JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

Demography
DOI 10.1007/s13524-015-0422-1

Promises and Pitfalls of Anchoring Vignettes in Health Survey Research

Hanna Grol-Prokopczyk[1] · **Emese Verdes-Tennant**[2] · **Mary McEniry**[3] · **Márton Ispány**[4]

Abstract Data harmonization is a topic of growing importance to demographers, who increasingly conduct domestic or international comparative research. Many self-reported survey items cannot be directly compared across demographic groups or countries because these groups differ in how they use subjective response categories. Anchoring vignettes, already appearing in numerous surveys worldwide, promise to overcome this problem. However, many anchoring vignettes have not been formally evaluated for adherence to the key measurement assumptions of vignette equivalence and response consistency. This article tests these assumptions in some of the most widely fielded anchoring vignettes in the world: the health vignettes in the World Health Organization (WHO) Study on Global AGEing and Adult Health (SAGE) and World Health Survey (WHS) (representing 10 countries; $n = 52{,}388$), as well as similar vignettes in the Health and Retirement Study (HRS) ($n = 4{,}528$). Findings are encouraging regarding adherence to response consistency, but reveal substantial violations of vignette equivalence both cross-nationally and across socioeconomic groups. That is, members of different sociocultural groups appear to interpret vignettes as depicting fundamentally different levels of health. The evaluated anchoring vignettes do not fulfill their promise of providing interpersonally comparable measures of health. Recommendations for improving future implementations of vignettes are discussed.

Electronic supplementary material The online version of this article (doi:10.1007/s13524-015-0422-1) contains supplementary material, which is available to authorized users.

✉ Hanna Grol-Prokopczyk
hgrol@buffalo.edu

[1] Department of Sociology, University at Buffalo, State University of New York, 430 Park Hall, Buffalo, NY 14260, USA

[2] World Health Organization, Geneva, Switzerland

[3] Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

[4] Faculty of Informatics, University of Debrecen, Debrecen, Hungary

Keywords Anchoring vignettes · Survey methods · Self-rated health · Comparative health research · Reporting heterogeneity

> I stare at the pain scale, a simple number line complicated by only two phrases. Under zero: "no pain." Under ten: "the worst pain imaginable."
>
> The worst pain imaginable . . . Whipped with nettles? Buried under an avalanche of sharp rocks? Impaled with hundreds of nails? . . .
>
> I chose thirty percent—three. Which seemed, at the time, quite substantial.
>
> "Three is nothing," my father [a doctor] tells me now. "Three is go home and take two aspirin."
>
> It would be helpful, I tell him, if that could be noted on the scale.
>
> —Eula Biss, "The Pain Scale" (2005)

As demography enters the era of big data, characterized by an "explosion . . . of individual-level population data" collected in a majority of the worlds' countries (Ruggles 2014:287), comparative research becomes increasingly common—and crucial. Kapteyn highlights the role of "harmonized microdata from different countries" in clarifying the relationship between national policies and health and aging outcomes (2010:S193); a National Institute of Aging (NIA) report argues that "cross-study comparative analysis" would accelerate research on genetic underpinnings of social and behavioral outcomes (2012:1); Burgard and Chen (2014) emphasize the role of comparison in understanding health disparities within and across countries; and Dong et al. argue that generally, "Comparison and comparability lie at the heart of social science" (2015:1062). In this context, data harmonization becomes critical, since differences in measurement cloud interpretation of cross-study or cross-population comparisons (National Institute on Aging 2012).

Since the early 2000s, anchoring vignettes have been promoted as a harmonization strategy to overcome a key challenge of comparative survey research, namely, the tendency of different groups to use subjective response categories in systematically different ways (e.g., more or less optimistically). (As discussed in more detail later, *anchoring vignettes* are brief hypothetical descriptions of fictional characters who exemplify the trait of interest—for example, pain—to a lesser or greater degree.) If effective, anchoring vignettes would enable harmonization of subjective variables, including those highlighted in the 2012 NIA report, such as well-being, depression, and stress. However, the method is predicated on at least one highly questionable assumption: cross-respondent vignette equivalence. is thus unclear whether anchoring vignettes function as intended.

This article assesses the validity of some of the most widely fielded health vignettes in the world, subjecting them to the most rigorous available tests of key measurement assumptions. Do anchoring vignettes, as currently formulated, fulfill their promise of enabling valid cross-group comparisons? If not, what improvements can be proposed?

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

Background and Motivations

Reporting Heterogeneity

Self-reports of health—including both overall health and specific domains of health—are often incomparable across national, racial/ethnic, and other demographic groups (e.g., King et al. 2004; Murray et al. 2002). In particular, accumulating evidence shows that when rating health using subjective ordinal categories (e.g., "none, mild, moderate, severe, or extreme" to describe pain or other health impairments), some groups use certain response categories more liberally or more sparingly than others. More formally, groups may differ in where on the latent health spectrum they locate the thresholds between adjacent response categories. Figure 1 depicts three populations with different understandings of how much pain constitutes mild pain, moderate pain, and so on; each population uses different cutpoints (marked with τs) to demarcate these categories. Each group's "mild" thus corresponds to a different portion of the latent pain scale. Such differences in rating style are referred to as reporting heterogeneity (e.g., Bago D'Uva et al. 2011b) or response category differential item functioning (DIF) (King et al. 2004).

Recent studies support the notion that health-related reporting heterogeneity is nontrivial across nationalities (e.g., Iburg et al. 2002; Jürges 2007; Jylhä et al. 1998; Murray et al. 2002; Zimmer et al. 2000), races/ethnicities (e.g., Menec et al. 2007; Shetterly et al. 1996; Smith 2003), and socioeconomic categories (e.g., Dowd and Zajacova 2007; Grol-Prokopczyk et al. 2011), and that failure to account for group

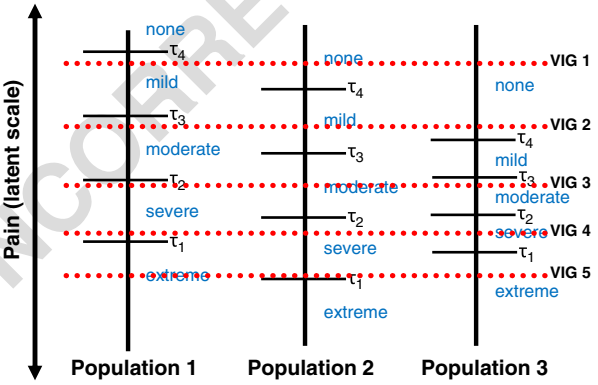

Fig. 1 Reporting heterogeneity and the anchoring vignette method. **Description:** Populations may differ in how they use subjective categories to describe pain (or other aspects of health), that is, they may demonstrate "reporting heterogeneity" (Bago D'Uva et al. 2011b; cf. King et al. 2004). Here, Population 1 uses systematically higher intercategory cutpoints (τs) than Population 2, while Population 3 shows a compression of cutpoints relative to the other groups. In this scenario, the three groups could have equal mean levels of pain, but nonetheless use different terms to refer to that level of impairment. By giving the same series of anchoring vignettes (here, "VIG 1" through "VIG 5", marked with dotted lines) to all respondents, researchers can determine how different groups use subjective response categories. Here, the pain in vignette 2 would be rated as "moderate" by Population 1, "mild" by Population 2, and "none" by Population 3. More formally, researchers can estimate where different groups locate intercategory thresholds (here, $\tau_1 - \tau_4$), and adjust for such different thresholds in subsequent analyses, enabling unbiased group comparison. To facilitate later comparison with other health domains, this pain scale is depicted as going from extreme pain at the lower end to no pain at the higher end. Higher levels of a construct thus consistently represent better health

differences in health-reporting style can lead to incorrect (and sometimes, highly implausible) research findings. Indeed, rankings of regions by self-reported health are frequently dramatically at odds with rankings based on objective measures. For example, based on subjective self-rated health, Indonesia and Nepal appear to be far healthier countries than France and Spain, despite the former's much lower life expectancies (Sadana et al. 2002; cf. Sen 2002). Such findings underscore the threat of reporting heterogeneity to comparative research validity.[1]

In statistical terms, researchers relying on subjective health assessments contend with an identification problem (Bago D'Uva et al. 2011b:879–880): one cannot simultaneously identify the location on the absolute, latent scale of respondents' (1) response-category thresholds and (2) perceived health. Standard survey analyses assume cross-group equivalence of the former to derive putatively comparable measures of the latter, leading to the questionable findings just mentioned. The challenge for survey researchers has been to find a way to circumvent this problem without prohibitive costs.

Anchoring Vignettes

In the early 2000s, researchers at the WHO systematically compared techniques for addressing reporting heterogeneity, and concluded that anchoring vignettes were "the most promising" of available strategies (Murray et al. 2002:429; cf. Tandon et al. 2003). As mentioned earlier, an anchoring vignette is a brief, hypothetical description of a fictional character who exemplifies the trait of interest (e.g., pain) to a lesser or greater degree. For example, "Laura has a headache once a month that is relieved one hour after taking a pill. During the headache she can carry on with her day to day affairs." (Online Resources 1 and 2 present additional vignette texts.) Respondents are asked to rate their own level of the trait and, using the same set of response categories, to also rate the fictional character's level. Respondents are given multiple vignettes per domain, each representing different points along the health spectrum. Since identical vignettes are given to all respondents, any differences in ratings of a given vignette are considered indicative of reporting heterogeneity. That is, vignette ratings can be used to determine what different groups mean by terms such as "mild" or "moderate," and to statistically estimate the locations of each group's intercategory thresholds (τs)—thereby overcoming the identification problem. Group differences in rating style can then be statistically accounted for, allowing for intergroup comparisons unbiased by reporting heterogeneity. This logic is depicted in Fig. 1. (For more formal overviews of vignette methodology, including of techniques for vignette-based adjustments, see King et al. 2004; King and Wand 2007; Rabe-Hesketh and Skrondal 2002; and van Soest and Vonkova 2014.)

