Using anchoring vignettes to assess the comparability of self-rated feelings of sadness, lowness or depression in France and Vietnam

G. EMMANUEL GUINDON¹,² & MICHAEL H. BOYLE³,⁴

¹ Département d'administration de la santé, Université de Montréal, Québec, Canada
² Centre for Health Economics and Policy Analysis, McMaster University, Hamilton, Ontario, Canada
³ Offord Centre for Child Studies, McMaster University, Hamilton, Ontario, Canada
⁴ Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada

Key words
France, Vietnam, depressive disorders, health measurement, reporting heterogeneity, self-reported health, vignettes

Abstract
General measures of self-rated health are collected routinely in national health surveys and widely used in the analyses of determinants of health and health care utilization. However, these subjective assessments can be influenced by health expectations (contextualized beliefs about health) that may vary systematically across individuals and be associated with their socio-demographic characteristics. Our objective is to contrast the impact of health expectations associated with respondent characteristics (reporting heterogeneity) on self-rated feelings of sadness, lowness or depression obtained from general population samples in France and Vietnam. Based on self ratings and ratings in response to common anchoring vignettes depicting different levels of depression, we used nationally representative data from the World Health Survey conducted in France (2002) and Vietnam (2002–2003) and a modification of the standard probit model to test and adjust for reporting heterogeneity associated with individual characteristics. We find evidence of reporting heterogeneity within France and Vietnam and across the two countries. In particular we find that, when adjusted for reporting heterogeneity, sex is no longer significantly associated with self-rated feelings of sadness, lowness or depression in France. Given the absence of clear biological markers in the definitions of depressive disorders and the substantial impact reporting heterogeneity is shown to have, measures of depressive disorders based on self-reports should be interpreted with caution. Copyright © 2012 John Wiley & Sons, Ltd

Background and objectives
General measures of self-rated health are collected routinely in national health surveys and widely used in the analyses of determinants of health and health care utilization. For example, the World Health Survey conducted by the World Health Organization (WHO) has the question, “In general, how would you rate your health today?” with response
Self-rated feelings of sadness, lowness or depression

categories “very bad, bad, moderate, good, very good”. This type of subjective assessment is easy to obtain and often used as a proxy indicator for more objectives measures of health that would be too onerous or impractical to collect. However, these subjective assessments can be influenced by health expectations (contextualized beliefs about health) that may vary systematically across individuals and be associated with their socio-demographic characteristics. These unobserved differences in subjective health ratings (reporting heterogeneity) may lead to distortions when comparing individuals or populations (King et al., 2004). This in turn could undermine cross-national comparison studies of population health and the measurement of important indicators such as health inequality (Bago d’Uva et al., 2008b).

Differences in the reporting of self-rated health can occur for a variety of reasons and be associated with socio-demographic features such as sex, age, place of residence or education. For example, the degree of mobility considered to be good health might decline as one gets older, leading to higher ratings (Salomon et al., 2004). Similarly, the conceptions of good health may vary by sex. The direction of this impact is unclear as it may differ by dimensions of health (e.g. physical versus mental health) (Bago d’Uva et al., 2008b).¹

Figure 1 presents the distribution of health across two different hypothetical populations (e.g. across two countries or across two hypothetical groups within the same country). The probability density function of the continuous health variable in both populations is identical suggesting individuals in both populations are in equivalent health (i.e. feelings of sadness, lowness or depression are the same). However, individuals across the two populations exhibit different response distributions when asked to evaluate their own health using a five-point ordinal scale (extreme, severe, moderate, mild, none). Individuals in population A (Figure 1) have a more positive view of their self-rated health (i.e. the cut-points are to the left of those of population B, indicating that more individuals in population A report being in better health even though there is no difference in the true health distribution between the two populations). Such cut-point shifts are not merely theoretical. For example, Canadians relative to Americans tend to avoid providing extreme responses (i.e. excellent/none and poor/extreme) (Evans, 2007). In Figure 1, such tendency would be represented by a shift outwards of the first and last cut-points of Canadians, relative to Americans. Other well known examples include the propensity for Aboriginals in Australia to report their own health as fair or poor to a lesser degree than the general Australian population even though the Aboriginal population scores worse on an array of more objective measures of morbidity and mortality (Mathers and Douglas, 1998). There is also the case of India where a strong negative association exists at the state level between reported morbidity and life expectancy (Sen, 2002).

