Categorical Borders Across Borders:

Can Anchoring Vignettes Identify Cross-National Differences in Health-Rating Style? *

Hanna Grol-Prokopczyk

University of Wisconsin-Madison

Mary McEniry

University of Wisconsin-Madison

Emese Verdes

World Health Organization

* This paper uses data from the WHO Study on Global AGEing and Adult Health (SAGE) and the WHO World Health Surveys (WHS). We thank Marton Ispany for statistical assistance, and Sarah Moen for assistance with table formatting and editing. Address correspondence to Hanna Grol-Prokopczyk, University of Wisconsin-Madison, Department of Sociology, 8128 Sewell Social Sciences Building, 1180 Observatory Drive, Madison, WI 53706 (email: hgrol@ssc.wisc.edu).

Categorical Borders Across Borders:

Can Anchoring Vignettes Identify Cross-National Differences in Health-Rating Style?

Abstract

Evidence indicates that self-reported measures of health cannot be directly compared across nations, because groups differ in how they use subjective response categories. Anchoring vignettes have been proposed as a solution to this problem, since they permit statistical adjustment for rating style, and thus enable valid intergroup comparisons. However, many anchoring vignettes have not been formally evaluated for adherence to key measurement assumptions, namely, *vignette equivalence* and *response consistency*. In this paper, we conduct such a formal evaluation by applying recently developed statistical tests to vignette data from the WHO Study on Global AGEing and Adult Health (SAGE) and the World Health Survey (WHS), representing a diverse set of ten countries (n=52,388) and covering eight domains of health (mobility, affect, pain, social relationships, vision, sleep, cognition, and self-care). We find substantial evidence that vignette equivalence is violated cross-nationally in all domains of health, with the exception of certain limited contexts, such as specific two-country comparisons. In contrast, our evidence is generally concordant with the assumption of response consistency. Nonetheless, existing WHO anchoring vignettes must be used with caution. We conclude with recommendations for future implementations and analyses of vignettes.

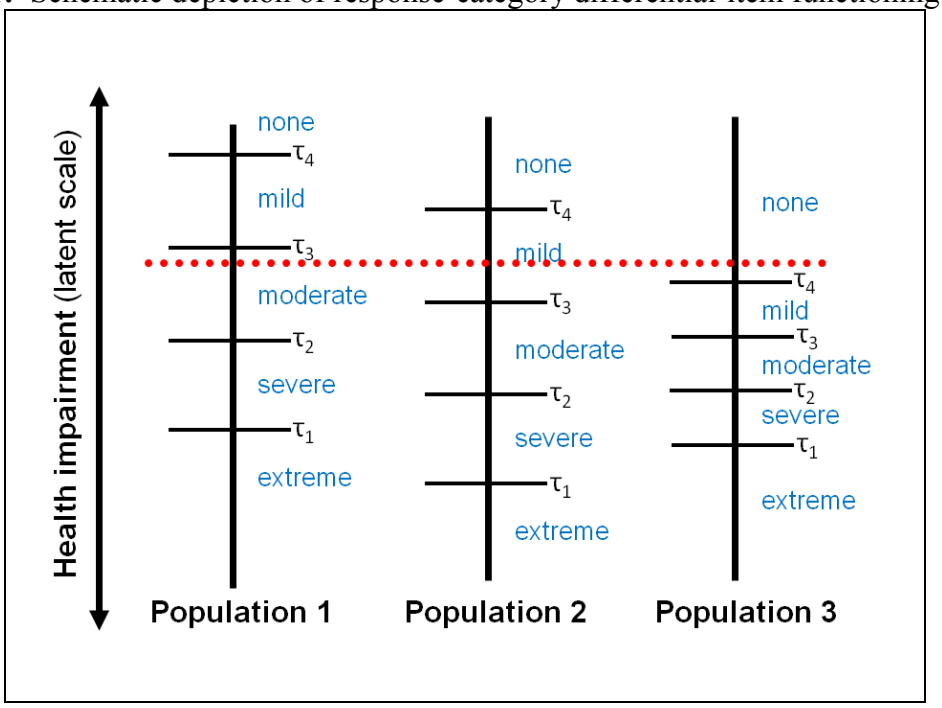
The past decade has a seen a dramatic growth of interest in health-related anchoring vignettes, as reflected in the growing availability of anchoring vignette data and the increasing number and sophistication of studies based on these vignettes. However, in many cases, the adherence of the vignettes to essential measurement assumptions has not been formally or rigorously tested, making it unclear whether anchoring vignettes are actually functioning as intended. This paper pairs recently released anchoring vignette data from World Health Organization (WHO) surveys with recently developed statistical techniques to test the validity of the most widely fielded health-related anchoring vignettes in the world.

ANCHORING VIGNETTES: USES AND EVALUATION

Reporting heterogeneity

For at least three decades, evidence has been mounting that self-reports of health are often incomparable across national, racial-ethnic, or other demographic groups, and that this problem is independent of issues of translation (e.g., King et al. 2004; Murray et al. 2002). In particular, there is accumulating evidence that, when rating health using subjective ordinal response categories (such as "excellent, very good, good, fair, or poor" for the general self-rated health item, or "none, mild, moderate, severe, or extreme" for questions about level of health impairment), some populations use certain response categories more liberally than do others. Phrased more formally, groups may differ in where on the latent health spectrum they locate the thresholds between adjacent response categories, as shown in Figure 1. Such differences in rating style are referred to as "response category differential item functioning" (DIF) (King et al. 2004) or "reporting heterogeneity" (e.g., Bago D'Uva et al. 2009).

Figure 1: Schematic depiction of response-category differential-item functioning (DIF).



Description: Populations may differ in how they divide the underlying health spectrum into categories. This has been referred to as "response-category differential item functioning," or DIF (King et al. 2004). Here, Population 1 has systematically higher intercategory cutpoints (τ's) than Population 2. Population 3 shows a compression of cutpoints relative to the other two groups. In this scenario, the three groups could have equal mean levels of health impairment (represented by the dotted horizontal line), but nonetheless use three different terms to refer to that level of impairment—here, "moderate," "mild," and "none," respectively.

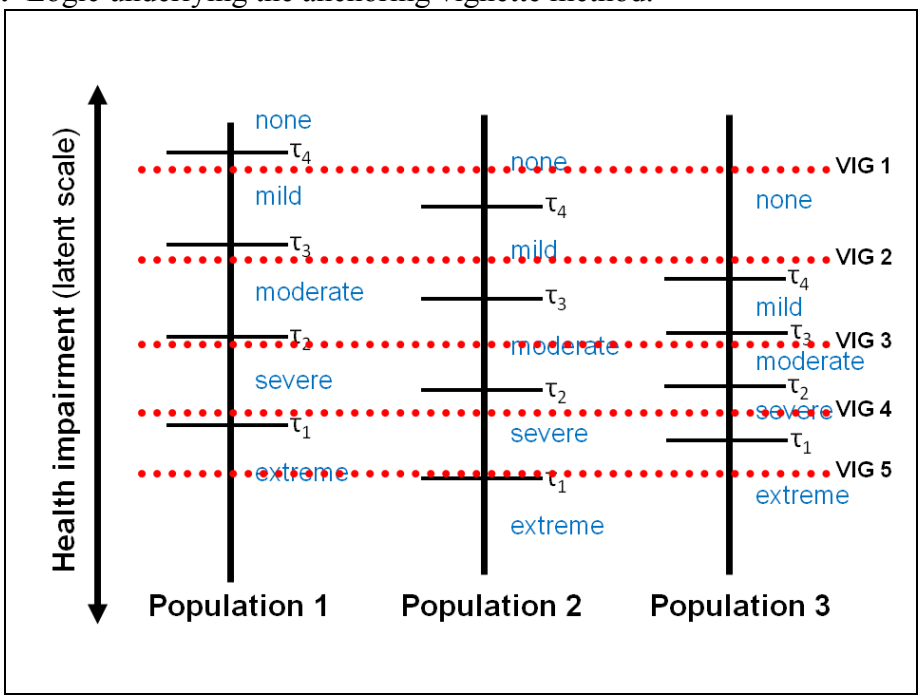
Banks et al. (2007), comparing American and English men's health, found that by any of several objective measures (including biomarkers), the English men had better health than the Americans. However, in their self-reports, the Americans rated their health more highly. Banks et al. conclude that this "contradiction most likely stems from different thresholds used by Americans and English. For the same 'objective' health status, Americans are much more likely to say their health is good than are the English" (28). Likewise, Sadana et al. (2002) describe implausible discrepancies among European countries in the proportion of respondents who rate themselves in very good versus good health (370), and present evidence that per capita health expenditures are *inversely* associated with self-reported health in a sample of 46 countries

(381)—findings which similarly suggest incomparability across nations in health-rating style. Many other examples suggesting cross-national differences in health category thresholds could be cited (e.g., Jürges 2007; Jylhä et al. 1998; Zimmer et al. 2000), along with evidence that *within* populations, reporting heterogeneity is observed by age, sex, race-ethnicity, and socioeconomic status (see Grol-Prokopczyk et al. 2011 for relevant citations).

Anchoring vignettes

In the early 2000s, researchers at the WHO undertook a systematic comparison of techniques for overcoming response-category DIF, and concluded that *anchoring vignettes* are "the most promising" of available strategies (Murray et al. 2002:429; cf. Tandon et al. 2003). An anchoring vignette is a brief, hypothetical description of a fictional character who exemplifies the trait of interest (e.g., mobility or vision) to a lesser or greater degree. Respondents are asked to rate their own level of the trait and then, using the same set of response categories, to rate the fictional character's health. Examples of health-related vignettes for four domains are presented in Appendix A. Respondents are given multiple vignettes per domain, each representing different points along the health spectrum. Since vignettes are held constant across respondents, any differences in ratings of a given vignette are considered indicative of DIF. 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 (τ's). Group differences in rating style can then be adjusted for via of any of several parametric or non-parametric techniques (King et al. 2004; King and Wand 2007), allowing for valid intergroup comparisons unbiased by DIF. The logic underlying the anchoring vignette method is depicted in Figure 2.

Figure 2: Logic underlying the anchoring vignette method.



Description: By giving the same series of anchoring vignettes (here, "VIG 1" through "VIG 5") to all respondents, researchers can determine how different groups use subjective response categories such as "mild" or "moderate." More formally, researchers can estimate where different groups locate intercategory thresholds (here, τ_1-τ_4), and adjust for different use of such thresholds in subsequent analyses.