For many health domains, anchoring vignettes represent a convenient alternative to expensive or inconvenient "gold standard" measures. For example, while Snellen or LogMAR eye exams are gold standard measures of visual acuity, they require in-person administration, adequate and standardized space and lighting, etc., and thus are not feasible in all surveys. Because anchoring vignettes depend on only those resources required for the survey itself (however administered), they may be a cost-saving alternative to measured tests or professional assessments (King et al. 2004). Some

[1] Although we focus on health (given the widespread use of anchoring vignettes in health surveys), similar issues arise whenever subjective self-ratings are used.

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

health conditions, however, have no "gold standard" measure beyond self-reports (e.g., pain; Schiavenato and Craig 2010). In such cases, anchoring vignettes may represent one of the only hopes for collecting internationally comparable measures. Vignettes could also potentially improve measurement in experimental and clinical settings.

Since the early 2000s, health-related anchoring vignettes have appeared in numerous regional, national, and international surveys, including but not limited to the Los Angeles Family and Neighborhood Survey (L.A.FANS); the Puerto Rican Elderly: Health Conditions (PREHCO) project; the Health and Retirement Study (HRS); the English Longitudinal Study of Ageing (ELSA); the Survey of Health, Ageing and Retirement in Europe (SHARE); the Study on Global AGEing and Adult Health (SAGE; Kowal et al. 2012); and the World Health Survey (WHS) (cf. Hopkins and King 2010:202–203). This represents an enormous quantity of data. The health vignettes in the WHO WHS and SAGE surveys alone reached nearly 350,000 respondents in 70 countries, and modified subsets of these vignettes have appeared in other large surveys including HRS, SHARE, and ELSA. Despite such widespread use, no systematic evaluation of the WHO vignettes or their variants has been conducted to date regarding adherence to the method's statistical assumptions.

Measurement Assumption 1: Vignette Equivalence

The anchoring vignette method depends on two key measurement assumptions. The first is vignette equivalence—or, more precisely, cross-respondent vignette equivalence (Grol-Prokopczyk 2014). *Vignette equivalence* (VE) refers to respondents perceiving the vignettes as representing the same absolute position on the latent health spectrum. (Thus, Fig. 1 depicts the vignettes as flat horizontal lines: a given vignette represents the same position on the latent scale for all populations.) Violations of VE may occur if groups interpret the vignette texts in systematically different ways. For example, if a vignette character's annual medical visit is interpreted by residents of rich countries as a beneficial, preventive check-up, and hence indicative of good health, but is interpreted by residents of poor countries as a sign of frequent medical need and hence of *poor* health, then VE has been violated.

VE is a critical assumption for any vignette-based adjustment of self-reports, parametric or nonparametric (King and Wand 2007:49[2]; King et al. 2004:194; van Soest and Vonkova 2014:116). If different groups do not interpret a vignette as representing the same absolute level of health, then the ability of anchoring vignettes to circumvent the identification problem disappears: level of health is no longer held constant, response thresholds for different groups cannot be compared, and self-ratings cannot be adjusted for comparability.

Anchoring vignette studies routinely acknowledge the necessity of VE, but rarely theorize the plausibility of the assumption. Yet, if we present the claim of VE in slightly different terms—that groups, even those differing in how they understand response categories, will *not* differ in how they understand descriptions of vignette characters—this proposition seems far from guaranteed.

On the one hand, the plausibility of VE could be defended by highlighting the contrast between short (often single-word) subjective response categories, and longer,

[2] King and Wand's nonparametric method contends with respondents misordering vignettes in a series, but treats such misorderings as "random measurement error," not as fundamental violations of VE (2007:49).

potentially more objective descriptions of health in base vignette texts (King et al. 2004:194). For example, van Soest et al. (2011) asked Irish university students to rate their own and vignette characters' drinking patterns as "Mild, Moderate, Some Cause for Concern, Excessive, [or] Extreme," but in the vignette texts, described characters who consumed a specific number of alcoholic drinks in a night. Subjective response categories, which might be used differently by different groups, were thus paired with concrete, quantified drinking scenarios, which presumably denote similar levels of alcohol consumption to all respondents. Here, the assumption of VE had *prima facie* plausibility. More generally, evidence that "objective rather than attitudinal" descriptions minimize cross-cultural misunderstanding (Pasick et al. 2001:240) suggests that if vignettes describe characters by using concrete, objective detail, the latent level of health could be understood similarly across groups.

On the other hand, it is not always obvious what details are concrete or "objective". Researchers have found that concepts as ostensibly straightforward as "household member" (Pasick et al. 2001:231), "cut" (Skevington 2002:138), and "chest pain" (Hanna et al. 2012) are interpreted differently by different cultural groups. Some concepts are very familiar to some populations but utterly unfamiliar to others (e.g., "routine check-up"; Pasick et al. 2001:233). Although some cross-group differences in survey interpretation reflect preventable "microlinguistic" translation problems (pertaining to word choice and grammar), others reflect more challenging "macrolinguistic" problems, in which cultural differences lead to incommensurability in conceptual understandings (Pan and Fond 2014:184). As phrased by Hunt and Bhopal (2004:618), "latent variables are not shared across languages". While some researchers appear optimistic that appropriate protocols can lead to "functionally equivalent" translations (Pan and Fond 2014:181), others are pessimistic, arguing that "the nature of language itself places limits on the extent to which complete equivalence can be achieved" (Angel 2013:228).

In short, theoretical and empirical evidence of intergroup incomparability in understandings of health concepts is sufficiently strong that VE should not be taken for granted. The anchoring vignette method is predicated on a questionable assumption.

Measurement Assumption 2: Response Consistency

The second key measurement assumption of anchoring vignettes—*response consistency* (RC)—refers to respondents rating themselves and vignette characters using the same thresholds (i.e., the τs in Fig. 1 are in the same positions for both self- and vignette-ratings). If respondents hold themselves to different standards than vignette characters, or use standards inconsistently across vignettes in a series, then RC is violated, and cutpoints calculated from vignettes will not correctly adjust self-ratings. Given RC's specificity to the anchoring vignette method, detailed discussions of why respondents may or may not adhere to this assumption are few. Bago D'Uva and colleagues suggest that external factors may affect self-ratings but not vignette ratings: for example, "[N]onworking individuals may experience social pressure and/or financial incentives to understate their own health but not that of hypothetical individuals" (2011b:87). Au and Lorgelly's post-survey interviews indicate that young respondents may have difficulty imagining certain problems among people their age (e.g., difficulty walking), or may use different scales for self-ratings because they have a "higher threshold for minor ailments than the average person" (2014:1724–1725). Although assessing RC is often

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

challenging, as discussed later, this article tests RC where possible, in order to present as complete an assessment of anchoring vignette validity as possible.

Testing Vignette Equivalence

Developing methods to test VE has proven conceptually and statistically challenging, as evidenced by the lack, until recently, of strong tests of this assumption. In initial pretests of vignettes, WHO researchers conducted only a minimal test of VE, namely, to check that most respondents correctly rank-ordered vignettes in a series (Murray et al. 2003:376). King et al.'s (2004) foundational article similarly relies on measures of rank-ordering to assess VE. Such tests are "weak", in that correct rank-ordering is a necessary but not sufficient condition for vignette equivalence. For several years, all tests of VE were based on examinations of rank-ordering, albeit with some variations—for example, looking for systematic patterns among nonnormative rankings, or for differences in ranking consistencies across national or other groups (e.g., Kristensen and Johansson 2008; Rice et al. 2011). The studies cited here all found support for VE.

A novel, more stringent approach to testing VE was proposed by Bago D'Uva et al. (2011b), and implemented using ELSA's mobility and cognition vignettes. Bago D'Uva et al. observe that, if VE holds, then the perceived distance (along the latent health spectrum) between any two vignettes in a series should be constant across groups. Models cannot simultaneously identify the locations on the latent spectrum of all vignettes in a series; however, if one vignette is constrained to be at the same position for all respondents—for example, by setting it to zero—then locations of other vignettes can be estimated relative to this reference vignette. The perceived locations of vignettes can then be compared across groups, to directly test VE. Referring to Fig. 1, this corresponds to testing whether the vignettes can in fact be depicted as flat horizontal lines, representing the same position on the latent (vertical) spectrum for all populations.

Bago D'Uva et al. (2011b) found strong evidence that VE was violated in the ELSA vignettes. Given the recency of that article, however, the method has yet to be widely applied.

Testing Response Consistency

Response consistency, too, has proven challenging to test rigorously, especially because assessing whether respondents rate vignette characters as they rate themselves depends on availability of data capturing respondents' "true" (objective) level of health. Initial tests of RC have been relatively informal. King et al. (2004) showed that vignette-adjusted self-ratings of vision corresponded better than unadjusted self-ratings with objective vision, but the strength of this correlation was not scrutinized. Grol-Prokopczyk et al. (2011) took a similar approach. Some researchers have conducted more compelling tests of RC, but with limited generalizability or feasibility. For example, van Soest et al.'s (2011) assessment hinges on a unique property of drinking behavior (that alcohol consumption can be quantified as number of drinks consumed; most health domains defy such straightforward quantification), and Kapteyn's (2010:S207) test requires at least two waves of data collection (with vignettes at Time 2 constructed from information from Time 1).

