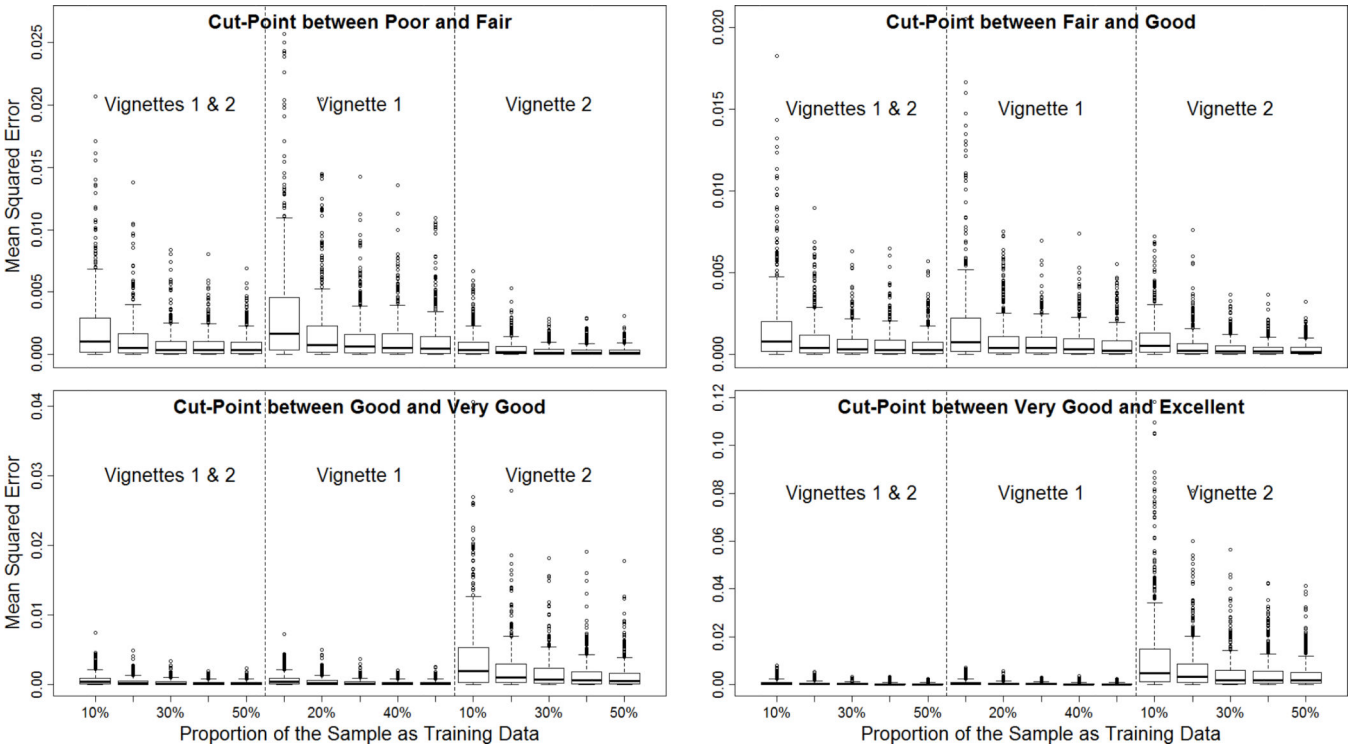

Figure 5.

Mean-squared error of predicted cut-points from cross-validation of HOPIT models by randomly selecting a subset of the CFPS-2012 sample as training data.

Note: Vignette 1 describes a healthier person compared to Vignette 2.