The two key measurement assumptions of the anchoring vignette method are *vignette equivalence* and *response consistency*. Vignette equivalence refers to the fact that respondents perceive the vignettes as representing the same absolute position on the underlying (latent) health spectrum. (In Figure 2, this is shown by the depiction of the vignettes as flat horizontal lines.) Violations of response consistency may occur if groups interpret the vignette texts in systematically different ways. For example, if a vignette character's obesity is interpreted by residents of rich countries as a sign of bad health, but is understood by residents of poor countries as a sign of wealth and good health, then vignette equivalence has been violated. Response consistency refers to the assumption that respondents rate themselves and vignette

characters using the same thresholds. If respondents hold themselves to higher or lower standards than the vignette characters, then response consistency is violated, and the cutpoints calculated from the vignettes will not correctly adjust self-ratings of health. Neither vignette equivalence nor response consistency can be taken for granted: some studies find evidence supportive of adherence to these assumptions (e.g., Grol-Prokopczyk et al. 2011, Rice et al. 2009, Van Soest et al. 2007), while others find the opposite (e.g., Bago D'Uva et al. 2009, Datta Gupta et al. 2010).

Anchoring vignettes pertaining to eight domains of health (mobility, affect, pain, social relationships, vision, sleep, cognition, and self-care) were fielded to approximately 300,000 respondents in 70 countries as part of the 2002 WHO World Health Survey (WHS). The same vignettes were also included in the WHO's 2007-2009 Study on Global AGEing and Adult Health (SAGE), which surveyed approximately 44,000 respondents in six countries. Though vignettes were subsampled, so that each respondents answered questions from only two of the eight health domains, this nonetheless represents an enormous quantity of data, making the WHO vignettes the most widely-fielded health vignettes in the world. (Furthermore, modified subsets of the WHO vignettes have been included in several other surveys, including the American Health and Retirement Study [HRS; http://hrsonline.isr.umich.edu/], the Survey of Health, Ageing and Retirement in Europe [SHARE; http://www.share-project.org/], and the English Longitudinal Study of Ageing [ELSA; http://www.ifs.org.uk/elsa/].) Despite such widespread use, to date, no systematic evaluation of the WHO vignettes has been conducted regarding their adherence to the statistical assumptions of the method.

Testing measurement assumptions

Developing methods to test adherence to vignette equivalence (VE) and response consistency (RC) has proven conceptually and statistically challenging, as evidenced by the lack, for most of the past decade, of strong tests of these assumptions. In their 2004 paper—widely regarded as the foundational work on the anchoring vignette method—King et al. conduct only a minimal test of vignette equivalence, namely, to check that most respondents correctly rank-order vignettes in a series. This is a "weak" test, in the sense 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, such as looking for systematic patterns among non-normative rankings, or looking for differences in ranking consistencies across groups (e.g., Kristensen and Johansson 2008; Rice et al. 2009). A novel and more stringent approach was proposed in Bago D'Uva et al.'s November 2009 paper, using ELSA data. Bago D'Uva et al. seize upon the observation 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. Though models cannot simultaneously identify the locations of all vignettes in a series, if one vignette is constrained to be the same for all respondents (e.g., by setting it to zero), then the 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. Given the recency of the Bago D'Uva (2009) paper, the method has yet to be widely applied.

Response consistency, too, has proved challenging to test, especially since assessing whether respondents rate vignette characters as they rate themselves depends on availability of data capturing respondents' "true" or 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 (as measured by Snellen eye chart exams), but the strength of this correlation was not scrutinized; a similar approach was taken by Grol-Prokopczyk et al. (2011). Van Soest et al.'s (2007) paper on binge drinking provided a more compelling test of RC, but its applicability has been limited, since it hinges upon a unique property of their data (specifically, the fact that drinking behavior can be easily quantified in terms of number of alcoholic drinks consumed; most domains of health defy such straightforward quantification). However, in the same 2009 paper mentioned just above, Bago D'Uva et al. propose a novel approach to testing RC as well: namely, to compare the locations of cutpoints estimated from vignette ratings with the locations of cutpoints estimated from objective measures of health. If the two sets of cutpoints line up closely, this supports the assumption of response consistency, as it shows that vignette-ratings and self-ratings are made using similar standards of evaluation.

PROJECT GOALS

The primary goal of this paper is to apply recently developed, stringent tests of vignette validity to health vignettes from the WHO's SAGE and WHS surveys, and thereby to evaluate the usefulness of the most widely-fielded health vignettes in the world. Specifically, we conduct two kinds of tests of vignette equivalence: one based on rank-orderings of vignettes, and one based on the Bago D'Uva et al. (2009) test of perceived vignette locations. By including both, we can assess whether "weak tests" and "strong tests" of VE yield similar results. We also conduct a version of Bago D'Uva et al.'s (2009) test of response consistency, based on comparison of cutpoint locations generated from vignettes versus from objective health

measures. We seek to clarify whether the WHO's vignettes function as intended, and thus whether they can serve as a useful tool to overcome cross-national reporting heterogeneity.

DATA AND ANALYTIC STRATEGIES

Data and variables

Our analyses are based primarily on data from the 2007-2009 (Wave 1) Study on Global AGEing and Adult Health (SAGE) (http://www.who.int/healthinfo/systems/sage/en/), which includes nationally representative samples from six countries: China, Ghana, India, Mexico, Russia, and South Africa. SAGE surveys include measured tests of vision, mobility, and cognition, as well as relatively objective self- or interviewer-reported measures of these and other health domains; such objective measures enable testing of response consistency. Since SAGE includes only low- or middle-income countries, we increase the socioeconomic and geographic diversity of our sample by also including data from four countries participating in the 2002 World Health Survey (WHS) (http://www.who.int/healthinfo/survey/en/): Brazil, France, Netherlands, and United Kingdom (UK). We thus include at least one country from each of the major regions of the Inglehart-Welzel Cultural Map of the World (Appendix B; Inglehart and Welzel 2005:64). The diversity of our sample of countries allows us to put vignette equivalence to a particularly rigorous test. Because the WHS did not include measured tests of health, however, WHS countries could not be included in tests of response consistency.

Descriptive statistics for our sample are shown in Table 1 below. We note the very different age structures of the countries in our sample. SAGE, designed as a survey of aging, focuses on adults aged 50 and older. However, some respondents under 50 were included as a comparison group, and the proportion of such younger respondents varies dramatically across

SAGE countries (from 9% for South Africa to 41% for India). The WHS countries, in contrast, include proportionate representation of adults aged 18 and over.

In all but the three European surveys, each vignette text was followed by *two* evaluation questions. For example, after each vision vignette, respondents rated the character's difficulty with both distance vision and near vision. The eight domains of health vignettes thus yielded 16 subdomains of vignette ratings, as follows: mobility (moving around, vigorous action), affect (depression, anxiety), pain (pain, discomfort), social relationships (relationships, conflict), vision (distance vision, near vision), sleep (sleep, energy), cognition (memory, learning new things), and self-care (self-care, appearance). In the European surveys, only the first of each pair of evolution questions was asked.

We conducted tests of vignette equivalence for all 16 subdomains of vignette ratings (excluding European countries where necessary). For tests of response consistency, we focused on vision and mobility, as these are the domains for which SAGE provides particularly good objective measures. In particular, for vision SAGE includes a measured visual acuity score (based on a standard optometry exam; scores are converted to decimal form, e.g., 20/20 is expressed as 1.0), and respondents' self-reports (yes/no) of cloudy vision and glares or halos. For mobility, SAGE includes two timed four-meter walks (one at normal walking speed, one at rapid walking speed), as well as the interviewer's evaluation of whether the respondent has difficulty walking.

In our analyses, age was grouped into five categories and education into six, as shown in Table 1. These variables, along with sex and country, were entered into models as dummy variables. All ages were included in analyses unless otherwise noted.

Table 1: Descriptive statistics for sample. All data are from SAGE Wave 1 except Brazil, UK, France, and Netherlands, from WHS.

Descriptive Statistics	Overall	Ghana	India	S Africa	China	Brazil	Russia	Mexico	UK	France	Netherlands
Sample Size	52,388	5,565	12,198	4,225	15,009	5,000	4,350	2,742	1,200	1,008	1,091
Sex											
Male (%)	42.57	50.60	38.60	42.53	46.59	43.76	35.59	38.29	36.83	40.08	32.54
Female (%)	57.43	49.40	61.40	57.47	53.41	56.24	64.41	61.71	63.17	59.92	67.46
Age Group											
18-49 (%)	27.05	15.08	41.38	9.11	10.94	70.16	9.59	15.65	49.71	67.23	57.75
50-59 (%)	29.90	33.85	26.06	40.12	38.69	13.48	33.75	15.84	13.62	15.59	18.24
60-69 (%)	22.64	23.46	20.13	29.16	26.44	9.08	24.62	34.05	15.79	7.75	15.77
70-79 (%)	15.12	19.25	9.41	15.67	18.67	5.72	23.40	22.60	14.12	6.45	7.33
80+ (%)	5.29	8.36	3.01	5.94	5.26	1.56	8.64	11.85	6.77	2.98	0.92
Education											
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
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
Primary school (%)	17.18	12.43	15.29	23.58	19.10	27.86	7.32	22.52	2.42	14.38	8.07
Secondary school (%)	16.18	5.50	12.42	14.80	21.28	14.88	18.14	10.61	50.83	21.83	7.06
High school completed (%)	17.93	17.08	10.66	8.15	13.87	21.34	51.76	3.80	17.00	27.68	59.85
College or more completed (%)	8.40	3.64	5.91	5.51	5.24	6.56	20.14	9.09	28.67	33.63	23.56
Rural/Urban Residence											
Rural (%)	48.09	59.05	74.32	33.43	50.97	17.36	24.25	26.19	7.17	45.04	n/a
Urban (%)	51.91	49.95	25.68	66.57	49.03	82.64	75.75	73.81	92.83	54.96	n/a
Marital Status											
Never married (%)	8.00	2.89	5.65	15.72	1.91	21.62	4.17	9.51	20.10	27.55	33.70
Married (%)	66.72	59.82	77.58	47.37	83.31	45.38	53.80	58.65	47.99	40.10	37.98
Cohabiting (%)	3.16	1.18	0.00	5.67	0.18	16.30	3.31	4.87	5.65	8.16	7.80
Separated/Divorced(%)	5.13	12.44	0.72	6.29	1.78	8.26	8.72	6.01	12.15	15.41	11.42
Widowed (%)	16.99	23.67	16.05	24.95	12.81	8.44	30.00	20.96	14.11	8.78	9.10

Analytic strategies and models

We conducted both "weak tests" and "strong tests" of vignette equivalence. The weaker tests were based on assessing respondents' rank-orderings of vignettes, to verify that respondents perceive the five severity levels in each domain in the expected order. The percentage of respondents showing the expected rank-ordering was calculated by country and by subdomain. These calculations were "benefit-of-the-doubt" calculations (as in Murray et al. 2003), meaning that ties in ratings were assumed to resolve consistently with the expected ordering.