However, Bago D'Uva et al. (2011b) also propose a relatively feasible approach to testing RC: namely, to compare the locations of cutpoints estimated from vignette

ratings with the locations of cutpoints estimated from self-ratings (paired with objective measures of health). If the two sets of cutpoints line up closely, this supports the assumption of RC, as it shows that vignette-ratings and self-ratings use similar standards of evaluation. The authors' results—unlike those of most earlier studies—indicate that RC is violated. Au and Lorgelly's (2014) interview-based findings also suggest that violations of RC are common.

Project Goals

This article assesses the validity of the most widely fielded health vignettes in the world. Specifically, we use WHO data from 10 geographically and socioeconomically diverse countries, as well as data from the HRS, to conduct two tests of vignette equivalence: "weak tests" based on rank-orderings of vignettes, and "strong tests" based on the Bago D'Uva et al. (2011b) test of perceived vignette locations. By conducting both, we assess whether weak and strong tests of VE yield similar results. Where data permit, we also conduct a version of Bago D'Uva et al.'s (2011b) test of response consistency. We seek to clarify whether these health vignettes function as intended, and thus whether they can enhance comparability of self-reported health.

Data and Methods

Data Sets and Variables

Core data for our study come from the 2007–2009 (Wave 1) WHO Study on Global AGEing and Adult Health (SAGE), which comprises nationally representative samples of older adults from six countries: China, Ghana, India, Mexico, Russia, and South Africa (combined n = 44,089; Table 1 describes individual country samples). SAGE enables testing of response consistency for vision and mobility, as it includes relatively objective measures of these domains: distance vision scores (we use the higher from left and right eye LogMAR scores), self-reports (yes/no) of cloudy vision and of glares/halos, scores from two timed walks (regular and rapid pace), and interviewers' assessments (yes/no) of whether respondents had difficulty walking.

Because SAGE includes only low- and middle-income countries, we increased the socioeconomic, geographic, and cultural diversity of the sample by also including four countries participating in the 2002 WHO World Health Survey (WHS): Brazil, France, Netherlands, and the United Kingdom (UK) (combined n = 8,299; see Table 1). We thus include at least one country from each major region of the Inglehart-Welzel Cultural Map of the World (Inglehart and Welzel 2005:64). The diversity of this sample allows us to put vignette equivalence to a particularly rigorous test. Due to a lack of appropriate data in the WHS, however, RC can be tested only with SAGE data.

SAGE and WHS surveys included identical vignettes for eight health domains: pain, mobility, depression, social relationships, distance vision, sleep, memory,

JmlID 13524_ArtID 422_Proof# 1 - 20/08/2015

t1.1 **Table 1** Descriptive statistics for analytic samples

t1.2		WHO Total	Ghana	India	South Africa	China	Brazil	Russia	Mexico	UK	France	Nether-lands	USA
t1.3	Sample Size[a]	52,388	5,565	12,198	4,225	15,009	5,000	4,350	2,742	1,200	1,008	1,091	4,528
t1.4	Sex												
t1.5	Male (%)	42.57	50.60	38.60	42.53	46.59	43.76	35.59	38.29	36.83	40.08	32.54	39.44
t1.6	Female (%)	57.43	49.40	61.40	57.47	53.41	56.24	64.41	61.71	63.17	59.92	67.46	60.56
t1.7	Age												
t1.8	18–49 (%)	27.05	15.08	41.38	9.11	10.94	70.16	9.59	15.65	49.71	67.23	57.75	3.05
t1.9	50–59 (%)	29.90	33.85	26.06	40.12	38.69	13.48	33.75	15.84	13.62	15.59	18.24	30.10
t1.10	60–69 (%)	22.64	23.46	20.13	29.16	26.44	9.08	24.62	34.05	15.79	7.75	15.77	33.55
t1.11	70–79 (%)	15.12	19.25	9.41	15.67	18.67	5.72	23.40	22.60	14.12	6.45	7.33	21.07
t1.12	80+ (%)	5.29	8.36	3.01	5.94	5.26	1.56	8.64	11.85	6.77	2.98	0.92	12.23
t1.13	Education, Highest Level												
t1.14	No formal schooling (%)	26.42	50.74	45.24	24.09	23.91	12.38	0.94	17.15	0.75	1.09	1.37	0.24
t1.15	Did not complete primary school (%)	13.89	10.62	10.48	23.86	16.61	16.98	1.70	36.82	0.33	1.39	0.09	1.86
t1.16	Primary school (%)	17.18	12.43	15.29	23.58	19.10	27.86	7.32	22.52	2.42	14.38	8.07	1.90
t1.17	Secondary school (%)	16.18	5.50	12.42	14.80	21.28	14.88	18.14	10.61	50.83	21.83	7.06	10.78
t1.18	High school (%)	17.93	17.08	10.66	8.15	13.87	21.34	51.76	3.80	17.00	27.68	59.85	61.13
t1.19	College or more (%)	8.40	3.64	5.91	5.51	5.24	6.56	20.14	9.09	28.67	33.63	23.56	24.10
t1.20	Race/Ethnicity												
t1.21	White, non-Hispanic (%)	—	—	—	—	—	—	—	—	—	—	—	79.84
t1.22	Black, non-Hispanic (%)	—	—	—	—	—	—	—	—	—	—	—	11.62
t1.23	Hispanic (%)	—	—	—	—	—	—	—	—	—	—	—	8.55

Notes: Data are from the WHO SAGE Wave 1 (Ghana, India, South Africa, China, Russia, and Mexico) and WHS (Brazil, UK, France, and the Netherlands), and from the HRS (USA). WHO countries are listed in reverse order of Human Development Index (United Nations Development Programme 2008:229–232).

[a] Because of subsampling in the WHO surveys, the sample size for any given domain of health vignettes was approximately 25 % of the total sample size shown. HRS did not subsample vignettes.

and self-care.[3] Online Resource 1 presents vignette texts for select domains. In this article, higher severities indicate worse health; thus, Severity 1 describes the healthiest vignette character in a series, and Severity 5 the least healthy. Due to 25 % subsampling, the size of the WHO analytic sample for each domain was just over 12,000.

Our final source of data was the (American) Health and Retirement Study (HRS) 2007 Disability Vignette Study (n = 4,528), which drew inspiration from the WHO vignettes but included only five domains (pain, mobility, depression, sleep, and memory); used only three vignettes per series; and often slightly modified the wording of WHO vignettes (see Online Resource 2). It was thus not possible to include HRS vignettes in the international analyses. Instead, we analyzed HRS data separately, and focused on VE across key demographic categories (age, sex, education, and race/ethnicity).

Both SAGE and HRS were designed as surveys of aging, and thus focus on adults older than age 50. SAGE included some respondents under age 50 for comparison, with this proportion varying from 9 % for South Africa to 41 % for India. HRS included respondents aged 50 and younger only if they were spouses of older respondents; the proportion of younger respondents is thus only 3 %. WHS surveys included proportionate representation of adults aged 18 and older. As explained later, our findings were insensitive to these differences in age distributions.

Table 1 provides descriptive characteristics for the analytic samples. In our analyses, respondent nationality and demographic characteristics were indicated with the dummy or categorical variables shown in Table 1, with the exceptions that (1) because of small cell size in some countries, the two lowest educational categories were combined in the WHO analyses; (2) HRS analyses used a four-category educational grouping: "No degree" (14.77 %), "High school diploma" (61.13 %), "College degree" (13.94 %), and "Graduate degree" (10.16%); and (3) HRS analyses treated "Under 60" (33.15%) as a single age category. The original HRS data included 100 respondents self-identifying as "Other, non-Hispanic"; due to their small number, these were excluded, yielding the shown sample size of 4,528.

Analytic Strategy and Models: Vignette Equivalence

We conducted weak tests and strong tests of vignette equivalence. Weak tests were based on respondents' rank-orderings of vignettes, to assess whether respondents perceived the five (in SAGE/WHS) or three (in HRS) severity levels in the expected order. The percentage of respondents showing the expected rank-ordering was calculated by country or subgroup. Ties in ratings were assumed to resolve consistently with the expected ordering, as in Murray et al.'s (2003:376) "benefit-of-the-doubt" calculations.

The stronger test of VE, following Bago D'Uva et al. (2011b), is based on a likelihood-ratio (LR) test comparison of two models, A and B. In each model, R_{ij} denotes respondent i's rating of vignette j, and V_{ij} is the unobserved perceived level of health of the vignette j character in the opinion of respondent i. The link between the observed discrete variable R_{ij} and the unobservable (latent) continuous variable V_{ij} is determined by the cutpoints (τs) as $R_{ij} = k$ if and only if V_{ij} is between cutpoints τ_i^{k-1} and τ_i^k. The

[3] In some surveys, two closely related evaluation questions (e.g., regarding "pain" and "discomfort") followed each vignette. Here, we present one question from each pair, given that pairs yielded extremely similar ratings, and European surveys included only the first question.

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

cutpoints are assumed to monotonically increase[4] between $\tau_i^0 = -\infty$ and $\tau_i^K = \infty$, where K denotes the number of available response categories; here, K=5. In Model A, the distribution of each vignette j's perceived location V_{ij} is assumed to be independent of all covariates, that is, each vignette location can be represented simply as a constant (α_j) plus a random error term (ε_{ij}; assumed to be normally distributed with mean zero):

$$\textbf{Model A}: \quad V_{ij} = \alpha_j + \varepsilon_{ij}$$

For model identification, α_1 is set to 0, and the variance of the random error term is set to 1.