The use of anchoring vignettes – fixed descriptions of a health domain rated on the same response continuum – has been proposed to identify and adjust for differences in expectations for health (King et al., 2004; Salomon et al., 2004; Tandon et al., 2003). To identify differences in expectations, survey respondents are asked to rate identical vignettes applicable to several hypothetical individuals using standard response options (e.g. extreme, severe, moderate, mild, none). To adjust for differences in expectations, respondents use these same standard responses to rate their own health on an identical question (King et al., 2004). Box 1 presents five anchoring vignettes from the World Health Survey which seek to measure a particular dimension of mental health. Variation in the rating of such vignettes can be attributed to differences in reporting since the health of hypothetical individuals described in the vignettes is the same for all individuals. These differences can then be inspected in relation to observed characteristics such

¹ See Bago d’Uva et al. (2008b) and Salomon et al. (2004) for a discussion of the likely direction of reporting heterogeneity across other characteristics such as residence, income and education.

Figure 1 Comparing self-reported health across two hypothetical populations: (A) population A; (B) population B. Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?
as sex, age, residence, income or education (Bago d’Uva et al., 2008b).

Box 1. Questions in the World Health Survey

Overall in the last 30 days, how much of a problem did [you/name] have with feeling sad, low or depressed? Extreme, Severe, Moderate, Mild, None.

Vignettes

[Ken] loves life and is happy all the time. He never worries or gets upset about anything and deals with things as they come.

[Henriette] enjoys her work and social activities and is generally satisfied with her life. She gets depressed every three 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.

[Jan] 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.

[Robert] 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.

[Vivian] 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.

The use of anchoring vignettes to identify systematic differences in reporting requires both response consistency and vignette equivalence (King et al., 2004; Murray et al., 2003; Salomon et al., 2004):

- Vignette equivalence requires that the health descriptions represented in each vignette are understood in the same way by all respondents, irrespective of their age, sex, education, country of residence, or other characteristics, apart from random measurement error.
- Response consistency requires individuals use the response categories for a particular question in a similar way when responding about their own health and when responding to a vignette about the health of hypothetical individuals.

It is the requirement of vignette equivalence that allows the identification of systematic variation in the ordinal reporting of a given level of function. If individual characteristics such as sex influence the understanding of each health description, then it is not possible to attribute differing vignette responses as differences in health expectations (Bago d’Uva et al., 2008b). The requirement of response consistency allows the variation to be used to remove reporting differences from self-report measures of health (Bago d’Uva et al., 2008b).

The vignette approach has been used recently to test for reporting heterogeneity in several areas. Kapteyn et al. (2007) examined self-reported work disability rates between the Netherlands and the United States and found heterogeneity in reporting across both countries and within the United States across categories of sex, age and education; they concluded overall that about half the difference in work disability rates could be attributed to heterogeneity in reporting. Lardjane and Dourgnon (2007) explored cross-country heterogeneity in the reporting of physical pain across eight European countries (Belgium, France, Germany, Greece, Italy, Netherlands, Spain and Sweden) and found evidence of heterogeneity, most notably for the Netherlands and Sweden. Bago d’Uva et al. (2008b) explored reporting heterogeneity across six dimensions of health (mobility, cognition, pain, self-care, usual activities, affective behaviour) in three Asian countries (China, India and Indonesia) and found some evidence of reporting heterogeneity across all demographic and socio-economic characteristics for all countries and health dimensions.

The inclusion in the World Health Survey of hypothetical anchoring vignettes that describe identical mental health states provides an opportunity to estimate the influence of health expectations on self-ratings of mental health. Accordingly, the objective of this study is to examine the comparability of self-rated feelings of sadness, lowness or depression across a sample of individuals from two heterogeneous countries – France and Vietnam – and across demographic and socio-economic characteristics within countries with an emphasis on the role played by sex. This study adds to the contribution of Bago d’Uva et al. (2008b) in the following ways. First, it compares self-rated feelings of sadness, lowness or depression within countries as well as across countries. Second, it uses data from national population samples. Third, it focuses specifically on heterogeneity in...
the self-reporting of depressive disorders and pays special attention to differing health expectations across sex. Heterogeneity in the reporting of mental health status is of particular interest given the substantial burden of disease associated with mental disorder and large differences in reported prevalence. For example, depression accounted for nearly 12% of all total years lived with disability worldwide in 2000 (Ustun et al., 2004) and the WHO World Mental Health Surveys estimated large cross-national differences in the prevalence of anxiety disorders: 18.2% in the United States, 5.8% in Italy and 3.3% in Nigeria. Estimates of the severity of disorders also show substantial differences across countries (Demyttenaere et al., 2004) as do differences in the prevalence of mental disorders by sex (Angst et al., 2002; Nolen-Hoeksema, 1990; Ustun et al., 2004; WHO International Consortium in Psychiatric Epidemiology, 2000; Wittchen and Jacobi, 2005). Estimating the extent to which mental disorders such as depression are influenced by health expectations could have important implications for our understanding of these conditions as well as their role in accounting for disability.