The stronger test of VE, following Bago D'Uva et al. (2009), is based on a likelihood-ratio (LR) test comparison of two models, A and B. In Model A, the distribution of each vignette's perceived location 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}):

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

In Model B, a selected reference vignette is set to a constant, as in Model A, but all other vignettes may now have their positions affected by a vector of parameters ($\lambda_j X_i$), which in our analyses included sex, age, education, and country:

Model B: As in Model A for reference vignette, but

$V_{ij} = \alpha_j + \lambda_j X_i + \varepsilon_{ij}$ for all other vignettes

If vignette equivalence holds, then $\lambda = 0$ for all j, and a LR test based on the log-likelihoods of the two models will fail to reject the hypothesis of no difference between models. If, however, groups differ in where they perceive vignettes to fall on the latent health spectrum, this test will reject VE (i.e., the associated LR test statistic will yield $p < .05$). Following Bago D'Uva et al. (2009), we refer to this model comparison test as the "global test" of vignette equivalence. We conducted such a global test for each of the 16 subdomains of health vignettes.

Both Models A and B were implemented by variations on the "hopit" (hierarchical ordered probit) model commonly used in vignette studies (described in Rabe-Hesketh and Skrondal 2002; cf. King et al. 2004. Some authors refer to this as "chopit"). Unlike a standard ordered probit, which assumes fixed response-category cutpoints, hopit models allow cutpoints to vary across groups (based on ratings of anchoring vignettes, unless otherwise specified). 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. For both Models A and B, we allowed cutpoints to vary by sex, age, education, and country. However, in Model A, only dummies for vignette severity (1-5, where 5 represents the worst state of health) entered into the equation for the perceived vignette locations. In contrast, in Model B, the equation also included multiple interaction terms, each representing the interaction between a given severity and a covariate. For example, the "Severity 1 * female" interaction would indicate whether the perceived distance between the Severity 1 vignette and the reference vignette was different for women than for men. Such interactions were included for each severity crossed with each covariate (sex, age, education, and country), excluding omitted categories. These interaction terms indicate which covariates drive violations of vignette equivalence.

Bago D'Uva et al. (2009:11) also propose a LR-based global test of response consistency, which compares a model that estimates intercategory cutpoints via vignettes with a model that estimates them via objective measures of health. However, this tests depends on vignette equivalence; that is, the null hypothesis of no difference between models will be rejected if RC is violated *or* if VE is violated. Given our findings regarding VE, this formal global test was not appropriate for our use. However, we use a somewhat less stringent test suggested in the same paper (2009:11-12), namely, to visually compare the cutpoints generated by the two models for

equidistance between cutpoints. Violations of VE may affect the apparent position of a given set of cutpoints (taken as a whole) along the latent health spectrum, but "the distance between any two true cut-points is [nonetheless] identified" (Bago D'Uva et al. 2009:12). Thus, observing similar "shapes" of cutpoints across the vignette-based and the objective health measure-based models would be supportive of RC (with the caveat that the relative positions of the two sets of cutpoints along the latent spectrum cannot be determined with certainty).

Concretely, to estimate intercategory cutpoints from vignette ratings alone, we used the same hopit model as for Model A above, except instead of focusing on the estimated vignette locations, we examine the estimated cutpoint locations. To estimate intercategory cutpoints from (relatively) objective measures of health, we used a third form of hopit model, Model C, which is identical to the other two, except that it estimates cutpoints by pairing self-ratings of health with objective measures of health (instead of vignette-ratings with vignette severities and possibly interaction terms). Table 2 summarizes the differences among Models A, B, and C. Stata code used to generate Models A-C (and all other code for this project) is available upon request from the first author.

Table 2: Comparison of hierarchical ordered probit (hopit) models used to test adherence to measurement assumptions.

	Covariates for cutpoint equation	**Outcome variable for second equation**	**Covariates for second equation**
Model A	Sex, age, education, country	Vignette-ratings	Vignette severities
Model B	Sex, age, education, country	Vignette-ratings	Vignette severities plus interactions of severities with sex, age, education, and country
Model C	Sex, age, education, country	Self-ratings	Objective measures of health

Note: Hopit models jointly estimate two equations: one for intercategory cutpoints, and one for vignette- or self-ratings. Vignette equivalence may be tested by comparing Models A and B. Response consistency may be tested by comparing Models A and C.

The three health measures used in Model C for our test of distance vision were distance vision scores (the higher from the left and right eye scores) and respondents' self-reports (yes/no) of cloudy vision and glares or halos. The three measures used in our test of "moving around" were the scores from the two timed walks, plus the interviewer's assessment (yes/no) of whether the respondent had difficulty walking. Given that these objective 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 the assumption of response consistency.

RESULTS

Vignette equivalence: weak tests

The percentage of respondents who ranked vignettes in each domain "correctly," that is, in the expected order by severity level, is shown by subdomain in Table 3, and by country in Table 4. In both tables, adherence to the expected ordering averages at about 90%, with a range of approximately 82-96%. Since some variation in vignette ordering is expected due to measurement error, these data appear reasonably consistent with the assumption of vignette equivalence.

Furthermore, vignette orderings that deviate from the expected global ordering (not shown) appear to be widely distributed across the range of all possible alternative orderings, rather than reflecting a small number of alternate orderings. Such non-systematic deviations from expected orderings are suggestive of measurement error, rather than systematic, alternate understandings of vignettes. We thus find little evidence of multidimensionality in these rank-order-based tests of VE.

Table 3: Percentage of rank-orderings consistent with expected ordering, by health subdomain.

Subdomain	Mean % consistent, across countries	Range across countries
Moving Around	91.51%	87.75 - 95.26%
Vigorous Action	91.57%	87.35 - 94.27%
Depression	92.18%	88.00 - 94.94%
Anxiety	91.80%	88.22 - 94.22%
Pain	91.60%	83.84 - 94.34%
Discomfort	90.68%	88.13 - 94.48%
Relationships	88.15%	82.17 - 90.99%
Conflict	87.63%	82.02 - 91.69%
Distance Vision	89.39%	86.06 - 92.87%
Near Vision	88.94%	86.00 - 92.18%
Sleep	89.58%	83.44 - 88.93%
Energy	85.84%	83.01 - 88.09%
Memory	92.83%	86.65 - 95.61%
Learning	91.59%	86.65 - 95.70%
Self-Care	88.99%	85.57 - 92.96%
Appearance	87.14%	83.67 - 88.78%

Table 4: Percentage of rank-orderings consistent with expected ordering, by country.

Country	Mean % consistent, across subdomains	Range across subdomains
Ghana	90.55%	83.67 - 93.55%
India	87.50%	84.37 - 90.77%
South Africa	91.09%	86.96 - 93.59%
China	91.07%	86.86 - 95.26%
Brazil	89.24%	84.34 - 92.74%
Russia	91.86%	87.24 - 95.70%
Mexico	85.90%	82.02 - 88.48%
UK	90.97%	88.08 - 95.33%
France	91.56%	88.89 - 94.54%
Netherlands	90.55%	88.91 - 93.31%

Vignette equivalence: strong tests

Table 5 shows the results of the global tests for vignette equivalence. As shown by the rightmost column, the assumption of vignette equivalence is rejected ($p < .001$) for all 16

subdomains when Model B includes sex, age, education, and country (interacted with vignette severities) as covariates. Alternate versions of Model B including various subsets of these covariates were also tested. However, with the exception of a model including only sex (which rejects VE for only six subdomains, namely, pain, discomfort, relationships, distance vision, self-care, and appearance [not shown]), VE equivalence is consistently rejected in all cases.

Table 5: Global tests of vignette equivalence, by health subdomain.

	Degrees of freedom	**LR test statistic**	***p*-value**
Moving Around	76	3029.67	< .001
Vigorous Action	64	2290.08	< .001
Depression	76	4362.42	< .001
Anxiety	64	3670.66	< .001
Pain	76	4282.32	< .001
Discomfort	64	3752.88	< .001
Relationships	76	3454.48	< .001
Conflict	64	3560.31	< .001
Distance Vision	76	4120.46	< .001
Near Vision	64	4333.15	< .001
Sleep	76	2734.23	< .001
Energy	64	2367.05	< .001
Memory	76	7238.12	< .001
Learning New Things	64	6693.15	< .001
Self-Care	76	2765.49	< .001
Appearance	64	2548.44	< .001

Note: Test is based on comparison of Models A and B. Covariates included in Model B are sex, age, education, and country interacted with each vignette severity. Degrees of freedom are lower for subdomains not included in European surveys.

To better understand which covariates drive the rejection of VE in the global tests, we next examine in more detail some Model B results. Due to space constraints, we present here only an extract of Model B output for the "moving around" subdomain, in Table 6. This table includes only coefficients predicting the location of the Severity 1 vignette, but these are indicative of findings for the other severities as well. A fuller version of the table including coefficients for all vignettes appears in Appendix C.

Table 6: Predictors of perceived vignette position for "moving around" subdomain. (Extract; fuller data shown in Appendix C table.)

	Ordered probit	
	β	SE
Severity 1	4.225***	.086
Severity 2	2.742***	.073
Severity 3	2.005***	.069
Severity 4	1.350***	.067
Sev 1 × Female	.069	.041
Sev 1 × Age 50-59	-.006	.057
Sev 1 × Age 60-69	.012	.061
Sev 1 × Age 70-79	-.214**	.068
Sev 1 × Age 80+	-.295**	.098
Sev 1 × Some Primary	.034	.066
Sev 1 × Primary Completed	-.144*	.061
Sev 1 × Secondary Completed	.102	.067
Sev 1 × High School Completed	.161*	.069
Sev 1 × College Completed	.336***	.089
Sev 1 × China	.490***	.083
Sev 1 × France	.075	.168
Sev 1 × UK	.939***	.165
Sev 1 × Ghana	-.257**	.094
Sev 1 × India	-1.371***	.076
Sev 1 × Mexico	-1.578***	.102
Sev 1 × Netherlands	-.221	.143
Sev 1 × Russia	1.309***	.118
Sev 1 × South Africa	-.279**	.103

Note: Perceived position of vignettes is calculated relative to the Severity 5 vignette. Other omitted reference categories are male (for sex), under age 50 (age), no formal schooling (education), and Brazil (country). Data generated by Model B hopit regression.