In Model B, a selected reference vignette is set to a constant (0), as in Model A, but all other vignettes may now have their positions affected by a vector of covariates ($\mathbf{X}_i$), which include sex, age, education, and either country (in the international analyses) or race/ethnicity (in the HRS analyses):

$$\textbf{Model B}: \quad \begin{aligned} &\text{As in Model A for reference vignette, but} \\ &V_{ij} = \alpha_j + \lambda_j \mathbf{X}_i + \varepsilon_{ij} \quad \text{for all other vignettes,} \end{aligned}$$

where the covariate vector $\mathbf{X}$ takes a linear functional form and does not include a constant term.

If vignette equivalence holds, then $\lambda_j = 0$ for all j, so that Model B reduces to Model A. This is consistent with an LR test failing to reject the hypothesis of no difference between models. If, however, the LR test rejects this hypothesis (i.e., yields $p < .05$), we interpret this as a rejection of VE, as it indicates that groups differ in where they perceive vignettes to lie on the latent health spectrum. (Online Resource 3 provides additional details about the LR test and likelihood function for Models A and B.) Following Bago D'Uva et al. (2011b), we refer to this model comparison as the "global test" of VE. Because of our large sample sizes, even substantively small violations of VE could lead to rejection of model equivalence. Thus, we assess VE based not only on the statistical significance of the global tests but also on the magnitude of the violations.

Concretely, Models A and B were implemented by variations on the hierarchical ordered probit (hopit) model common in vignette studies (e.g., Rabe-Hesketh and Skrondal 2002).,[5] [6] Unlike standard ordered probit models, which assume fixed response-category cutpoints, hopit models allow cutpoints to vary across groups (based on ratings of anchoring vignettes). These calculated differences in cutpoints are then accounted for in a second set of calculations, which, in the cases of Models A and B, estimate perceived vignette locations. In both models, we allow cutpoints to vary by sex, age, education, and country (for SAGE/WHS) or race/ethnicity (for HRS). However, in Model A, only dummy variables for vignette severity enter into the equation for

[4] To ensure sequential increases in cutpoints, exponential coding is used: that is, $\tau_i^1 = \gamma_1 \mathbf{X}_i$ and $\tau_i^k = \tau_i^{k-1} + \exp(\gamma_k \mathbf{X}_i)$, $k = 2, \ldots, K-1$ (as in, e.g., van Soest and Vonkova (2014)). Note that in the cutpoint parametrization, the covariate vector $\mathbf{X}$ includes a constant term.

[5] Some refer to this as "chopit" (with "c" standing for "compound"; Rabe-Hesketh and Skrondal (2002)); others use "chopit" only when multiple ratings of each vignette enable calculation of individual-level random effects. We do not calculate random effects, so use "hopit" to avoid ambiguity.

[6] van Soest and Vonkova (2014) present an extension of the hopit model allowing for unobserved heterogeneity, and recommend other model variants as well, which future researchers may wish to consider. We are confident that our main (parametric) conclusions regarding VE are not artifacts of modeling assumptions, however, since our entirely nonparametric weak tests support the same conclusions.

perceived vignette locations. In contrast, in Model B, the equation also includes multiple terms representing the interaction between a given severity and a covariate. For example, the "Severity 1 × female" interaction indicates whether the perceived distance between the Severity 1 (least severe) vignette and the reference vignette was different for women than for men. Such interactions were included for each severity crossed with each covariate (excluding omitted categories). These interaction terms indicate which covariates drive violations of VE. Visually, significant interactions indicate that vignettes *cannot* be depicted as flat horizontal lines across groups.

Analytic Strategy and Models: Response Consistency

Bago D'Uva et al. (2011b) propose an LR-based global test of response consistency, which compares a model estimating intercategory cutpoints via vignettes with a model estimating them via objective measures of health. However, this test depends on vignette equivalence; the LR test will be rejected if RC *or* VE is violated. Given our upcoming findings regarding VE, this global test was not appropriate here. Instead, we use a somewhat less-stringent test suggested in the same article (2011b:884), namely, to graph cutpoints generated from vignette ratings next to cutpoints generated from objective measures of health (paired with self-ratings), and then visually compare the two. Observing similar "shapes" of cutpoints in both models would indicate that similar standards of evaluation are used for vignette- and self-ratings, and thus would be supportive of RC (with the caveat that the relative positions of the two sets of cutpoints along the latent spectrum cannot be definitively determined).

Concretely, to estimate intercategory cutpoints from vignette ratings, we used hopit Model A (described earlier), except instead of presenting estimated vignette locations, we present estimated cutpoint locations. To estimate intercategory cutpoints from (relatively) objective measures of health, we used a third form of hopit, Model C, which is identical to Model A except that it estimates cutpoints by pairing self-ratings of health with objective measures of health (instead of pairing vignette-ratings with vignette severities).

We tested RC for two domains, distance vision and mobility, because SAGE includes relatively objective measures of these (see the earlier "Data Sets and Variables" section). Given that these measures are unlikely to fully capture true health, we would consider high, even if imperfect, concordance between vignette-generated and health measure-generated cutpoints to be encouraging regarding RC.

The Stata 13 code used to generate Models A–C and all other code for this project is available as Online Resource 4.

Results

Results: Weak Tests of Vignette Equivalence

Table 2 shows that the percentage of respondents who ranked the WHO vignettes correctly—that is, consistently with the expected order—ranged from 44.39 to 69.63 %, depending on domain. Examining countries individually, correct rank-orderings ranged from 29.94 % (for Mexicans' ratings of sleep vignettes) to 84.18 % (for Russians' ratings of memory). Although some variation in orderings is expected due to measurement error, and no precise cut-off for acceptable rates of correct rank-ordering has been established

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

t2.1 **Table 2** Percentage of respondents ordering vignettes consistently with expected ordering

t2.2	Domain	% Consistent, Full WHO Sample (n = 52,388)	Range (as %) Across WHO Countries	% Consistent, HRS Sample (n = 4,528)
t2.3	Pain	63.38	39.48–79.67	90.75
t2.4	Mobility	69.63	47.91–80.40	78.00
t2.5	Depression	68.58	48.99–82.26	85.84
t2.6	Relationships	48.58	33.20–64.35	—
t2.7	Distance Vision	55.37	40.52–68.37	—
t2.8	Sleep	44.39	29.94–55.38	70.55
t2.9	Memory	66.36	41.50–84.18	91.39
t2.10	Self-care	51.58	38.56–71.94	—

Notes: WHO vignette series consist of five vignettes each, subsampled at 25 %. HRS series consist of three vignettes, with wording close to but often not identical to WHO vignettes.

(Rice et al. 2011:147), one could argue that percentages below 80 % are worrisome, and those below 60 % are dire, or even ruinous.

Space restrictions prevent detailed analysis of rank-orderings in all 80 country-domain pairings (data available upon request), but in 40 of these—a full half—fewer than 60 % of respondents gave concordant rankings. Only in five of the 80 did the percentage exceed 80 %. Substantial rank-order violations were not restricted to particular domains or countries (although violations were particularly common among Mexicans, who showed concordance below 50 % in every domain). There was no discernible association between countries' level of socioeconomic development and rates of misordering: for example, Ghana's overall percentage of correct orderings (58.94 %) was scarcely different from the Netherlands' (60.36 %). Moreover, different domains performed particularly poorly in different countries. In other words, the violations of VE revealed here appear due neither to a few isolated "bad vignettes", nor to respondent incomprehension in particular countries. Rather, variation in interpretation of vignettes appears to be high both within and across all examined countries.

Rank-order violations were less frequent in the HRS (Table 2, right). The percentage of correct rank-orderings was above 85 % for pain, depression, and memory; and between 70 % and 80 % for sleep and mobility (similar to van Soest and Vonkova's (2014:122–123) figures based the same vignettes in SHARE).

Results: Strong Tests of Vignette Equivalence

Table 3 presents the results of the global test of VE. As shown, the assumption of VE was rejected ($p < .001$) for all series of vignettes, in both WHO and HRS data. Alternate versions of Model B including subsets of covariates were also tested. For four WHO vignette series (mobility, depression, sleep, and memory), VE was not rejected in models including respondent sex as the only covariate. In all other specifications, VE was consistently rejected.

Results from Models B indicate which demographic variables drive the global rejection of VE in Table 3. Due to space limits, we focus on the example of pain. Table 4 shows predictors of perceived vignette position (i.e., location on the latent health spectrum) for

t3.1 **Table 3** Global tests of vignette equivalence

t3.2		WHO (SAGE/WHS)		HRS	
t3.3	Domain	Degrees of Freedom	LR Test Statistic	Degrees of Freedom	LR Test Statistic
t3.4	Pain	72	4,274.07***	20	428.41***
t3.5	Mobility	72	3,028.75***	20	249.60***
t3.6	Depression	72	4,355.36***	20	710.15***
t3.7	Relationships	72	3,443.05***	—	—
t3.8	Distance vision	72	4,115.94***	—	—
t3.9	Sleep	72	2,728.08***	20	497.79***
t3.10	Memory	72	7,235.53***	20	387.97***
t3.11	Self-care	72	2,762.60***	—	—

Notes: Test is based on likelihood ratio (LR) comparison of Models A and B, described in the text. Covariates, interacted with vignette severities, are sex, age, education, and country for WHO analyses; and sex, age, education, and race/ethnicity for HRS analyses.