**Data**

World Health Survey data were used. Because of their striking differences in prevalence of mental health disorders and levels of economic development, we selected France and Vietnam to illustrate the potential importance of reporting heterogeneity. In both countries, probability sampling was used to identify and select individual respondents – one per household – for a personal interview conducted in the home. Response rates were high: at the household level they exceeded 99% in both countries; at the individual level, they were 83.6% (n = 3490) in Vietnam and 99.3% (n = 1001) in France. For more details about the World Health Surveys’ methodology, see Usten et al. (2003).

The measures of own health and “vignette health” are detailed in Box 1. The five response categories for both own health and vignettes are: extreme, severe, moderate, mild and none. Only a random sub-sample of individuals were asked the vignette questions on mental health (251 in France and 856 in Vietnam). In all vignettes, country and sex specific first names were used to match the country of residence and sex of respondents. We examined differences in reporting across the following demographic and socio-economic characteristics: sex, age, urban/rural status, education and income. Age was measured in years. We followed Bago d’Uva et al. (2008b) and used a flexible measure of education through a series of four dummies representing the highest level of education completed: less than primary (reference category), primary, secondary and high school or above. Such dummies allow the relationship between education and self-rated health to be non-linear. The income variable was specified as the natural logarithm of total household spending in the last four weeks adjusted for household size (number of adults in household + 0.5 × number of children in household)$^{0.75}$ and expressed in international dollars (IMF, 2009).

**Methods**

First, we examined the frequency of depressive episodes by sex for both France and Vietnam. The classifications of depression into “brief episode” and “major” represent an adaptation of the International Classification of Diseases, 10th revision (ICD-10) research diagnostic criteria applied to questions about depressed mood, loss of interest, loss of energy, loss of appetite and slowed thinking, taken from the Composite International Diagnostic Interview (Kessler and Ustun, 2004) for use in the World Mental Health Survey. Second, we explored graphically the variation in vignette ratings across sex within France and Vietnam and overall across France and Vietnam. Third, we followed the approach suggested by Tandon et al. (2003) and implemented recently by Bago d’Uva et al. (2008b) which proposed a modification of the standard probit model – the hierarchical ordered probit – that incorporates information obtained through the use of vignettes. There are two components. One, subject ratings in response to the standard vignettes are used to model the cut-points as a function of the covariates. In contrast to the standard probit model, the cut-points are allowed to vary as a function of the covariates. This modification makes it possible: (1) to test for homogeneity of responses across cut-points, and (2) to test for parallel cut-point shift in responses (i.e. whether or not the shift in response is equal across all cut-points). Two, subject self-ratings of their

---

3 Bago d’Uva et al. (2008b) focus on income and education.

4 The World Health Survey has been conducted in 70 countries. The analysis of 70 large national datasets is beyond the scope of this paper.
own health are used to adjust for reporting heterogeneity to remove the effect of health expectations. Details of this approach appear in the Appendix.

All statistical analyses were conducted using Stata/SE 10.0 for Macintosh with statistical codes developed by Bago d’Uva and colleagues (Jones et al., 2007) and listwise deletion of missing data (i.e. individuals with missing data on any of own health, covariates or vignettes are dropped). As recommended by Kapteyn et al. (2007), the two components of the model were estimated using a one step procedure.

**Results**

Table 1 presents the frequency of depressive episodes and distribution of responses to self-rated mental health and descriptive statistics for covariates. Striking differences exist between France (8.5%) and Vietnam (0.4%) in the estimated 12-month prevalence of major depressive disorder. Within-country differences in prevalence between women and men are also evident (0.6 versus 0.1% in Vietnam and 10.9 versus 5.0% in France). Consistent with these differences in prevalence, women in both Vietnam and France (76.9 and 56.4%) are less likely than men (82.5 and 72.1%) to report the highest health category (i.e. no feeling of sadness, lowness or depression in the last 30 days).