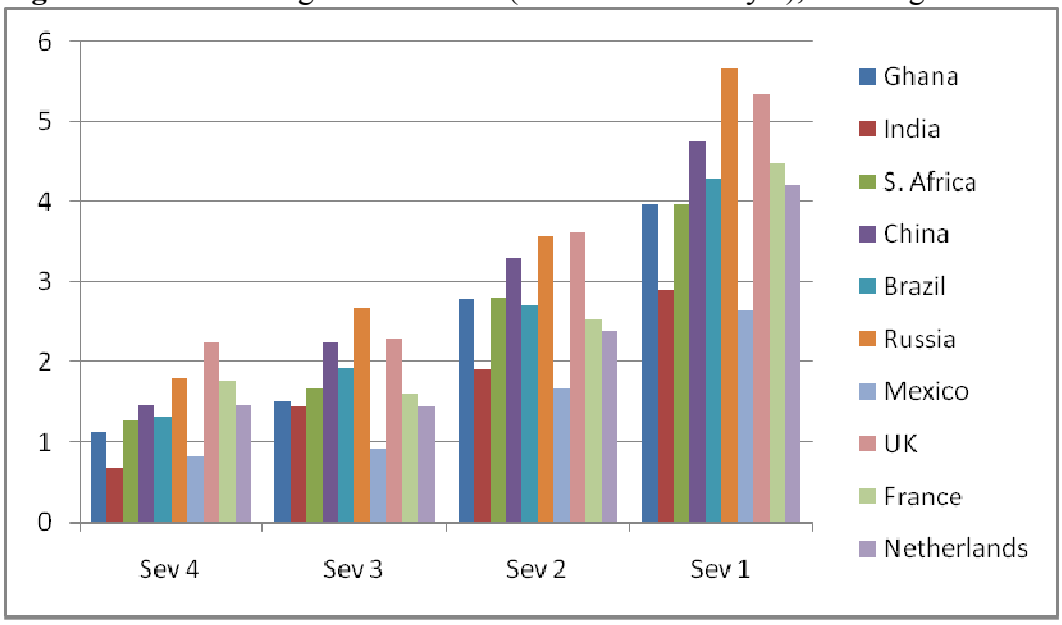
As shown by the lack of a statistically significant interaction between "Female" and Severity 1, there is no significant difference between men's and women's perceptions of Vignette 1's location relative to the reference vignette. (This is true for Severities 2-4 as well.) There is some evidence that older respondents perceive vignettes as falling lower on the latent health spectrum than do younger ones, as shown by the negative and statistically significant coefficients for the age interactions for respondents aged 70-79 or 80 and above. (This pattern,

too, holds for Severities 2-4.) The effect of education is inconsistent, with those completing primary school perceiving lower vignette locations than those with no formal schooling, while college graduates perceive higher vignette locations, at least for vignettes distant from the reference vignette. The largest and most consistently statistically significant coefficients in the model, however, are those for country interactions. Indeed, the coefficients for the countries are often several times larger than even the largest age or education coefficients—and this is true for all severities. It thus appears that cross-national differences in understandings of vignettes are substantially larger than differences across sexes, age groups, or educational categories.

Furthermore, this is the case not only for the "moving around" subdomain, but for all 16 subdomains. The effects of other covariates are domain-dependent, and thus cannot be easily summarized. (For example, while sex appears unrelated to perceived vignette location for "moving around", women appear to perceive *pain* vignettes as higher along the health spectrum than do men. In contrast, no age effects are found when Model B is run for pain, though they were found for "moving around".) However, across all subdomains, differences across countries appear consistently; these cross-national differences are often both statistically significant (with $p<.001$ in many cases) and substantively large.

Graphs of perceived vignette locations by country may provide a clearer sense of the extent to which vignette equivalence is violated cross-nationally. To generate such graphs, we used coefficients from Models B to predict perceived vignette locations for all respondents in our sample, for several subdomains. Countries are listed in reverse order of Human Development Index (HDI). Figure 3 shows estimates of perceived vignette location for "moving around" based on the data from Table 6 above. The zero on the y-axis represents the mean of the reference vignette, and units are standard deviations of the reference vignette.

Figure 3: Predicted vignette locations (relative to Severity 5), "moving around".

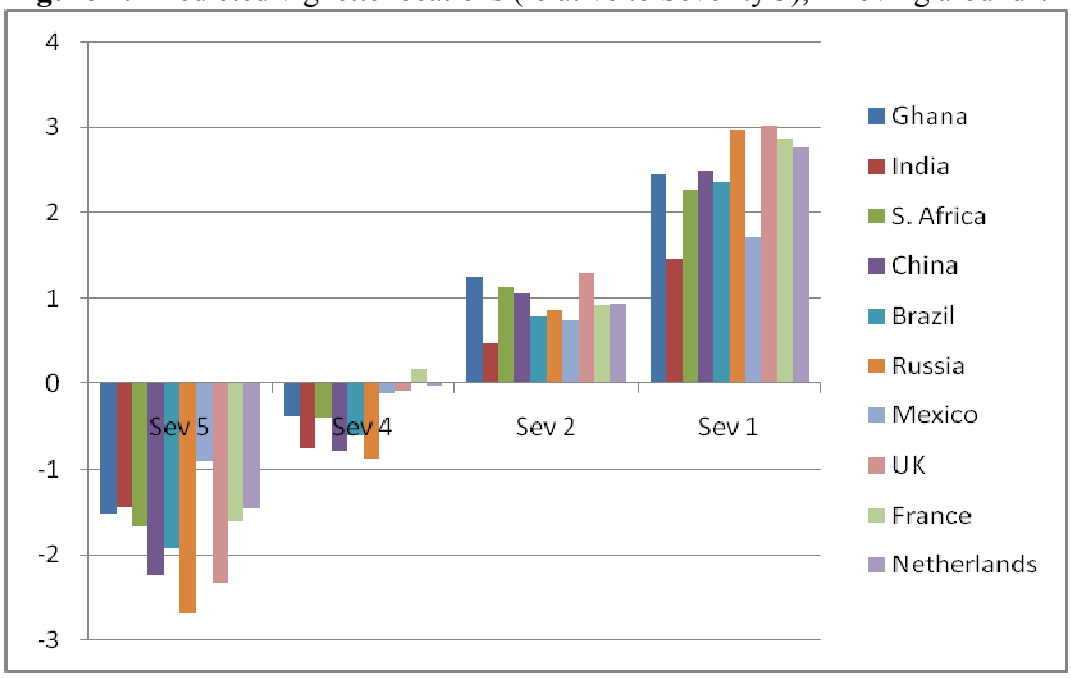


If perfect vignette equivalence were observed, the bars for each severity would all be exactly the same height, so that graph would resemble four table-tops. However, as Figure 3 shows, the predicted vignette locations vary widely across countries. The standard deviations for the country estimates (ranging from.05-.08 for Severity 4 to .14-.20 for Severity 1) are relatively small, meaning that the country estimates often do not have overlapping confidence intervals. These are, then, genuinely large cross-national differences.

To ensure that these findings were not driven by differences in national age-distributions, we created another graph that included only respondents age 50 and above. However, the graph (not shown) was visually nearly indistinguishable from the above. Next, to test the sensitivity of our findings to choice of reference vignette, we re-ran the Model B hopit regressions to use Severity 3 rather than Severity 5 as the omitted vignette. However, as shown in Figure 4, cross-national differences still appear large (and some of the apparent shrinking of differences reflects

the new reference vignette's larger standard deviation, making the current units larger than the previous ones).

Figure 4: Predicted vignette locations (relative to Severity 3), "moving around".



Large cross-national differences in perceived vignette locations are also observable in graphs for other subdomains, e.g., distance vision and memory, shown in Figures 5 and 6, respectively.

Figure 5: Predicted vignette locations (relative to Severity 5), for distance vision.

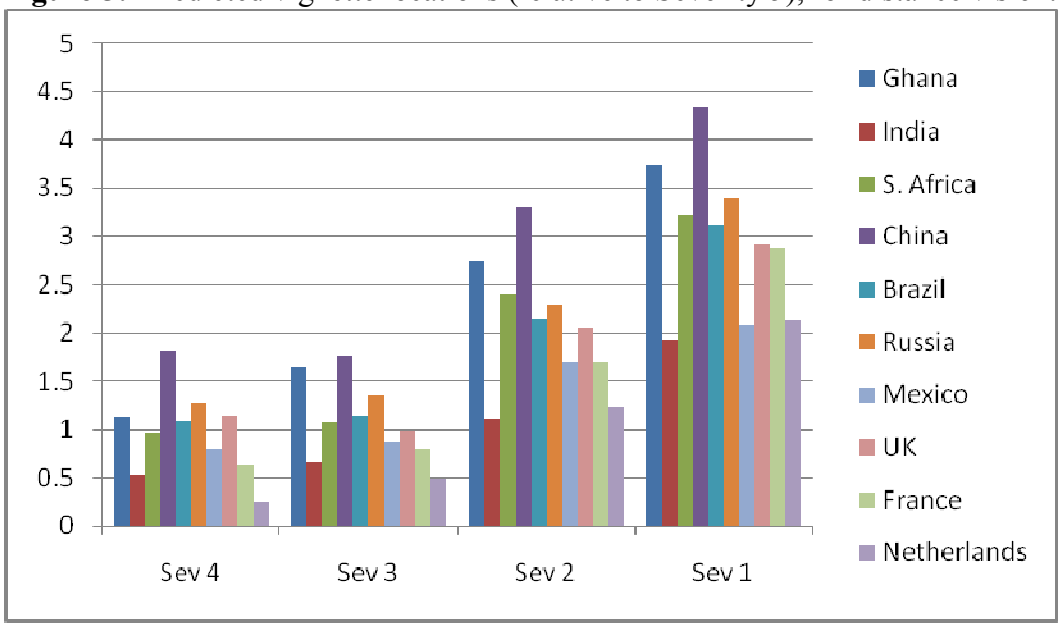
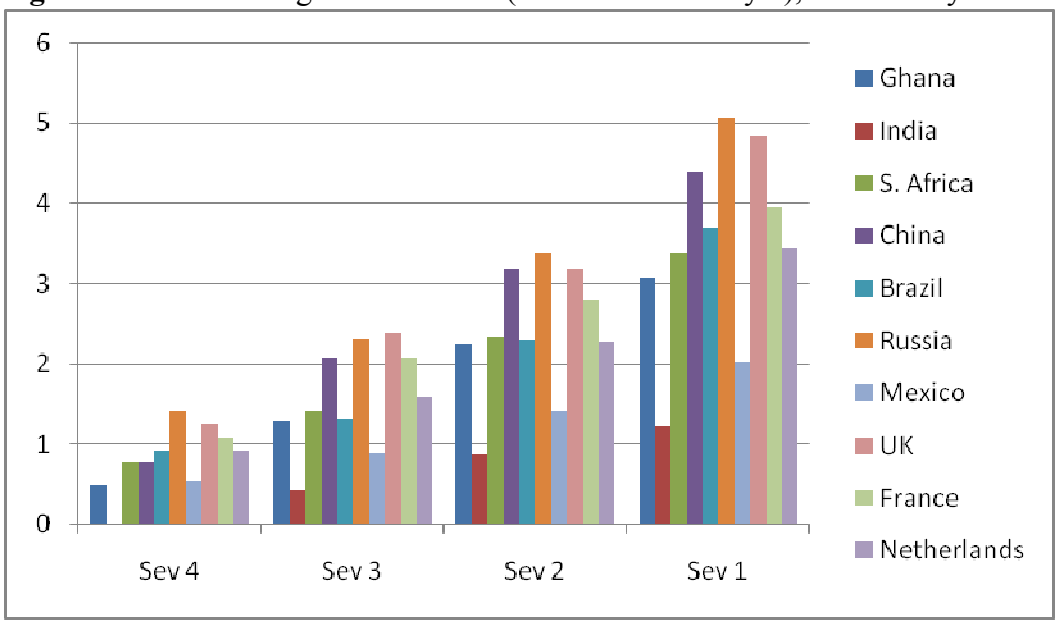