***$p < .001$

pain vignettes, estimated from the WHO data. Positive coefficients for interaction terms indicate that the given group perceives the given vignette to be farther from the reference vignette (the vignette representing the worst health—vignette 5 in WHO data and vignette 3 in HRS). Thus, in Table 4, the positive, statistically significant interactions between female sex and each vignette severity indicate that compared with men, women see vignettes 1–4 as being farther from the reference vignette (i.e., as representing comparatively better health—here, relatively less pain). Similarly, respondents who completed high school or college perceived these vignettes as being more distant from vignette 5. Respondent age, in contrast, did not significantly predict pain vignettes' relative locations on the latent spectrum. The largest coefficients in the model, for all four severities, are for country interactions. Cross-national differences in understandings of vignettes thus often appear substantially larger than differences across sex, age group, or educational category. As discussed shortly, this is true across all tested health domains.[7]

Similar analyses of other WHO vignette series reveal that the effects of sex, age, and education were inconsistent across domains, and thus cannot be easily summarized (results available upon request). For example, while women perceived pain vignettes 1–4 to be farther from the reference vignette than did men (Table 4), respondent sex appeared unrelated to perceived vignette location for mobility. Conversely, significant age effects were found in the mobility series, but not the pain series. However, across all domains, cross-national differences in understandings of vignettes were consistently both statistically significant and substantively large—indeed, constituting the largest coefficients in their respective models. This suggests that cross-national vignette-based comparison is particularly fraught. We hypothesize that similar issues would arise across culturally or linguistically distinct groups within a country (e.g., immigrant groups; see Pan and Fond 2014).

[7] One-way analysis of variance (ANOVA) confirms that between-country variation in perceived vignette locations dwarfs within-country variance, in all WHO vignette series ($p < .001$).

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

t4.1 **Table 4** Predictors of perceived vignette position for pain vignettes, WHO analysis

t4.2		Ordered Probit	
t4.3		β	SE
t4.4	Severity 1	3.50***	0.08
t4.5	Severity 2	2.38***	0.07
t4.6	Severity 3	1.53***	0.07
t4.7	Severity 4	0.74***	0.07
t4.8	Sev 1 × Female	0.23***	0.04
t4.9	Sev 1 × Age 50–59	–0.08	0.05
t4.10	Sev 1 × Age 60–69	–0.07	0.05
t4.11	Sev 1 × Age 70–79	–0.07	0.06
t4.12	Sev 1 × Age 80+	–0.11	0.09
t4.13	Sev 1 × Less Than Primary School	0.05	0.05
t4.14	Sev 1 × Secondary Completed	0.11	0.06
t4.15	Sev 1 × High School Completed	0.36***	0.06
t4.16	Sev 1 × College Completed	0.72***	0.08
t4.17	Sev 1 × India	–1.34***	0.07
t4.18	Sev 1 × South Africa	0.00	0.09
t4.19	Sev 1 × China	0.30***	0.07
t4.20	Sev 1 × Brazil	0.30**	0.09
t4.21	Sev 1 × Russia	1.01***	0.10
t4.22	Sev 1 × Mexico	–2.21***	0.08
t4.23	Sev 1 × UK	0.90***	0.15
t4.24	Sev 1 × France	–0.10	0.15
t4.25	Sev 1 × Netherlands	–0.55***	0.13
t4.26	Sev 2 × Female	0.12***	0.03
t4.27	Sev 2 × Age 50–59	–0.03	0.05
t4.28	Sev 2 × Age 60–69	–0.01	0.05
t4.29	Sev 2 × Age 70–79	–0.06	0.06
t4.30	Sev 2 × Age 80+	–0.08	0.08
t4.31	Sev 2 × Less Than Primary School	0.02	0.05
t4.32	Sev 2 × Secondary Completed	0.07	0.05
t4.33	Sev 2 × High School Completed	0.15**	0.06
t4.34	Sev 2 × College Completed	0.32***	0.07
t4.35	Sev 2 × India	–0.98***	0.06
t4.36	Sev 2 × South Africa	–0.25**	0.08
t4.37	Sev 2 × China	0.02	0.06
t4.38	Sev 2 × Brazil	–0.15*	0.08
t4.39	Sev 2 × Russia	0.36***	0.09
t4.40	Sev 2 × Mexico	–1.64***	0.08
t4.41	Sev 2 × UK	0.88***	0.14
t4.42	Sev 2 × France	–0.18	0.13
t4.43	Sev 2 × Netherlands	1.16	0.12
t4.44	Sev 3 × Female	0.08*	0.03

t4.45 **Table 4** (continued)

t4.46		Ordered Probit β	SE
t4.47	Sev 3 × Age 50–59	–0.04	0.04
t4.48	Sev 3 × Age 60–69	–0.03	0.05
t4.49	Sev 3 × Age 70–79	–0.08	0.05
t4.50	Sev 3 × Age 80+	–0.05	0.08
t4.51	Sev 3 × Less Than Primary School	0.06	0.04
t4.52	Sev 3 × Secondary Completed	0.09	0.05
t4.53	Sev 3 × High School Completed	0.11*	0.04
t4.54	Sev 3 × College Completed	0.24***	0.07
t4.55	Sev 3 × India	–0.32***	0.06
t4.56	Sev 3 × South Africa	–0.02	0.08
t4.57	Sev 3 × China	0.77***	0.06
t4.58	Sev 3 × Brazil	0.16*	0.07
t4.59	Sev 3 × Russia	0.78***	0.08
t4.60	Sev 3 × Mexico	–0.90***	0.08
t4.61	Sev 3 × UK	0.73***	0.13
t4.62	Sev 3 × France	0.06	0.12
t4.63	Sev 3 × Netherlands	–0.08	0.11
t4.64	Sev 4 × Female	0.06*	0.03
t4.65	Sev 4 × Age 50–59	–0.02	0.04
t4.66	Sev 4 × Age 60–69	–0.02	0.05
t4.67	Sev 4 × Age 70–79	0.01	0.05
t4.68	Sev 4 × Age 80+	–0.01	0.07
t4.69	Sev 4 × Less Than Primary School	–0.01	0.04
t4.70	Sev 4 × Secondary Completed	0.05	0.05
t4.71	Sev 4 × High School Completed	0.14**	0.05
t4.72	Sev 4 × College Completed	0.25***	0.07
t4.73	Sev 4 × India	–0.17**	0.06
t4.74	Sev 4 × South Africa	0.04	0.07
t4.75	Sev 4 × China	1.04***	0.06
t4.76	Sev 4 × Brazil	0.29***	0.07
t4.77	Sev 4 × Russia	0.41***	0.08
t4.78	Sev 4 × Mexico	–0.55***	0.07
t4.79	Sev 4 × UK	0.74***	0.12
t4.80	Sev 4 × France	0.42***	0.12
t4.81	Sev 4 × Netherlands	0.11	0.11

Notes: Results are from Model B hopit regression (n = 12,380). Perceived position of vignettes is calculated relative to the Severity 5 vignette. Other omitted reference categories are male (for sex), under age 50 (age), primary school completed (education), and Ghana (country). Countries are listed in reverse order of Human Development Index (United Nations Development Programme 2008:229–232).

*p < .05; **p < .01; ***p < .001 (two-tailed tests)

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

Graphs of perceived vignette locations by country provide a clearer sense of the extent to which VE is violated cross-nationally. In the following figures, if VE were perfectly upheld, the bars for each severity would be exactly the same height across all 10 countries, so that each figure would resemble four flat tabletops. In actuality, however, when coefficients from Model B (Table 4) are applied to the WHO sample to predict perceived vignette locations, the resulting figures take a very different shape.

As Fig. 2 shows, differences in perceived pain vignette locations across countries are enormous—often on the scale of 1, 2, or even 3 standard deviations of the reference vignette (the *y*-axis unit). That is, residents of different countries appear to interpret the relative severity of the pain described in the vignettes in dramatically different ways. These findings are not driven by (indeed, are scarcely affected by) differences in national age distributions, as supplementary analyses confirm (not shown). Graphs restricted to respondents aged 50 and older are visually indistinguishable from those shown here.

Moreover, the bumpiness of Fig. 2 was replicated across all domains, although there was variation in which countries served as peaks or valleys: see, for example, graphs for mobility (Fig. 3) and distance vision (Fig. 4). Countries were listed in reverse order of Human Development Index (HDI) to see whether interpretation of vignettes was correlated with socioeconomic development, but no such association emerged. Despite these large cross-national differences—and the numerous rank-order violations in individual respondents' ratings, discussed earlier—mean vignette locations within a country were usually in the expected order.