Figure 2 presents unadjusted ratings for self-assessments and for all five vignettes in France and Vietnam. Differences in self-assessed health are large. Residents of Vietnam are less likely to rate their own health as being poor (i.e. use response categories “extreme” and “severe”). Reporting heterogeneity across France and Vietnam is evident for some, but not all vignettes. The Wilcoxon–Mann–Whitney test indicates that there are statistically significant differences in the response distributions associated with the vignettes Henriette and Jan ($z = 5.66, p < 0.000; z = -2.541, p = 0.011$). In testing for variation in vignette ratings by sex in France and Vietnam, we find evidence of reporting heterogeneity by sex in France for some vignettes but none in Vietnam (not shown). The Wilcoxon–Mann–Whitney test indicates that there are statistically significant difference by sex in the response distributions associated with the vignettes Roberta and Vivian ($z = 2.84, p = 0.005; z = -2.68, p = 0.007$).

**Table 1** Descriptive statistics

<table>
<thead>
<tr>
<th></th>
<th>Vietnam</th>
<th></th>
<th>France</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female (%)</td>
<td>Male (%)</td>
<td></td>
<td>Female (%)</td>
</tr>
<tr>
<td><strong>Depressive episodes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brief episode</td>
<td>2.4</td>
<td>0.9</td>
<td>11.3</td>
<td>8.4</td>
</tr>
<tr>
<td>Major episode</td>
<td>0.6</td>
<td>0.1</td>
<td>10.9</td>
<td>5.0</td>
</tr>
<tr>
<td><strong>Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extreme</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Severe</td>
<td>1.0</td>
<td>0.7</td>
<td>4.3</td>
<td>2.5</td>
</tr>
<tr>
<td>Moderate</td>
<td>5.3</td>
<td>3.3</td>
<td>13.0</td>
<td>8.5</td>
</tr>
<tr>
<td>Mild</td>
<td>16.7</td>
<td>13.4</td>
<td>25.6</td>
<td>16.8</td>
</tr>
<tr>
<td>None</td>
<td>76.9</td>
<td>82.5</td>
<td>56.4</td>
<td>72.1</td>
</tr>
<tr>
<td><strong>Covariates</strong></td>
<td>n</td>
<td>Percentage</td>
<td>n</td>
<td>Percentage</td>
</tr>
<tr>
<td>Sex: Female</td>
<td>1916</td>
<td>55.0</td>
<td>573</td>
<td>60.5</td>
</tr>
<tr>
<td>Residence: Urban</td>
<td>748</td>
<td>21.5</td>
<td>517</td>
<td>54.6</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than primary</td>
<td>743</td>
<td>21.3</td>
<td>23</td>
<td>2.4</td>
</tr>
<tr>
<td>Primary</td>
<td>816</td>
<td>23.4</td>
<td>141</td>
<td>14.9</td>
</tr>
<tr>
<td>Secondary</td>
<td>1111</td>
<td>31.9</td>
<td>209</td>
<td>22.1</td>
</tr>
<tr>
<td>High school and above</td>
<td>813</td>
<td>23.3</td>
<td>574</td>
<td>60.6</td>
</tr>
<tr>
<td>Mean Age (in years)</td>
<td>40.0</td>
<td>14.2</td>
<td>43.9</td>
<td>16.7</td>
</tr>
<tr>
<td>Household income (in natural log)</td>
<td>5.0</td>
<td>0.6</td>
<td>10.0</td>
<td>3.8</td>
</tr>
</tbody>
</table>

N = 3483 947
Table 2 presents the results of tests of homogeneity (i.e. whether or not differences in the cut-points for self-reported health arising from the standard vignettes are associated with individual covariates). The data presented are \( p \)-values of Wald tests of significance for each covariate for the four cut-points. When all covariates are considered jointly, the results provide evidence of reporting heterogeneity in both France and Vietnam. When covariates are examined individually, the results provide evidence of reporting heterogeneity by sex, urban/rural status and income in France but not in Vietnam and by education attainment in both France and Vietnam (the education dummies are jointly significant, \( \chi^2(12) = 36.1, p = 0.0003; \ \chi^2(12) = 22.6, p = 0.0314 \), respectively).