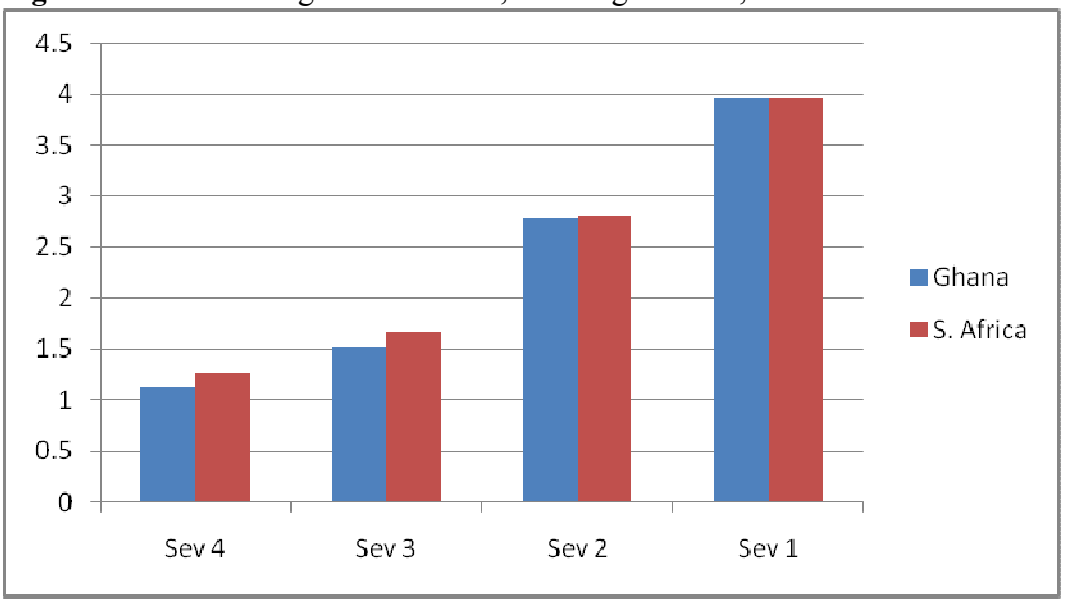
Figure 6: Predicted vignette locations (relative to Severity 5), for memory.



It appears, then, that vignette equivalence is unambiguously violated in highly diverse sets of countries such as those in our sample. However, when select subsets of countries are analyzed, vignette equivalence may be upheld, or violated only slightly. For example, when

Figure 3 is redrawn to include only Ghana and South Africa, we see a striking concordance in perceived vignette locations (Figure 7). Indeed, this particular pairing of countries passes the global test of VE (df=44, LR test statistic=47.02; p=.35).

Figure 7: Predicted vignette locations, "moving around", Ghana and South Africa only.



France and Netherlands do not officially pass the global test of VE for moving around (df=44, LR test statistic=90.0; p<.001; we note that they do pass the test for other domains, such as sleep and self-care). However, they appear to have relatively similar perceptions of vignette locations, shown in Figure 8. Indeed, the same could be said for five of our ten countries, shown in Figure 9. Sensitivity analyses in specific research contexts could clarify whether vignettes may still be useful when only such relatively minor deviations from VE are observed.

Figure 8: Predicted vignette locations, "moving around", France and Netherlands only.

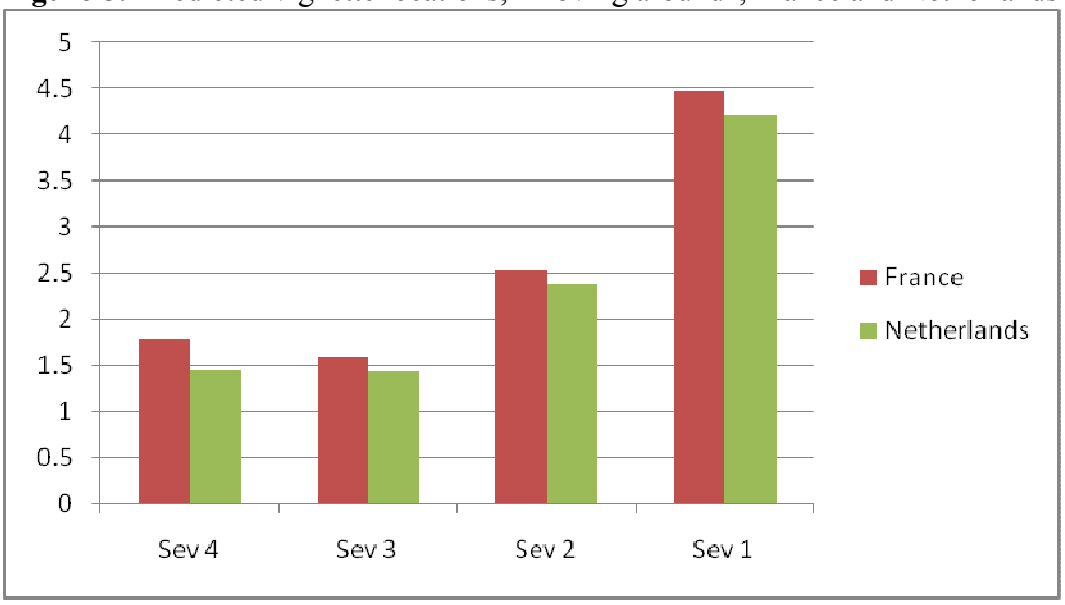
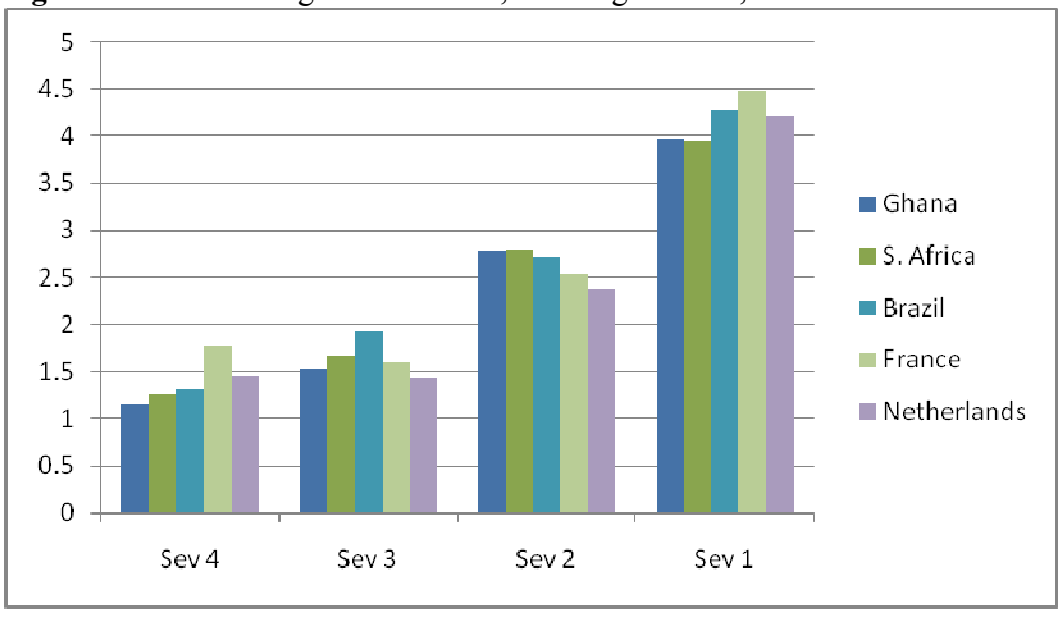


Figure 9: Predicted vignette locations, "moving around", selected countries.



To summarize this section: All subdomains appear to pass weak (rank-order based) tests of vignette equivalence reasonably well. However, all subdomains fail stricter (LR-based) tests of VE, at least when all ten countries are included. When analyses are limited to specific subsets

of countries, they may fail to reject VE, or the deviations from VE may appear substantively small. Nonetheless, our results make clear that vignette equivalence cannot be assumed, especially across diverse sets of countries.

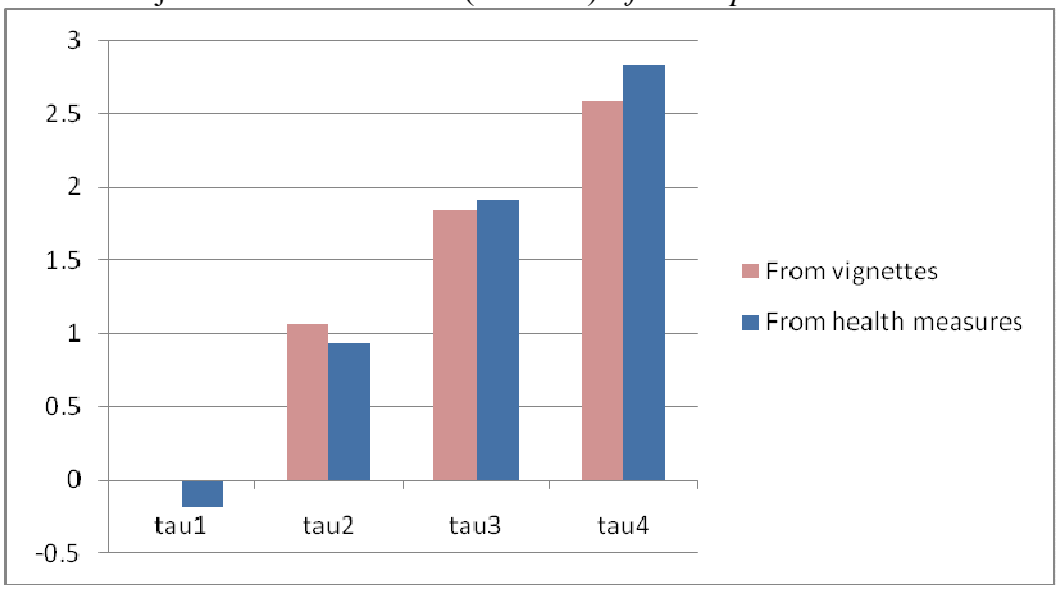
Tests of response consistency

As described earlier, we could not use the Bago D'Uva (2009) global test of response consistency in our full-sample analyses, as the test is not appropriate when vignette equivalence is violated. (In the single case when it was appropriate, namely, a test of "moving around" for Ghana and South Africa only, RC was rejected [df=48, LR test statistic=12,166; p<.001].)