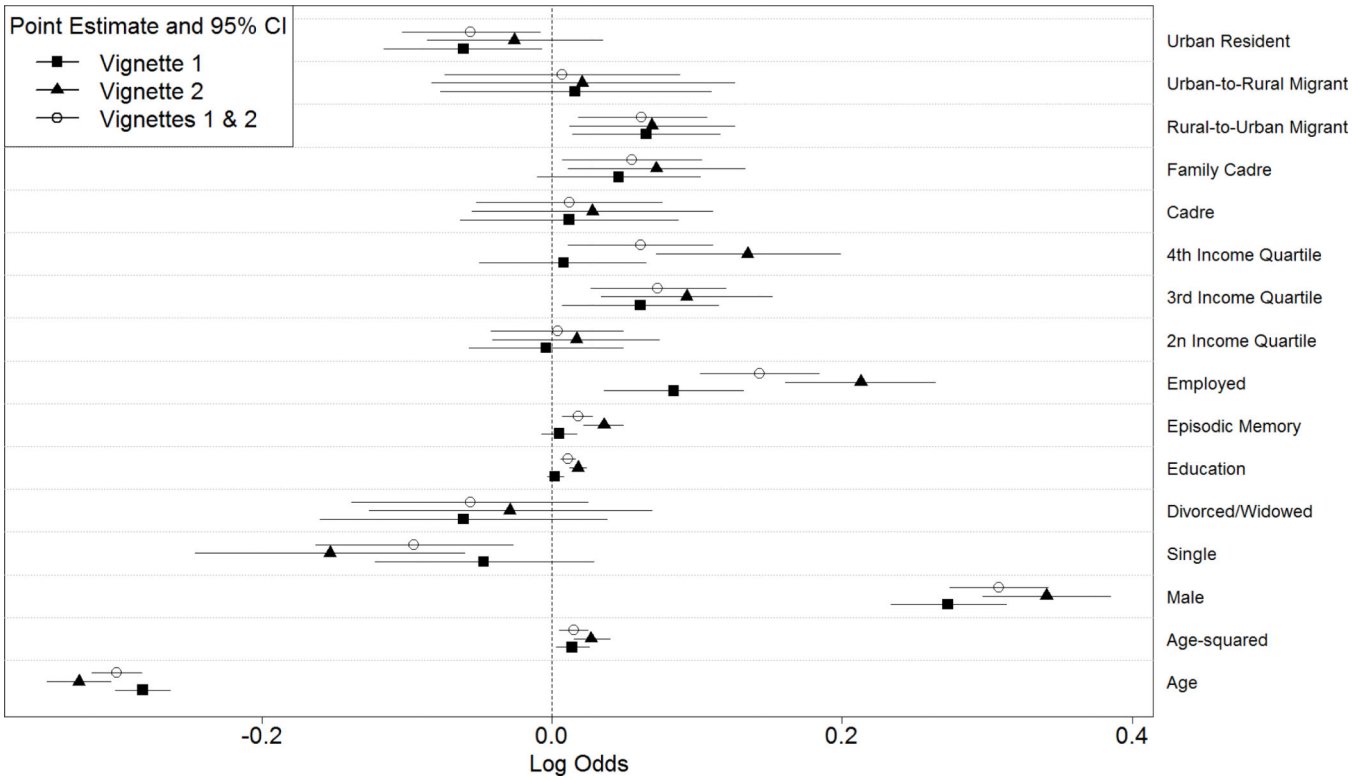

Figure 6.

Comparisons of coefficient estimates for the health component of HOPIT models by using different vignettes: CFPS-2012.

Note: Vignette 1 describes a healthier person compared to Vignette 2.

Table 1

Frequency distributions of self-rated health and vignette ratings

	CFPS-2012			CHARLS-2011
	Self-Rated Health (%)	**Vignette 1 (%)**	**Vignette 2 (%)**	**Self-Rated Health (%)**
Poor	16.4	0.0	60.6	26.1
Fair	18.4	4.5	23.6	48.1
Good	34.8	27.3	15.0	16.5
Very Good	20.5	40.0	0.8	8.7
Excellent	10.0	28.1	0.0	0.7
N		23,207		5,928

Note: Vignette 1 describes a person in better health status compared to Vignette 2 by design.