In several domains, including pain, Mexico appears to be an outlier, with Mexicans locating vignettes much closer together on the latent spectrum than do other respondents. Although this result may genuinely correspond to Mexican understandings of vignettes, close review of the data suggests another possibility, namely, that Mexicans misinterpret the "Extreme/Cannot do" response category. "Cannot do" is intended to describe vignette characters' limitations (e.g., a blind character might elicit a rating of "Cannot do" regarding her capacity to see things), but respondents might instead interpret it as describing their own capacity to answer the question: that is, "I cannot do this question." Mexicans choose "Extreme/Cannot do" much less frequently than do other respondents (e.g., only 14 % of Mexicans rate the Severity 5 pain vignette with "Extreme/Cannot do," versus 49 % of all other WHO respondents). Because requests to view local-language

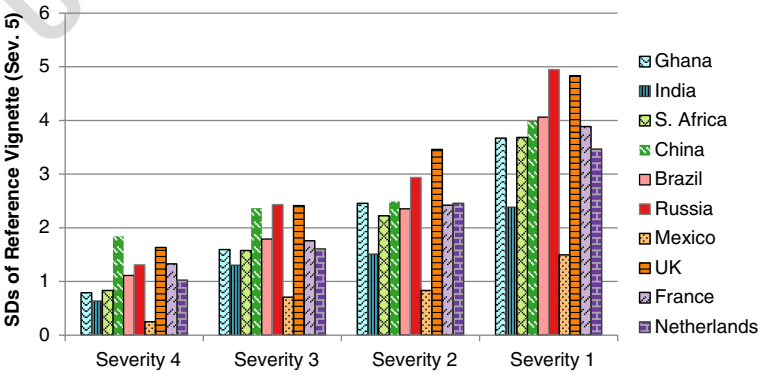

Fig. 2 Estimated pain vignette locations (on latent health spectrum; relative to Severity 5), WHO data. Zero on the *y*-axis represents the mean of the reference (least healthy) vignette; higher numbers represent better perceived health

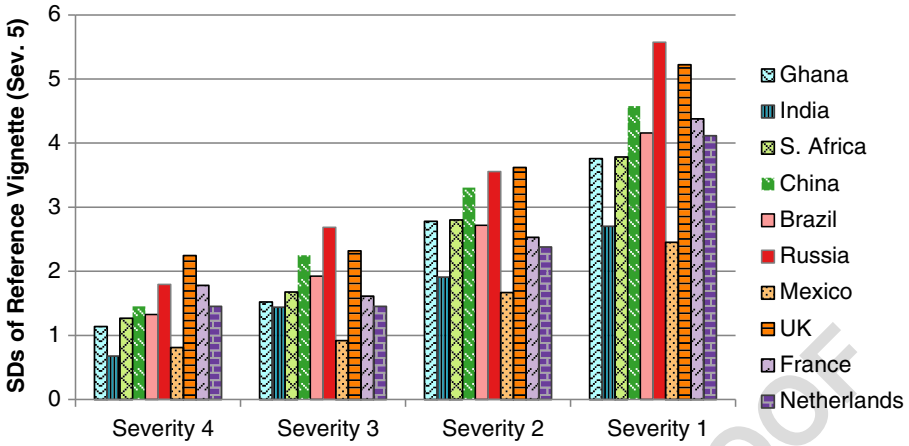

Fig. 3 Estimated mobility vignette locations (on latent health spectrum; relative to Severity 5), WHO data. Zero on the *y*-axis represents the mean of the reference (least healthy) vignette; higher numbers represent better perceived health

versions of the SAGE survey have been unsuccessful, it is unclear whether or why such an interpretation should be more common among Mexicans (or whether other country surveys were also affected). However, if this misinterpretation is in fact widespread among Mexican respondents, then they are effectively working with a truncated set of response categories, which would lead to reduced dispersion of vignette locations.

Although VE is unambiguously violated across the highly diverse countries in this sample, specific subsets of countries violate VE less egregiously. For example, reexamining Fig. 2, one can see that Ghana and South Africa show substantively very minor discrepancies in perceived pain vignette locations; Brazil, France, and the Netherlands also appear to interpret the pain vignettes as representing similar levels of pain. Though no pairing of countries in this set of five actually passes the global test of VE for pain, the degree of violation might be forgivable, depending on the application.

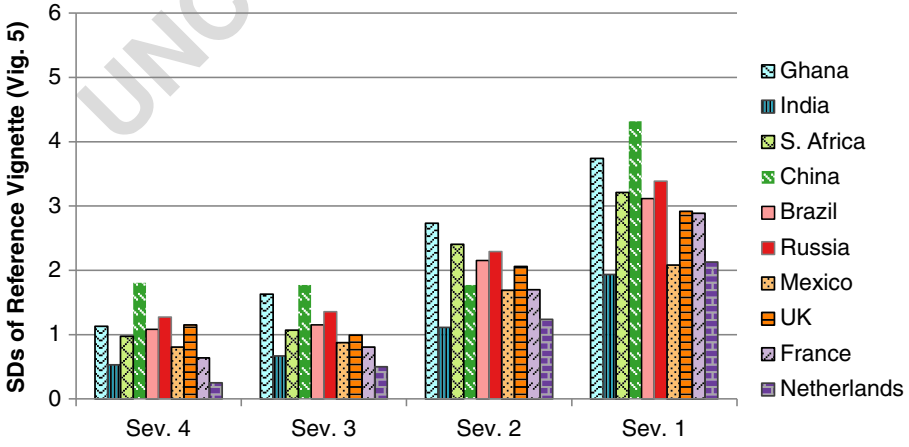

Fig. 4 Estimated distance vision vignette locations (on latent health spectrum; relative to Severity 5), WHO data. Zero on the *y*-axis represents the mean of the reference (least healthy) vignette; higher numbers represent better perceived health

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

For other WHO vignette series, the global test occasionally fails to reject VE for specific pairings of countries: Ghana and South Africa in the mobility vignette series, and France and the Netherlands in the sleep and self-care series. Such concordant pairings of countries were rare, however.

Violations of VE across demographic groups in the HRS were less dramatic than cross-national ones in SAGE/WHS, but not negligible. Table 5 shows that women and more highly educated respondents perceived a significantly greater distance between the Severity 1 and 3 pain vignettes, and that nonwhites and older respondents perceived a significantly lesser distance. The distance along the latent health spectrum between Severity 2 and 3 pain vignettes was seen as significantly greater by respondents 80+ and Hispanics.

Figure 5 (applying coefficients from Model B to the HRS sample; n = 4,258) presents these findings visually, showing nontrivial differences in perceived pain vignette locations by level of education and between whites and nonwhites. It should be underscored that choice of reference vignette is arbitrary. Thus, one cannot conclude from the relatively flat appearance of the Severity 2 bars in Fig. 5 that only the Severity 1 vignette is problematic, because when a different reference vignette is chosen, vignette nonequivalence manifests itself through different contrasts (graphs available upon request). In other words, one cannot simply discard certain vignettes and salvage the rest, since the Bago D'Uva method provides no mechanism for identifying the most problematic vignettes in a series (assuming that such vignettes exist; nonequivalence could stem from different interpretations of *all* vignettes in a series). The method can diagnose nonequivalence, but cannot cure it.

Results: Tests of Response Consistency

As described earlier, our assessment of RC is based on a visual comparison of (1) cutpoints generated from anchoring vignette ratings (Model A), and (2) cutpoints generated from self-ratings paired with objective measures of health (Model C),[8] using six-country SAGE data. As Fig. 6 shows, the cutpoints predicted by the two models look extremely similar in a full sample analysis, for both distance vision and mobility. For vision, the slope for the health measure–based cutpoints is only slightly higher than that for the vignette-based cutpoints; for mobility, the difference is even slighter (in the other direction). Although calculated from entirely different types of data, the two sets of cutpoints show impressively concordant shapes, consistent with the assumption of RC.

These full-sample data mask some heterogeneity among countries. For example, although the two sets of distance vision cutpoints show near perfect concordance for India, they are obviously misaligned for Russia, with other countries falling at various points in between. At the same time, Russia's mobility vignette results are largely congruent (data and graphs available upon request). Response consistency for a given domain may thus be more problematic in some regions than in others, and may also vary across health domains for a given country. Despite some exceptions, however, most countries in our sample show close alignment of cutpoints in both tested domains.

[8] To align and facilitate comparison of the two sets of bars, Model C units (standard deviation of the self-rating) were converted to Model A units (standard deviation of the reference vignettes), and a constant was added to Model C's predicted cutpoints. Graphs reflect these conversions.

Table 5 Predictors of perceived vignette position for pain vignettes, HRS analysis

	Ordered Probit	
	β	SE
Severity 1	2.56***	0.11
Severity 2	0.92***	0.09
Sev 1 × Female	0.31***	0.06
Sev 1 × Age 60–69	–0.17*	0.08
Sev 1 × Age 70–79	–0.24**	0.09
Sev 1 × Age 80+	–0.34**	0.10
Sev 1 × High School Diploma	0.58***	0.08
Sev 1 × College Degree	1.01***	0.12
Sev 1 × Graduate Degree	1.09***	0.14
Sev 1 × Black, non-Hispanic	–0.42***	0.09
Sev 1 × Hispanic	–0.30**	0.11
Sev 2 × Female	0.04	0.05
Sev 2 × Age 60–69	–0.03	0.06
Sev 2 × Age 70–79	0.03	0.07
Sev 2 × Age 80+	0.18*	0.08
Sev 2 × High School Diploma	0.11	0.07
Sev 2 × College Degree	0.14	0.09
Sev 2 × Graduate Degree	0.11	0.10
Sev 2 × Black, non-Hispanic	0.13	0.08
Sev 2 × Hispanic	0.24**	0.01

Notes: Data are from Model B hopit regression (n = 4,528). Perceived position of vignettes is calculated relative to the Severity 3 vignette. Other omitted reference categories are male (for sex), under age 60 (age), no degree (education), and white (race/ethnicity).
*p < .05; **p < .01; ***p < .001 (two-tailed tests)

Thus, with the caveat that the exact vertical alignment of the two sets of cutpoints cannot be guaranteed (as mentioned previously), our findings suggest relatively minor violations of response consistency in SAGE.