We thus base our assessment of RC on a visual comparison of 1) cutpoints generated from anchoring vignette ratings, and 2) cutpoints generated from self-ratings paired with objective measures of health. These represent cutpoints estimated from Model A and Model C, respectively. In raw model output, the units for the two sets of cutpoints are different (namely, the unit for Model A cutpoints is the standard deviation of the omitted vignette; the unit for Model C cutpoints is the standard deviation of the self-rating). For the sake of comparability, Model C units were converted to Model A units. A constant was also added to cutpoints predicted by Model C to better align the two sets of cutpoints and further facilitate comparison. The graphs in this section reflect these conversions.

As shown in Figure 10, the cutpoints predicted by the two models for distance vision look extremely similar in a full sample analysis, with the slope for the health measure-based cutpoints only slightly higher than that for the vignette-based cutpoints. The two sets of cutpoints thus appear to show impressively concordant shapes, despite being calculated from entirely different types of data.

Figure 10: Predicted cutpoint locations for distance vision, from vignettes (Model A) and from objective health measures (Model C)—*full sample*.



The full-sample graph masks some heterogeneity among countries. For example, concordance is very high for India (Figure 11), but much lower for Russia (Figure 12). Response consistency may thus be more problematic in some regions than in others.

Figure 11: Predicted cutpoint locations for distance vision, from vignettes (Model A) and from objective health measures (Model C)—*India*.

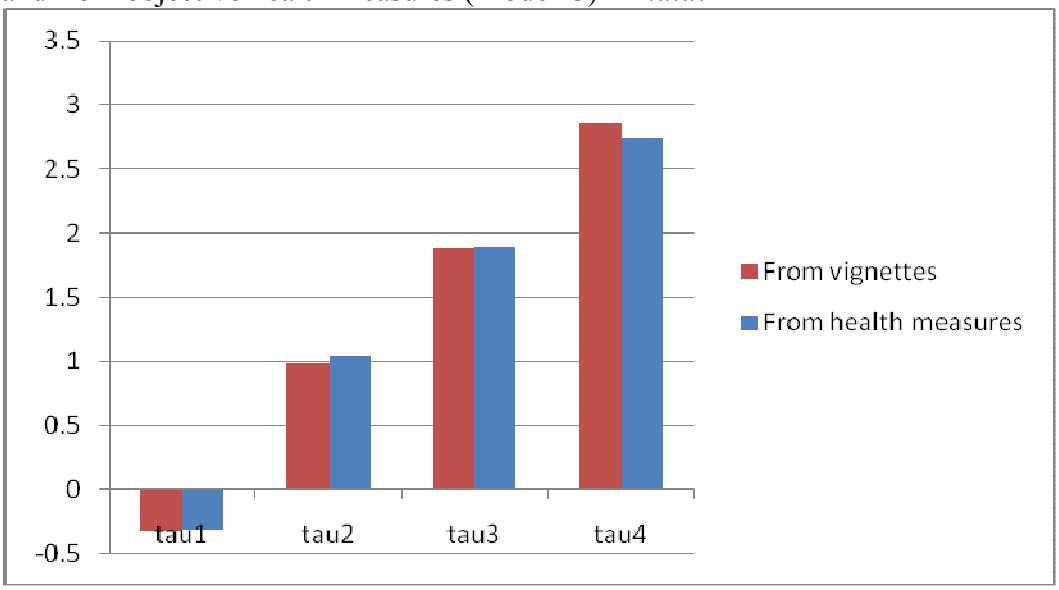
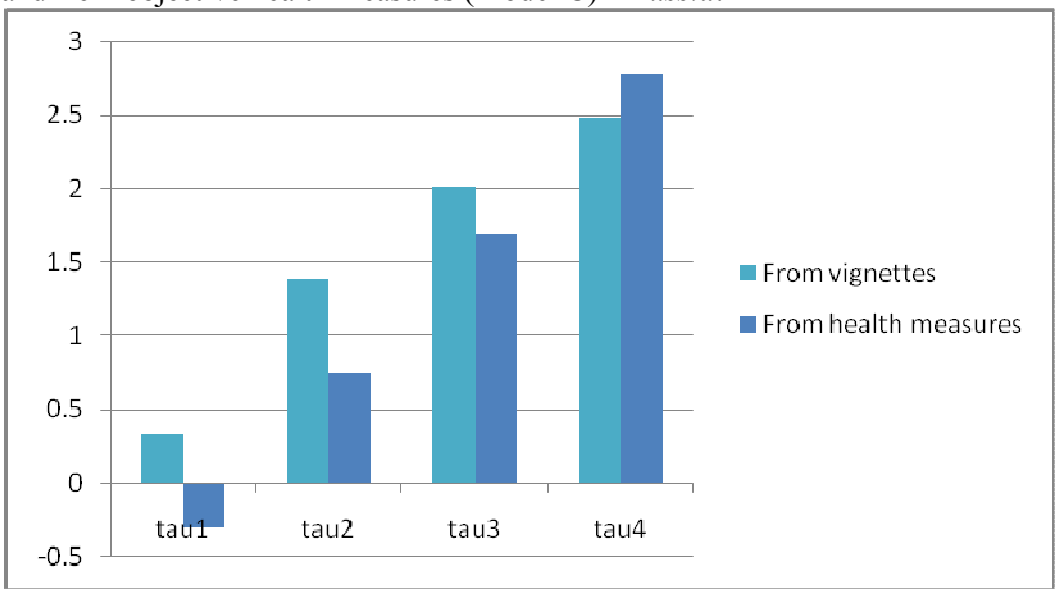


Figure 12: Predicted cutpoint locations for distance vision, from vignettes (Model A) and from objective health measures (Model C)—*Russia*.



Full-sample results for "moving around" are shown in Figure 13. Again, the distances between cutpoints are extremely similar in the two sets of data, suggesting adherence to the assumption of response consistency. While Russia showed violations of RC for distance vision, its "moving around" results are more in line with RC, as shown in Figure 14. Response consistency, it appears, may vary across health domain for a given country.

Figure 13: Predicted cutpoint locations for "moving around", from vignettes (Model A) and from objective health measures (Model C)—*full sample*.

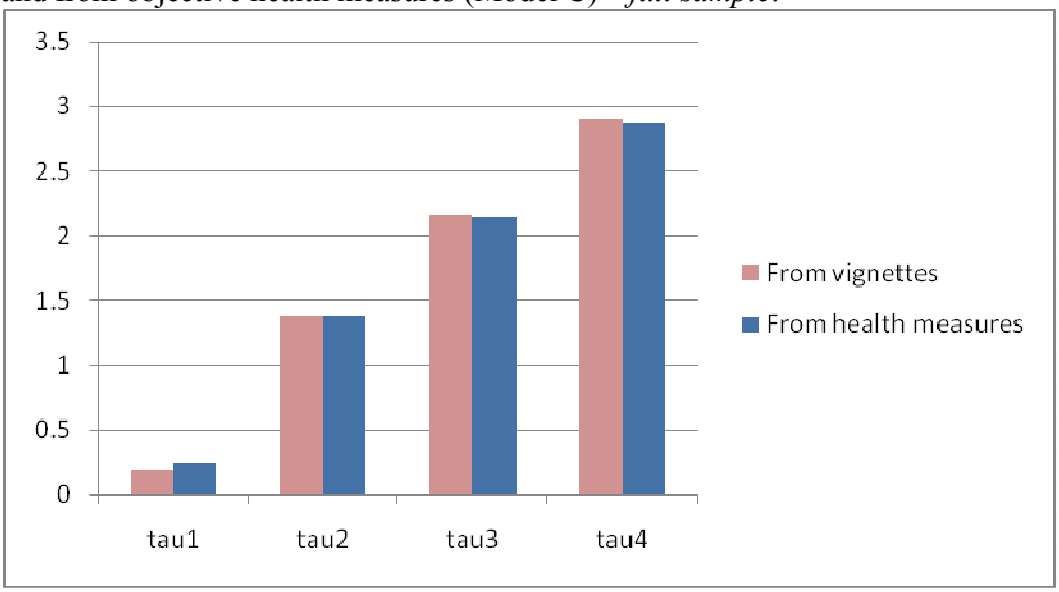
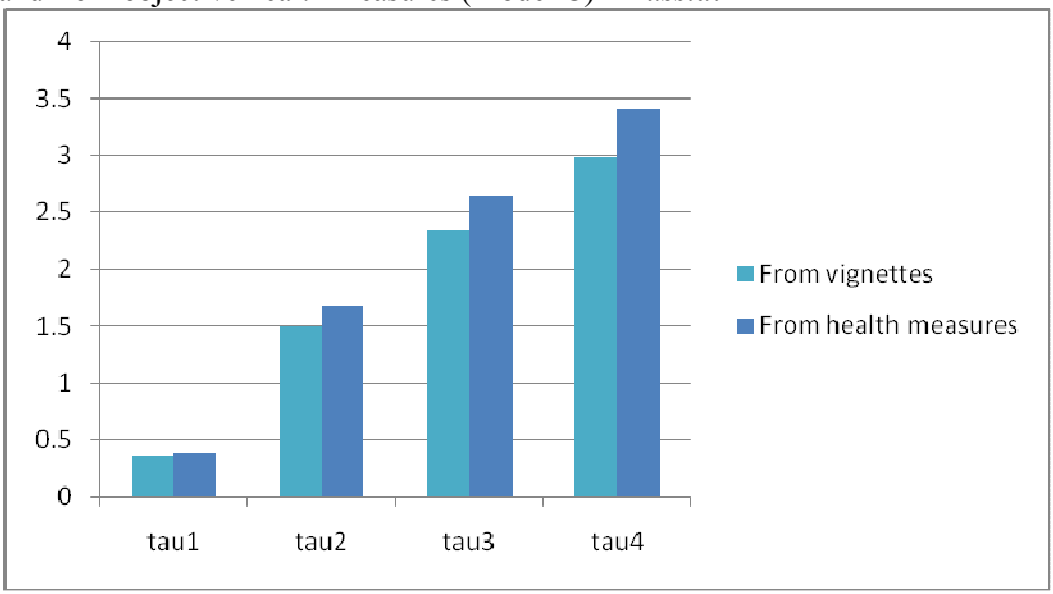


Figure 14: Predicted cutpoint locations for "moving around", from vignettes (Model A) and from objective health measures (Model C)—*Russia*.