Table 2 also presents the results of tests of parallel cut-point shift. Tests for parallel cut-point shift examine whether or not the effect of a covariate is equal across all thresholds (i.e. if reporting heterogeneity is stronger at some levels of health than others). Parallel cut-point shift for all covariates is rejected in France but not in Vietnam (first column in Table 4). This suggests that the covariates do not affect all cut-points by the same magnitude in France: heterogeneity in the reporting of health is stronger at some levels of health status than others. For example, there is evidence of non-parallel cut-point shift for urban residency, education and income in France. Although the omnibus test is non-significant in Vietnam, there is evidence of non-parallel cut-point shift for educational attainment.

When both country samples are pooled to allow testing for heterogeneity in the reporting of health between France and Vietnam, homogeneity of reporting and parallel cut-point shift by country is rejected (final column in Table 2).

Table 3 presents the estimated shifts at each of the four cut-points by sex within Vietnam and France and between thresholds (i.e. if reporting heterogeneity is stronger at some levels of health than others). Parallel cut-point shift for all covariates is rejected in France but not in Vietnam (first column in Table 4). This suggests that the covariates do not affect all cut-points by the same magnitude in France: heterogeneity in the reporting of health is stronger at some levels of health status than others. For example, there is evidence of non-parallel cut-point shift for urban residency, education and income in France. Although the omnibus test is non-significant in Vietnam, there is evidence of non-parallel cut-point shift for educational attainment.

When both country samples are pooled to allow testing for heterogeneity in the reporting of health between France and Vietnam, homogeneity of reporting and parallel cut-point shift by country is rejected (final column in Table 2).

Table 3 presents the estimated shifts at each of the four cut-points by sex within Vietnam and France and between
the two countries. Higher health expectations are represented by positive shifts in the cut-points (a positive shift in cut-point corresponds to a lower probability of endorsing a higher level). Individuals with higher expectations for health are less likely to endorse positive ratings for a standard vignette (e.g. more likely to identify a low level of sadness, lowness or depression in the present study). For example, in France, the covariate “female” has positive coefficients across all cut-points (Table 3). It indicates that women have higher health expectations. In the pooled sample, the country dummy representing France has a positive shift at cut-point one and negative shifts at cut-points three and four. This indicates that respondents residing in France, relative to Vietnam, are more likely to rate a given vignette as corresponding to extreme feeling of sadness or depression and are more likely to rate a given vignette as corresponding to mild or no feeling of sadness or depression.

Table 4 presents estimated coefficients before and after adjustment. In order to be comparable, the before adjustment coefficients are multiplied by the estimated scale parameters from the hierarchical probit model (Jones et al., 2007). Quantitatively, each coefficient represents a change in z-score; qualitatively, a positive coefficient indicates a stronger probability of reporting a higher category of health status. The first column (before) represents the association between covariates and self-assessed health status assuming homogeneous reporting behaviour (i.e.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Female</th>
<th>Urban</th>
<th>Education (2)</th>
<th>Education (3)</th>
<th>Education (4)</th>
<th>Income</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnam Homogeneity</td>
<td>0.012</td>
<td>0.823</td>
<td>0.065</td>
<td>0.092</td>
<td>0.005</td>
<td>0.023</td>
<td>0.484</td>
<td>0.061</td>
</tr>
<tr>
<td>Vietnam Parallel cut-point shift</td>
<td>0.105</td>
<td>0.679</td>
<td>0.341</td>
<td>0.692</td>
<td>0.002</td>
<td>0.013</td>
<td>0.349</td>
<td>0.165</td>
</tr>
<tr>
<td>France Homogeneity</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000</td>
<td>0.585</td>
<td>0.033</td>
<td>0.348</td>
<td>0.519</td>
<td>0.010</td>
</tr>
<tr>
<td>France Parallel cut-point shift</td>
<td>0.000</td>
<td>0.168</td>
<td>0.001</td>
<td>0.418</td>
<td>0.018</td>
<td>0.333</td>
<td>0.402</td>
<td>0.007</td>
</tr>
<tr>
<td>France/Vietnam Homogeneity</td>
<td>0.000</td>
<td>0.453</td>
<td>0.037</td>
<td>0.763</td>
<td>0.078</td>
<td>0.042</td>
<td>0.642</td>
<td>0.642</td>
</tr>
<tr>
<td>France/Vietnam Parallel cut-point shift</td>
<td>0.000</td>
<td>0.851</td>
<td>0.470</td>
<td>0.980</td>
<td>0.197</td>
<td>0.023</td>
<td>0.573</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Note: n/a, not applicable. Education 1 = less than primary (reference category), Education 2 = primary, Education 3 = secondary and Education 4 = high school or above.