Table 2

Descriptive statistics of independent variables

	CFPS-2012		CHARLS-2011	
	Mean	**SD**	**Mean**	**SD**
Age (years)	42.7	14.7	56.5	7.1
Male (%)	49.4	—	47.3	—
Marital status (%)				
Single	14.5	—	0.9	—
Married/Cohabitation	81.2	—	91.2	—
Divorced/Widowed	4.4	—	8.0	—
Years of education	7.6	4.7	5.9	4.4
Episodic memory	4.3	1.9	3.7	1.7
Employed (%)	72.8	—	74.0	—
Family income (RMB)	14,490	24,795	6,048	9,478
Cadre/Party member (%)	7.7	—	4.4	—
Had a family member as cadre/Party (%)	13.8	—	—	—
Residence and *hukou* status (%)				
Rural resident with rural *hukou*	51.9	—	57.3	—
Rural-to-urban migrant	18.7	—	20.2	—
Rural resident with urban *hukou*	4.6	—	2.1	—
Urban resident with urban *hukou*	24.9	—	20.4	—
Region (%)				
Northeast	14.7	—	7.8	—
Northern coast	12.3	—	14.8	—
Eastern coast	10.9	—	8.8	—
Southern coast	10.1	—	8.2	—
Yellow River middle reach	18.3	—	20.2	—
Yangtze River middle reach	8.8	—	16.8	—
Southwest	13.5	—	19.6	—
Northwest	11.4	—	3.8	—
N	23,207		5,928	

Table 3

Coefficient estimates of cut-point shift (parallel versus non-parallel) from the vignette data: CFPS-2012.

		Non-Parallel Shift				
	Parallel Shift	**Poor - Fair**	**Fair - Good**	**Good - Very Good**	**Very Good - Excellent**	**Wald Test of Parallel Shift (df=3)**
Age (centered)	−0.019***	−0.022**	−0.054***	0.005	−0.013	68.96***
Age-square	−0.019***	−0.035***	−0.029***	−0.011*	−0.002	33.67***
Male (ref: female)	0.103***	0.135***	0.121***	0.092***	0.056***	15.40**
Marital status (ref: married/cohabitation)						
Single	−0.080***	−0.036	−0.149***	−0.103***	−0.059	16.66***
Divorced/Widowed	0.014	0.081*	0.029	−0.014	−0.043	7.39
Years of education	0.009***	−0.010***	−0.003	0.028***	0.023***	260.62***
Episodic memory	−0.022***	−0.032***	−0.039***	−0.009*	−0.010*	46.02***
Employed (ref: no)	−0.006	0.012	−0.015	−0.006	−0.021	2.97
Family income quartiles (ref: poorest)						
2nd	−0.003	−0.011	0.007	0.014	−0.022	3.57
3rd	0.045**	0.040*	0.072***	0.056**	0.018	6.09
4th (richest)	0.031	0.030	0.027	0.052*	0.011	3.53
Cadre/Party member (ref: no)	−0.083***	−0.098***	−0.101***	−0.113***	−0.018	9.82*
Family cadre/Party member (ref: no)	−0.020	−0.043*	−0.036	−0.017	0.017	4.62
Residence and *hukou* status (ref: rural)						
Rural-to-urban migrant	0.032*	−0.048*	0.046*	0.044*	0.103***	42.13***
Rural resident with urban *hukou*	0.019	−0.041	0.029	0.034	0.074	6.92
Urban resident with urban *hukou*	−0.038*	−0.158***	−0.018	0.035	0.018	73.60***

Notes: Estimates of ancillary parameters are not shown. Coefficients in the HOPIT models have been rescaled to be comparable to those in the ordered probit model.

* $p<0.05$;

** $p<0.01$;

*** $p<0.001$.

Sociol Methodol. Author manuscript; available in PMC 2017 August 01.

Table 4

Coefficient estimates from standard ordered probit versus hierarchical ordered probit (HOPIT) models of self-rated health: CFPS-2012 and CHARLS-2011.