In sum, subjective evaluations of the alignment of cutpoints (derived from vignettes vs. derived from objective health measures) are encouraging regarding adherence to response

consistency. In most individual SAGE countries, and in the SAGE sample as a whole, the two sets of cutpoints take on very similar shapes (i.e., the distances between comparable cutpoints appear very similar). The limitation of this visual test of response consistency is that, in the absence of vignette equivalence, we cannot be sure how the two cutpoint distributions actually line up vertically. Apart from this, results are quite promising.

SUMMARY AND DISCUSSION

Our results show that, while the WHO's health-related vignettes appear to pass "weak" (rank-ordering-based) tests of vignette equivalence (VE) reasonably well, they routinely fail stricter tests positing equidistance between perceived vignette locations across countries. That is, respondents in different countries appear to understand base vignette texts as representing fundamentally different levels of health. We identified some subsets of countries that, for specific health domains, appear to adhere to or only minimally violate the stricter version of VE, but these were relatively rare exceptions. Our tests of response consistency (RC) were more encouraging, often showing a striking concordance between cutpoints generated from vignette ratings and cutpoints generated from objective measures of health. Overall, our findings suggest that violations of VE are more egregious and more likely to undermine the anchoring vignette method than violations of RC. Existing WHO vignettes may still be useful in specific, limited applications, but their validity should be tested first rather than simply assumed. In brief, these anchoring vignettes must be used with caution.

We note that while our focus in this paper has been on anchoring vignettes as a tool to enhance cross-national comparability, vignettes could potentially have many useful applications within individual countries, to adjust for reporting heterogeneity between the sexes and/or across

age groups, educational groups, or other socioeconomic strata. Vignettes might also be fruitfully used in within-person (longitudinal) analyses (as in Angelini et al. 2010). Vignette validity may be higher in such contexts; in particular, one might imagine that vignette equivalence is less likely to be violated in within-person analyses, though this has not yet been empirically tested. The techniques reviewed in this paper could be used for precisely such testing, that is, to identify those contexts in which vignettes are most likely to correctly adjust for reporting heterogeneity.

Going forward, what could be done to improve the validity of new fieldings of vignettes? Recent studies make a number of suggestions in this regard. King and Wand (2007) present statistical techniques to aid in selecting optimally informative vignettes, while rejecting less useful ones, thereby minimizing the rank-order ties and inconsistencies that result from a crowded vignette field. Based on recent experimental findings, Hopkins and King (2010) argue that placing self-assessment questions immediately after vignette assessments improves response consistency, by "clarify[ing] the meaning of the self-assessment question and familiariz[ing] the respondents with the response scale" (208). Grol-Prokopczyk et al. (2011) argue against the mention of specific diseases or conditions (such as high blood pressure) in vignettes, since personal or familial experiences with such conditions appears to lead to different evaluations of such vignettes. (We note that the WHO vignettes do mention specific conditions, including stroke and obesity.)

As this last example suggests, attending closely to vignette wording may be a key to improving vignette equivalence. Despite the great importance of well-worded vignettes that both accurately capture the trait of interest and do so in a universally comprehensible way, vignette studies to date have almost without exception analyzed vignettes in the aggregate, without examination of individual vignette texts. This is despite the fact that, a priori, we can imagine

many reasons that vignettes mentioning obesity, or suicide, or pain caused by excessive computer use—as the WHO vignettes do—would be interpreted differently by some national, religious, or socioeconomic groups than by others. Methods described in this paper could be used to test, e.g., whether vignettes mentioning suicide are interpreted differently in Catholic or highly religious countries than in others, or whether obesity is interpreted differently in contexts of predominant overnutrition versus in contexts of food insecurity. Given the problems with vignette equivalence demonstrated here, such careful attention to avoid culturally-specific references is warranted. Whether the very sorts of cultural and linguistic differences that lead to differences in uses of response categories can be overcome in interpretations of vignette texts remains an open question—one which existing vignettes, it appears, have not yet been optimally designed to answer.

REFERENCES

Angelini, Viola, Danilo Cavapozzi, and Omar Paccagnella. 2010. "Dynamics of work disability reporting in Europe". Paper presented at the Royal Statistical Society conference on "Anchoring Vignettes in Social Science Research", London, U.K., 17 November 2010. Available at http://membership.rss.org.uk/main.asp?group=&page=1321&event=1194&month=11&year=2010&date=17%2F11%2F2010.

Bago D'Uva, Teresa, Maarten Lindeboom, Owen O'Donnell, and Eddy van Doorslaer. 2009 (November). "Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity". HEDG Working Paper 09/30. Available at http://www.york.ac.uk/res/herc/documents/wp/09_30.pdf.

Banks, James, Michael Marmot, Zoë Oldfield, and James P. Smith. 2007. "The SES Health Gradient on Both Sides of the Atlantic". No. WP07/04. The Institute for Fiscal Studies, UCL (University College London). Available at http://eprints.ucl.ac.uk/2653/1/2653.pdf.

Datta Gupta, Nabanita, Nicolai Kristensen, and Dario Pozzoli. 2010. "External Validation of the Use of Vignettes in Cross-Country Health Studies". *Economic Modelling* 27:854–865.

Grol-Prokopczyk, Hanna, Jeremy Freese, and Robert M. Hauser. 2011. "Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health." *Journal of Health & Social Behavior* 52(2). NIHMSID: NIHMS258919. In press; available in interim as working paper at http://www.ssc.wisc.edu/cde/cdewp/2010-09.pdf.

Hopkins, Daniel J. and Gary King. 2010. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly* 74(2): 201–222.

Inglehart, Ronald, and Christian Welzel. 2005. *Modernization, Cultural Change and Democracy*. New York: Cambridge University Press.

Jürges, Hendrik. 2007. "True Health vs Response Styles: Exploring Cross-country Differences in Self-Reported Health". *Health Economics* 16(2):163-178.

Jylhä, Marja, Jack M. Guralnik, Luigi Ferrucci, Jukka Jokela, and Eino Heikkinen. 1998. "Is Self-Rated Health Comparable Across Cultures and Genders?" *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 53(3):S144-S152.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004 (Feb). "Enhancing the Validity and Cross-Cultural Comparability of Survey Research". *American Political Science Review* 98(1):191-207.

King, Gary and Jonathan Wand. 2007 (Winter). "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes". *Political Analysis*15(1):46-66.

Kristensen, Nicolai and Edvard Johansson. 2008. "New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes". *Labour Economics* 15(1):96-117.

Murray, Christopher J.L., Ajay Tandon, Joshua A. Salomon, Colin D. Mathers, and Ritu Sadana. 2002. "New approaches to enhance cross-population comparability of survey results". Ch. 8.3 (pp. 421-431) in Murray, Christopher J.L., Joshua A. Salomon, Colin D. Mathers, and Alan D. Lopez (eds.), *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. 2002. Geneva: World Health Organization.

Murray, Christopher J.L., Emre Özaltin, Ajay Tandon, Joshua A. Salomon, Ritu Sadana, and Somnath Chatterji. 2003. "Empirical Evaluation of the Anchoring Vignette Approach in Health Surveys". Ch. 30 (pp. 369-399) in Murray, Christopher J.L., and David B. Evans (eds.). *Health Systems Performance Assessment: Debates, Methods and Empiricism*. 2003. Geneva: World Health Organization.

Rabe-Hesketh, Sophia and Anders Skrondal. 2002. "Estimating Chopit Models in gllamm: Political Efficacy Example from King et al." Retrieved January 20, 2009 (http://www.gllamm.org/chopit.pdf).

Rice, Nigel, Silvana Robone, and Peter Smith. 2009. "Analysis of the Validity of the Vignette Approach to Correct for Heterogeneity in Reporting Health System Responsiveness." HEDG Working Paper 09/28.

Sadana, Ritu, Colin D. Mathers, Alan D. Lopez, Christopher J. L. Murray, and Kim Moesgaard Iburg. 2002. "Comparative Analyses of More than 50 Household Surveys on Health Status". Ch. 8.1 (pp. 369-386) in Murray, Christopher J.L., Joshua A. Salomon, Colin D. Mathers, and Alan D. Lopez (eds.), *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. 2002. Geneva: World Health Organization.

Tandon, Ajay, Christopher J.L Murray, Joshua A. Salomon, Gary King. 2003. "Statistical Models for Enhancing Cross-Population Comparability". Ch. 55 (pp. 727-741) in Murray, Christopher J.L., and David B. Evans (eds.). *Health Systems Performance Assessment: Debates, Methods and Empiricism*. 2003. Geneva: World Health Organization.

Van Soest, Arthur, Liam Delaney, Colm Harmon, Arie Kapteyn, James P. Smith. 2007. "Validating the Use of Vignettes for Subjective Threshold Scales." *IZA Discussion Paper* No. 2860.

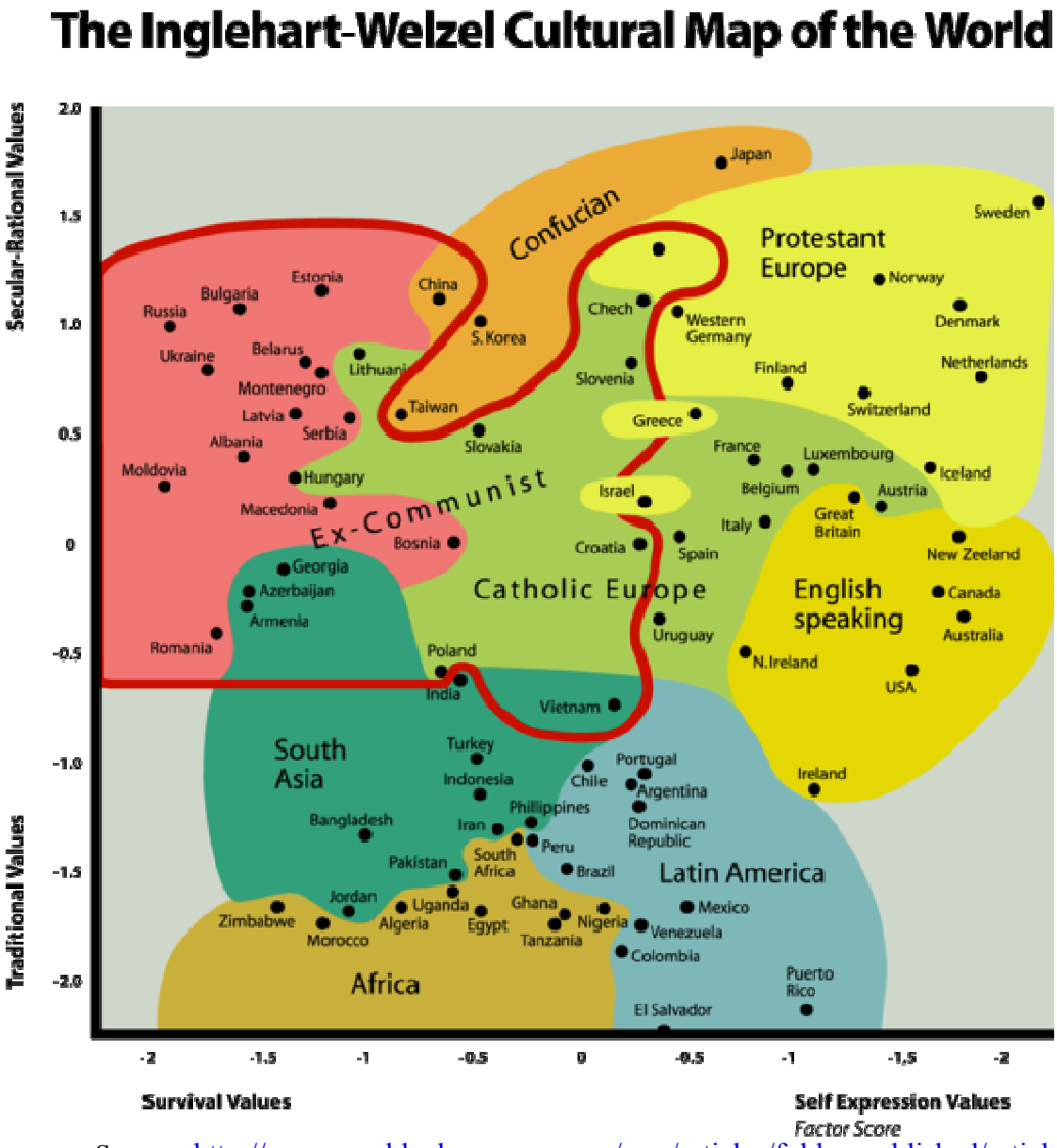
Zimmer, Zachary, Josefina Natividad, Hui-Sheng Lin, and Napaporn Chayovan. 2000. "A Cross-National Examination of the Determinants of Self-Assessed Health". *Journal of Health and Social Behavior* 41(4):465-481.