Discussion

Anchoring vignettes have been lauded as a simple, inexpensive way to harmonize subjective survey questions, and in the process to enable much-needed comparative

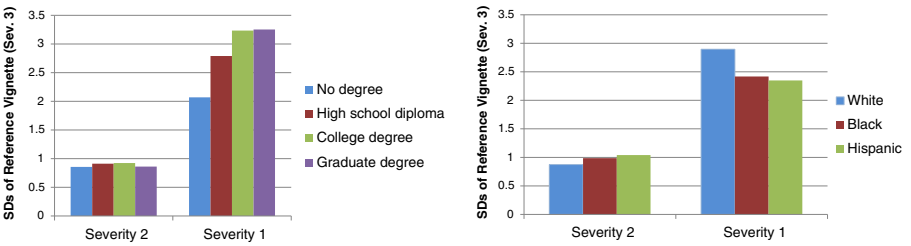

Fig. 5 Estimated pain vignette locations by education and race/ethnicity, HRS data. The zero on the y-axis represents the mean of the reference (least healthy) vignette (Severity 3); higher numbers represent better perceived health

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

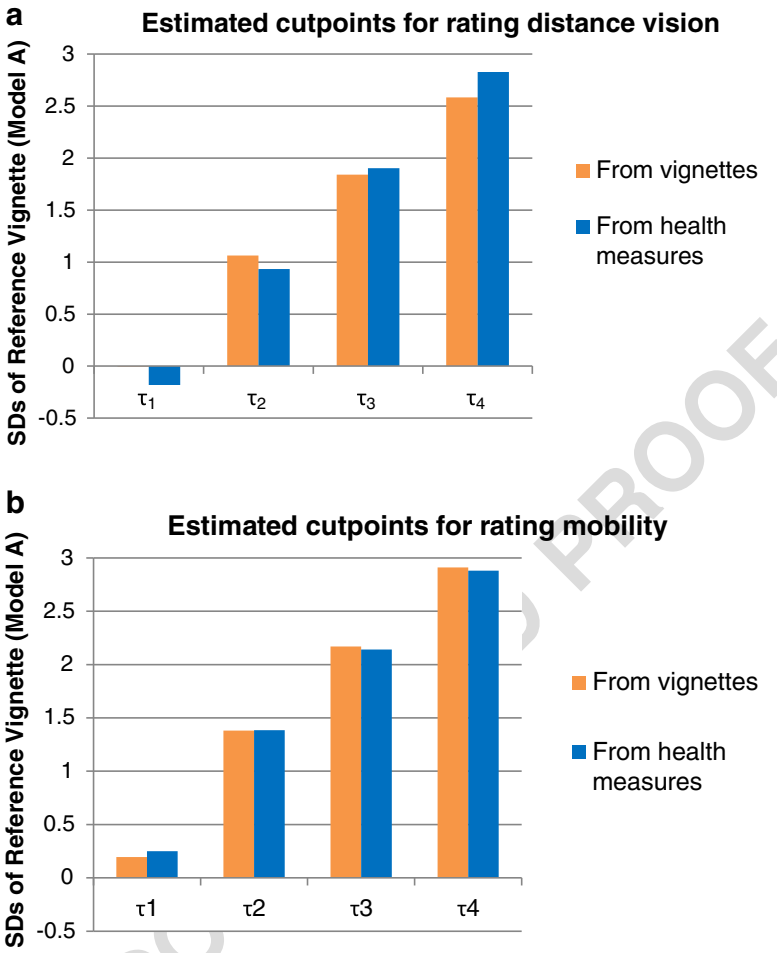

Fig. 6 Estimated cutpoint locations for distance vision and mobility, from vignettes (Model A) and from objective health measures (Model C), full SAGE sample

research (Kapteyn 2010). However, our findings show that existing WHO health vignettes, and some HRS health vignettes, fail weak (rank-order–based) tests of vignette equivalence. Moreover, they routinely—and egregiously—fail stricter tests positing equidistance between latent vignette locations across countries or socioeconomic categories. Respondents in different demographic groups appear to understand vignette texts as representing fundamentally different levels of health, meaning that vignette ratings cannot be used to identify different styles of using response categories. The solution to the identification problem promised by anchoring vignettes is, in these data, discredited. Although our tests of response consistency were more encouraging—often showing a striking concordance between cutpoints generated from vignette- and self-ratings—our findings as a whole undercut the legitimacy of the anchoring vignette method, at least for these vignettes.

That VE is violated in the 10-country WHO data is perhaps unsurprising: the countries were selected for geographic and socioeconomic diversity, in order to

constitute a maximally stringent test of VE. Some subsets of the countries appear to violate VE relatively minimally, and in rare cases, such as specific two-country pairings, VE is not rejected at all. This is consistent with Corrado and Weeks (2010), who used the Bago D'Uva et al. (2011b) technique to assess VE for life satisfaction vignettes from SHARE. Although VE was rejected across the 11 countries as a whole, understandings of vignettes appeared comparable in certain small subsets of countries.

Such findings suggest that there are specific cases in which existing anchoring vignettes could be legitimately used for cross-group comparison, but they also underscore the need to explicitly test VE for each potential analysis, rather than assuming it *a priori* (especially in cross-national contexts). In general, weak tests should not be used in isolation, since violations of VE may not manifest themselves primarily through rank-order inconsistencies; the strong test of Bago D'Uva et al. (2011b) provides valuable additional information.

What might cause violations of VE? Although we earlier identified a potential ambiguity in survey wording ("Cannot do") that, in some translations, might lead to misunderstanding of response categories, it appears unlikely that violations of VE are primarily due to microlinguistic mistranslation. The WHO translation protocol is sufficiently careful (Angel 2013:233) that gross errors in word choice or syntax are likely rare. However, as discussed in our overview of VE, grammatically correct renditions of source text do not guarantee cross-group conceptual equivalence. In Pan and Fond's (2014) schema, good translation requires attending not only to local (1) linguistic rules, but also to (2) cultural norms and (3) social practices. Doing so may require substantially deviating from the source text's word order or even its content, as additional information or alternate examples may be required to achieve "functional equivalence" across groups. This is because, for example, health symptoms may be expressed differently in different cultures, reflecting local "idioms of distress" (Angel 2013:233); and because concepts may have different connotations across cultures (see, e.g., Pan and Fond (2014:187) on Vietnamese speakers' interpretation of "nursing home" as "a luxurious resort").

In short, adherence to local linguistic rules does not correct or compensate for culturally specific content. Yet existing WHO vignettes seem in numerous ways to invite different interpretations across national, religious, and/or socioeconomic groups. The description of pain caused by excessive computer use (pain vignette 3) may have different meaning in a technology-based economy than in one in which computer work is rare. The mention in four vision vignettes of reading (e.g., newspapers) may elicit different interpretations in countries with dramatically different literacy rates. Similar examples include the mention of suicide in the pain and depression series, of obesity in the mobility series, of exercise in the pain and mobility series, of hospital admissions in the depression series, and of stroke in the relationships series (cf. Grol-Prokopczyk et al.'s (2011) argument against mentioning specific diseases in vignettes). Even if WHO translations are grammatically correct, they appear to inadequately account for local cultural contexts, which may ascribe very different meanings to ostensibly similar vignette descriptions. The result is a failure to achieve functional equivalence—or, in this study's terms, vignette equivalence.

What do these findings imply for researchers considering anchoring vignettes? Those developing vignettes *de novo* (or modifying existing vignettes) are potentially in the best position. They can incorporate recent findings on improving vignette implementation (e.g., Grol-Prokopczyk (2014) on presenting characters' age and sex;

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

Hopkins and King (2010) on placement of vignettes *vis-à-vis* self-assessments). More crucially, they can strive to optimize vignette content to maximize vignette equivalence.[9] Indeed, attending closely to details of wording may be the key to improving the validity of future vignettes. Despite the great importance of vignettes that accurately capture the trait of interest and do so in as universally comprehensible a way as possible, vignette studies to date have almost without exception analyzed vignettes in the aggregate, without examining, comparing, or validating individual vignette texts. Research on anchoring vignettes is dominated by highly statistically oriented scholars. The method, however, represents an opportunity for quantitative researchers to collaborate with experts in translation and in local cultures to generate vignettes that achieve "semantic, conceptual, and technical equivalence" across groups (Skevington 2002:138; cf. Hunt and Bhopal 2004). Recent advances in classifying and overcoming translation problems, based on cognitive interviewing of survey respondents, may help achieve this goal (e.g., Pan and Fond 2014). Admittedly, however, such undertakings are likely to require substantial investment of resources (Pasick et al. 2001).[10]

Researchers conducting secondary data analyses with vignettes might begin by conducting the tests of VE and RC discussed earlier, and hope for no or only minor violations of measurement assumptions. Where measurement assumptions are substantially violated, however, we provisionally argue against use of vignettes. This is because, in cross-national or other cross-group comparisons, experts in local languages or cultures may be able to make educated guesses about the direction of bias in simple self-reports (as when Angel (2013:230) compares the nuances of "fair" in English versus "regular" in Spanish). After vignette-based adjustments are made, however, the direction and extent of bias becomes much less amenable to educated guesswork. It may be preferable to use original self-ratings than to add another, less predictable source of error into the mix.

This is a provisional argument, however, pending stronger evidence to identify which is the lesser of two evils: unadjusted self-ratings, or ratings adjusted via imperfect vignettes. For vision and mobility—domains in which SAGE provides objective measures—we tested whether raw or vignette-adjusted self-ratings yield country rankings that are closer to objective rankings. The results were equivocal: country rankings based on self-ratings were identical to rankings from vignette-adjusted (hopit) models—and both were quite distant from rankings based on objective measures.[11] In these data, then, vignettes appeared neither to help overcome reporting heterogeneity nor to exacerbate the problem. Researchers with access to other objective measures may be able to conduct more definitive assessments of whether vignettes (even flawed ones) lead to more accurate group comparisons than unadjusted self-reports.