	CFPS-2012 (N = 23,207)			CHARLS-2011 (N = 5,928)		
		HOPIT			HOPIT	
	Ordered Probit	Parallel Shift	Non-Parallel Shift	Ordered Probit	Parallel Shift	Non-Parallel Shift
Age (mean-centered and divided by 10)	−0.279***	−0.294***	−0.301***	−0.252**	−0.285**	−0.303***
Age-square	0.031***	0.015**	0.015**	0.045	0.026	0.020
Male (ref: female)	0.229***	0.312***	0.308***	0.152***	0.270***	0.298***
Marital status (ref: married/cohabit)						
Single	−0.051*	−0.114***	−0.095***	−0.499**	−0.555**	−0.557***
Divorced/Widowed	−0.086*	−0.074	−0.057	0.097	0.122*	0.162**
Years of education	0.004*	0.012***	0.011***	0.010*	0.017***	0.009*
Episodic memory	0.036***	0.018***	0.017***	0.057***	0.030**	0.021*
Employed (ref: no)	0.143***	0.138***	0.143***	0.347***	0.313***	0.351***
Family income quartiles (poorest)						
2nd	0.003	0.000	0.003	0.177***	0.165***	0.172***
3rd	0.032	0.068**	0.072**	0.254***	0.297***	0.313***
4th (richest)	0.037	0.061*	0.061*	0.369***	0.402***	0.425***
Cadre/Party member (ref: no)	0.080**	0.013	0.012	0.149*	0.061	0.065
Family cadre/party member (ref: no)	0.066**	0.050*	0.055*	—	—	—
Residence and *hukou* status (ref: rural)						
Rural-to-urban migrant	0.032	0.057*	0.063**	0.151***	0.166**	0.143***
Rural resident with urban *hukou*	−0.018	−0.003	0.007	0.114	0.126	0.118
Urban resident with urban *hukou*	−0.033	−0.064**	−0.054*	0.169***	0.108*	0.099
σ	—	1.244***	1.231***	—	0.914***	0.917***

Notes: Estimates of ancillary parameters are not shown. All the models control for regional variation. Coefficients in the HOPIT models have been rescaled to be comparable to those in the ordered probit model.

* $p<0.05$;
** $p<0.01$;
*** $p<0.001$.

Table 5

Rural-urban stratified coefficient estimates from standard ordered probit versus hierarchical ordered probit (HOPIT) models of self-rated health: CFPS-2012.

	Rural Sample (N = 16,002)			Urban Sample (N = 7,205)		
		HOPIT			HOPIT	
	Ordered Probit	Parallel Shift	Non-Parallel Shift	Ordered Probit	Parallel Shift	Non-Parallel Shift
Age (mean-centered and divided by 10)	−0.285***	−0.304***	−0.312***	−0.261***	−0.261***	−0.265***
Age-square	0.022***	0.004	0.005	0.057***	0.048***	0.042***
Male (ref: female)	0.238***	0.310***	0.306***	0.204***	0.303***	0.296***
Marital status (ref: married/cohabit)						
Single	−0.034	−0.054	−0.034	−0.072	−0.218***	−0.199**
Divorced/Widowed	−0.065	−0.046	−0.033	−0.102	−0.096	−0.067
Years of education	0.003	0.011***	0.011***	0.009*	0.013**	0.010*
Episodic memory	0.032***	0.020**	0.019**	0.046***	0.017	0.015
Employed (ref: no)	0.164***	0.163***	0.157***	0.165***	0.206***	0.210***
Family income quartiles (poorest)						
2nd	−0.005	−0.004	0.000	0.024	−0.045	−0.041
3rd	0.017	0.068*	0.073**	0.064	0.000	0.008
4th (richest)	0.007	0.044	0.044	0.091*	0.035	0.038
Cadre/Party membership (ref: no)	0.098*	0.001	0.000	0.049	0.008	0.007
Family cadre/Party member (ref: no)	0.041	0.032	0.039	0.096**	0.078*	0.076*
Urban *hukou* (ref: rural *hukou*)	−0.020	−0.033	−0.026	−0.083*	−0.117**	−0.101*
Σ	—	1.284***	1.271***	—	1.154***	1.157***

Notes: Estimates of ancillary parameters are not shown. All the models control for regional variation. Coefficients in the HOPIT models have been rescaled to be comparable to those in the ordered probit model.

*$p<0.05$;

**$p<0.01$;

*** $p<0.001$.

Table 6

Correlation coefficients among predicted self-rated health anchored by different vignettes: CFPS-2012.

	Latent Health Scale			Predicted Self-Rated Health based on Modal Item-Response Probability		
	Vignette 1	Vignette 2	Vignettes 1 & 2	Vignette 1	Vignette 2	Vignettes 1 & 2
Vignette 1	1	—	—	1	—	—
Vignette 2	0.97	1	—	0.92	1	—
Vignettes 1 & 2	0.99	0.99	1	0.87	0.88	1

Note: Vignette 1 describes a healthier person compared to Vignette 2.

Sociol Methodol. Author manuscript; available in PMC 2017 August 01.