APPENDIX A: Texts of mobility, vision, cognition, and affect vignettes from the WHO's World Health Survey (WHS) and Study on Global AGEing and Adult Health (SAGE Wave 1).

Domain / severity	Vignette text
Mobility, severity 1	[Mary] has no problems with walking, running or using her hands, arms and legs. She jogs 4 kilometres twice a week.
Mobility, severity 2	[Yusuf] is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometre or climbing up more than one flight of stairs. He has no problems with day-to-day physical activities, such as carrying food from the market.
Mobility, severity 3	[Margaret] does not exercise. He cannot climb stairs or do other physical activities because he is obese. He is able to carry the groceries and do some light household work.
Mobility, severity 4	[Gabriel] has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feel heavy.
Mobility, severity 5	[Abdul] is paralyzed from the neck down. He is unable to move his arms and legs or to shift body position. He is confined to bed.
Rating question 1	Overall in the last 30 days, how much of a problem did [name of person] have with moving around?
Rating question 2	In the last 30 days, how much difficulty did [name of person] have in vigorous activities, such as running 3 km (or equivalent) or cycling?
Vision, severity 1	[Hector] can read words in newspaper articles (and can recognize faces on a postcard size photograph). He can recognize familiar people's faces all the time and picks out most details in pictures from across 20 metres.
Vision, severity 2	[Antonio] can read words in newspaper articles (and can recognize faces on a postcard size photograph). He can recognize shapes and colours from across 20 metres but misses out the fine details.
Vision, severity 3	[Norman] needs a magnifying glass to read small print and look at details on pictures. He also takes a while to recognize objects if they are too far from him.
Vision, severity 4	[Jennifer] only reads if the text is in very large print, such as 10 lines per page. Otherwise she does not read anything. Even when people are close to her, she sees them blurred.
Vision, severity 5	[Sebastian] cannot detect any movement close to the eyes or even the presence of a light.
Rating question 1	In the last 30 days, how much difficulty did you think [name of person] have in seeing and recognizing a person she knows across the road (i.e. from a distance of about 20 meters)?
Rating question 2	In the last 30 days, how much difficulty did you think [name of person] have in seeing and recognizing an object at arm's length or in reading?
Cognition, severity 1	[Rob] is very quick to learn new skills at his work. He can pay attention to the task at hand for long uninterrupted periods of time. He can remember names of people, addresses, phone numbers and such details that go back several years.

Cognition, severity 2	[Malcolm] can concentrate while watching TV, reading a magazine or playing a game of cards or chess. He can learn new variations in these games with small effort. Once a week he forgets where his keys or glasses are, but finds them within five minutes.
Cognition, severity 3	[Sue] can find her way around the neighborhood and know where her own belongings are kept, but struggles to remember how to get to a place she has only visited once or twice. She is keen to learn new recipes but finds that she often makes mistakes and has to reread several times before she is able to do them properly.
Cognition, severity 4	[Theo] cannot concentrate for more than 15 minutes and has difficulty paying attention to what is being said to him. Whenever he starts a task, he never manages to finish it and often forgets what he was doing. He is able to learn the names of people he meets but cannot be trusted to follow directions to a store by himself.
Cognition, severity 5	[Peter] does not recognize even close relatives and gets lost when he leaves the house unaccompanied. Even when prompted, he shows no recollection of events or recognition of relatives. It is impossible for him to acquire any new knowledge as even simple instructions leave him confused.
Rating question 1	Overall in the last 30 days, how much difficulty did [name] have with concentrating or remembering things? [None, mild, moderate, severe, extreme/cannot do?]
Affect, severity 1	[Samson] loves life and is happy all the time. He never worries or gets upset about anything and deals with things as they come.
Affect, severity 2	[Jane] enjoys her work and social activities and is generally satisfied with her life. She gets depressed every 3 weeks for a day or two and loses interest in what she usually enjoys but is able to carry on with her day to day activities.
Affect, severity 3	[Lucas] feels nervous and anxious. He worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests him. When he is alone he tends to feel useless and empty.
Affect, severity 4	[Susan] feels depressed most of the time. She weeps frequently and feels hopeless about the future. She feels that she has become a burden on others and that she would be better dead.
Affect, severity 5	[Scholastica] has already had five admissions into the hospital because she has attempted suicide twice in the past year and has harmed herself on three other occasions. She is very distressed every day for the most part of the day, and sees no hope of things ever getting better. She is thinking of trying to end her life again.
Rating question 1	Overall in the last 30 days, how much of a problem did [name] have with feeling sad, low, or depressed? [None, mild, moderate, severe, extreme/cannot do?]
Rating question 2	In the last 30 days, how much of a problem did [name of person] have with worry or anxiety? [None, mild, moderate, severe, extreme/cannot do?]

APPENDIX B: Inglehart-Welzel Cultural Map of the World.



Source: http://www.worldvaluessurvey.org/wvs/articles/folder_published/article_base_54 (cf. Inglehart and Welzel 2005:64).

APPENDIX C: Predictors of perceived vignette position for "moving around" subdomain. (Fuller version of Table 6).

	Ordered probit	
	β	SE
Severity 1	4.225***	.086
Severity 2	2.742***	.073
Severity 3	2.005***	.069
Severity 4	1.350***	.067
Sev 1 × Female	.069	.041
Sev 1 × Age 50-59	-.006	.057
Sev 1 × Age 60-69	.012	.061
Sev 1 × Age 70-79	-.214**	.068
Sev 1 × Age 80+	-.295**	.098
Sev 1 × Some Primary	.034	.066
Sev 1 × Primary Completed	-.144*	.061
Sev 1 × Secondary Completed	.102	.067
Sev 1 × High School Completed	.161*	.069
Sev 1 × College Completed	.336***	.089
Sev 1 × China	.490***	.083
Sev 1 × France	.075	.168
Sev 1 × UK	.939***	.165
Sev 1 × Ghana	-.257**	.094
Sev 1 × India	-1.371***	.076
Sev 1 × Mexico	-1.578***	.102
Sev 1 × Netherlands	-.221	.143
Sev 1 × Russia	1.309***	.118
Sev 1 × South Africa	-.279**	.103
Sev 2 × Female	.033	.035
Sev 2 × Age 50-59	-.102*	.049
Sev 2 × Age 60-69	-.048	.053
Sev 2 × Age 70-79	-.251***	.059
Sev 2 × Age 80+	-.257**	.085
Sev 2 × Some Primary	-.041	.056
Sev 2 × Primary Completed	-.113*	.053
Sev 2 × Secondary Completed	.027	.058
Sev 2 × High School Completed	.073	.058
Sev 2 × College Completed	.200**	.075
Sev 2 × China	.659***	.069
Sev 2 × France	-.265*	.130
Sev 2 × UK	.851***	.140
Sev 2 × Ghana	.134	.080
Sev 2 × India	-.787***	.065
Sev 2 × Mexico	-.962***	.088
Sev 2 × Netherlands	-.421***	.118
Sev 2 × Russia	.846***	.094
Sev 2 × South Africa	.167	.088
Sev 3 × Female	-.007	.031
Sev 3 × Age 50-59	-.083	.046
Sev 3 × Age 60-69	-.075	.050
Sev 3 × Age 70-79	-.131*	.055
Sev 3 × Age 80+	-.201*	.080
Sev 3 × Some Primary	-.022	.053
Sev 3 × Primary Completed	-.132**	.050
Sev 3 × Secondary Completed	-.032	.054

Sev 3 × High School Completed	-.046	.055
Sev 3 × College Completed	.132	.070
Sev 3 × China	.369***	.064
Sev 3 × France	-.360**	.122
Sev 3 × UK	.363**	.129
Sev 3 × Ghana	-.368***	.074
Sev 3 × India	-.482***	.062
Sev 3 × Mexico	-.960***	.084
Sev 3 × Netherlands	-.499***	.111
Sev 3 × Russia	.793***	.089
Sev 3 × South Africa	-.199*	.082
Sev 4 × Female	.008	.031
Sev 4 × Age 50-59	-.020	.044
Sev 4 × Age 60-69	-.018	.048
Sev 4 × Age 70-79	-.051	.053
Sev 4 × Age 80+	-.085	.078
Sev 4 × Some Primary	-.031	.051
Sev 4 × Primary Completed	-.112*	.048
Sev 4 × Secondary Completed	.010	.052
Sev 4 × High School Completed	.043	.053
Sev 4 × College Completed	.043	.053
Sev 4 × China	.141*	.062
Sev 4 × France	.405**	.123
Sev 4 × UK	.871***	.128
Sev 4 × Ghana	-.182*	.073
Sev 4 × India	-.653***	.060
Sev 4 × Mexico	-.487***	.083
Sev 4 × Netherlands	.068	.110
Sev 4 × Russia	.436***	.085
Sev 4 × South Africa	-.037	.080

Note: Perceived position of vignettes is calculated relative to the Severity 5 vignette. Other omitted reference categories are male (for sex), under age 50 (age), no formal schooling (education), and Brazil (country). Data generated by Model B hopit regression.