[9] To this end, patterns in rank-order violations may have diagnostic utility. For example, in WHO self-care vignettes, 35.71 % of respondents misordered Severities 3 versus 4, while fewer than 10 % misordered all other adjacent vignette pairs. Vignettes 3 and 4 thus particularly invite further investigation and refinement.

[10] A full cost-benefit analysis of anchoring vignettes would consider both challenges of vignette development/assessment and subsequent challenges of analysis. At present, vignette analyses are often time-consuming to run, and typically adjust only dependent variables. Bago D'Uva et al. (2011a:641) reported that theirs was "only the second study" to use vignette-adjusted independent variables.

[11] For example, both raw and vignette-adjusted self-ratings of distance vision yield this ranking of SAGE countries (best to worst; respondents 50+): China, Mexico, Russia, South Africa, Ghana, and India. An objective ranking, based on LogMAR vision tests, is Ghana, South Africa, China, Russia, India, and Mexico.

The present study does not argue for abandonment of the anchoring vignette method, given its potential utility. Rather, it underscores the need for more carefully constructed, culturally sensitive vignettes. However, if concerted efforts to develop valid vignettes fail (or are deemed too challenging or costly), anchoring vignettes may need to lose their status as the "most promising" solution (Murray et al. 2002:429) to reporting heterogeneity. One potential alternative is suggested by Schenker et al. (2010), who use objective (clinical) health information from one survey to improve analyses of self-reported data in a second. A related strategy would be to collect both objective measures and self-ratings for a subset of survey respondents, and use these to adjust self-ratings in the sample as a whole. Given the great importance of comparative research, we must hope that some combination of creativity and perseverance will lead to methods enabling valid cross-group comparisons in survey-based research. The anchoring vignettes evaluated here appear not to have met this challenge; perhaps future ones will.

Acknowledgments This article uses data from the World Health Organization (WHO) Study on Global Ageing and Adult Health (SAGE), the WHO World Health Surveys (WHS), and the Health and Retirement Study (HRS). SAGE is supported by the National Institute on Aging (NIA) Division of Behavioral and Social Research through interagency agreements and research grants, and the WHO Department of Health Statistics and Information Systems. The HRS is sponsored by the NIA (grant number NIA U01AG009740) and is conducted by the University of Michigan. The first author's work on this research was supported by a Robert Wood Johnson Foundation Health & Society Scholars Dissertation Grant, and by research assistantships through the Center for Demography of Health and Aging and the Center for Demography and Ecology (core grants P30 AG017266 and R24 HD047873) at the University of Wisconsin–Madison. We thank Joan Fujimura, Robert M. Hauser, Pamela Herd, Cameron Macdonald, and James R. Walker for helpful comments on earlier versions of this manuscript.

References

Angel, R. (2013). After Babel: Language and the fundamental challenges of comparative aging research. *Journal of Cross-Cultural Gerontology, 28,* 223–238.

Au, N., & Lorgelly, P. K. (2014). Anchoring vignettes for health comparisons: An analysis of response consistency. *Quality of Life Research, 23,* 1721–1731.

Bago D'Uva, T., Lindeboom, M., O'Donnell, O., & van Doorslaer, E. (2011a). Education-related inequity in healthcare with heterogeneous reporting of health. *Journal of the Royal Statistical Society, Series A, 174,* 639–664.

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

Biss, E. (2005). The pain scale. *Seneca Review, 35*(1), 5–25.

Burgard, S. A., & Chen, P. V. (2014). Challenges of health measurement in studies of health disparities. *Social Science & Medicine, 106,* 143–150.

Corrado, L., & Weeks, M. (2010). *Identification strategies in survey response using vignettes* (Cambridge Working Papers in Economics No. 1031). Cambridge, UK: Cambridge University. Retrieved from http://ideas.repec.org/p/cam/camdae/1031.html

Dong, H., Campbell, C., Kurosu, S., Yang, W., & Lee, J. Z. (2015). New sources for comparative social science: Historical population panel data from East Asia. *Demography, 52,* 1061–1088.

Dowd, J. B., & Zajacova, A. (2007). Does the predictive power of self-rated health for subsequent mortality risk vary by socioeconomic status in the US? *International Journal of Epidemiology, 36,* 1214–1221.

Grol-Prokopczyk, H. (2014). Age and sex effects in anchoring vignette studies: Methodological and empirical contributions. *Survey Research Methods, 8,* 1–17.

JrnlID 13524_ArtID 422_Proof# 1 - 20/08/2015

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

Hanna, L. C., Hunt, S. M., & Bhopal, R. S. (2012). Using the Rose Angina Questionnaire cross-culturally: The importance of consulting lay people when translating epidemiological questionnaires. *Ethnicity & Health, 17,* 241–251.

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

Hunt, S. M., & Bhopal, R. (2004). Self report in clinical and epidemiological studies with non-English speakers: The challenge of language and culture. *Journal of Epidemiology and Community Health, 58,* 618–622.

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

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change and democracy.* New York, NY: Cambridge University Press.

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

Jylhä, M., Guralnik, J. M., Ferrucci, L., Jokela, J., & Heikkinen, E. (1998). Is self-rated health comparable across cultures and genders? *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 53,* S144–S152.

Kapteyn, A. (2010). What can we learn from (and about) global aging? *Demography, 47*(Suppl), S191–S209.

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

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

Kowal, P., Chatterji, S., Naidoo, N., Biritwum, R., Fan, W., Lopez Ridaura, R., … Boerma, J. T. (2012). Data resource profile: The World Health Organization Study on Global AGEing and Adult Health (SAGE). *International Journal of Epidemiology, 41,* 1639–1649.

Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics, 15,* 96–117.

Menec, V. H., Shooshtari, S., & Lambert, P. (2007). Ethnic differences in self-rated health among older adults: A cross-sectional and longitudinal analysis. *Journal of Aging and Health, 19,* 62–86.

Murray, C. J. L., Tandon, A., Salomon, J. A., Mathers, C. D., & Sadana, R. (2002). New approaches to enhance cross-population comparability of survey results. In C. J. L. Murray, J. A. Salomon, C. D. Mathers, & A. D. Lopez (Eds.), *Summary measures of population health: Concepts, ethics, measurement and applications* (pp. 421–431). Geneva, Switzerland: World Health Organization.

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

National Institute on Aging (NIA). (2012). *Harmonization strategies for behavioral, social science, and genetic research* (Workshop Summary Report). Retrieved from https://www.nia.nih.gov/sites/default/files/nia_bssg_harmonization_summary_version_2-5-20122.pdf

Pan, Y., & Fond, M. (2014). Evaluating multilingual questionnaires: A sociolinguistic perspective. *Survey Research Methods, 8,* 181–194.

Pasick, R. J., Stewart, S. L., Bird, J. A., & D'Onofrio, C. N. (2001). Quality of data in multiethnic health surveys. *Public Health Reports, 116*(Suppl. 1), 223–243.

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

Rice, N., Robone, S., & Smith, P. (2011). Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *The European Journal of Health Economics: HEPAC: health economics in prevention and care, 12,* 141–162.

Ruggles, S. (2014). Big microdata for population research. *Demography, 51,* 287–297.

Sadana, R., Mathers, C. D., Lopez, A. D., Murray, C. J. L., & Moesgaard Iburg, K. (2002). Comparative analyses of more than 50 household surveys on health status. In C. J. L. Murray, J. A. Salomon, C. D. Mathers, & A. D. Lopez (Eds.), *Summary measures of population health: Concepts, ethics, measurement and applications* (pp. 369–386). Geneva, Switzerland: World Health Organization.

Schenker, N., Raghunathan, T. E., & Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine, 29,* 533–545.

Schiavenato, M., & Craig, K. D. (2010). Pain assessment as a social transaction: Beyond the "gold standard.". *Clinical Journal of Pain, 26,* 667–676.

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

Shetterly, S. M., Baxter, J., Mason, L. D., & Hamman, R. F. (1996). Self-rated health among Hispanic vs non-Hispanic white adults: The San Luis Valley Health and Aging Study. *American Journal of Public Health, 86,* 1798–1801.

Skevington, S. M. (2002). Advancing cross-cultural research on quality of life: Observations drawn from the WHOQOL development. *Quality of Life Research, 11,* 135–144.

Smith, T. W. (2003). Developing comparable questions in cross-national surveys. In J. A. Harkness, F. J. R. van der Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 69–91). Hoboken, NJ: John Wiley & Sons.

Tandon, A., Murray, C. J. L., Salomon, J. A., & King, G. (2003). Statistical models for enhancing cross-population comparability. In C. J. L. Murray & D. B. Evans (Eds.), *Health systems performance assessment: Debates, methods and empiricism* (pp. 727–741). Geneva, Switzerland: World Health Organization.

United Nations Development Programme (UNDP). (2008). *Human development report 2007/2008*. Retrieved from http://hdr.undp.org/en/reports/global/hdr2007-2008/

van Soest, A., & Vonkova, H. (2014). Testing the specification of parametric models by using anchoring vignettes. *Journal of the Royal Statistical Society, Series A, 177,* 115–133.

van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society, Series A, 174,* 575–595.

Zimmer, Z., Natividad, J., Lin, H.-S., & Chayovan, N. (2000). A cross-national examination of the determinants of self-assessed health. *Journal of Health and Social Behavior, 41,* 465–481.