<table>
<thead>
<tr>
<th></th>
<th>Cut-point 1</th>
<th>Cut-point 2</th>
<th>Cut-point 3</th>
<th>Cut-point 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Vietnam</td>
<td>0.040</td>
<td>0.010</td>
<td>0.009</td>
<td>−0.045</td>
</tr>
<tr>
<td></td>
<td>(0.59)</td>
<td>(0.19)</td>
<td>(0.17)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Female France</td>
<td>0.284</td>
<td>0.145</td>
<td>0.210</td>
<td>0.340</td>
</tr>
<tr>
<td></td>
<td>(2.43)</td>
<td>(1.56)</td>
<td>(2.34)</td>
<td>(3.52)</td>
</tr>
<tr>
<td>Country Vietnam/France</td>
<td>0.260</td>
<td>−0.004</td>
<td>−0.156</td>
<td>−0.395</td>
</tr>
<tr>
<td></td>
<td>(2.49)</td>
<td>(−0.04)</td>
<td>(−1.96)</td>
<td>(−4.70)</td>
</tr>
</tbody>
</table>

Note: t-ratios in parentheses. Italic indicates significance at 5%. Cut-point one is the lowest cut-point determining probability of extreme distress, cut-point four is the highest cut-point determining probability of no sadness, lowness or depression.
imposing the same cut-points). These coefficients reflect both the true health effects and the effects of reporting heterogeneity. Coefficients presented in the second column (after) are adjusted for reporting heterogeneity. Before adjustment, the ordered probit results indicate significant negative associations between sex and the measure of mental health status in both France and Vietnam. When adjusted for reporting heterogeneity, sex is no longer significantly associated with health status in France.

Partial effects – change in the predicted probabilities for a given change in an explanatory variable, while holding all other variables constant – have more intuitive interpretation than changes in z-scores and can be calculated for each health category and for each individual. Females versus males, in the absence of any adjustments, have a lower probability of reporting the highest health category (none) and a higher probability of reporting the lowest three categories combined: extreme, severe, moderate. Before adjustments, being female decreases the probability of reporting the highest health category by 4.8 and 15.5% in Vietnam and France, respectively. After adjustments, being female increases the probability by 1.0 and 3.3%, respectively.

We also calculate the partial effects of residing in France versus Vietnam on the probability of reporting the highest health category (none) and the lowest three categories combined. On average we find, before adjustments, that residing in France relative to Vietnam, decreases the probability of reporting the highest health category by 24.5%. After adjustments, residing in France decreases the probability by 31.6%. Finally, before adjustments, residing in France increases the probability of reporting the bottom three health categories by 12.2% and after adjustments, residing in France increases the probability by 15.2%.

**Discussion**

**Principal findings**

We have examined the comparability of self-rated feelings of sadness, lowness or depression across a sample of individuals from two heterogeneous countries – France and Vietnam – and across demographic and socio-economic characteristics within these two countries. We found evidence of reporting heterogeneity in both France and Vietnam and across the

---

Table 4 Purging reporting bias: estimated coefficients before and after adjustment

<table>
<thead>
<tr>
<th></th>
<th>Vietnam Before</th>
<th>Vietnam After</th>
<th>France Before</th>
<th>France After</th>
<th>Vietnam/France Before</th>
<th>Vietnam/France After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>−0.233</td>
<td>−0.267</td>
<td>−0.473</td>
<td>−0.174</td>
<td>−0.289</td>
<td>−0.235</td>
</tr>
<tr>
<td></td>
<td>(−3.65)</td>
<td>(−3.26)</td>
<td>(−5.21)</td>
<td>(−1.39)</td>
<td>(−5.56)</td>
<td>(−3.50)</td>
</tr>
<tr>
<td>Urban</td>
<td>−0.044</td>
<td>0.100</td>
<td>−0.183</td>
<td>0.123</td>
<td>0.097</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>(−0.51)</td>
<td>(0.92)</td>
<td>(−2.10)</td>
<td>(1.00)</td>
<td>(−1.64)</td>
<td>(0.78)</td>
</tr>
<tr>
<td>Age</td>
<td>−0.018</td>
<td>−0.014</td>
<td>−0.002</td>
<td>0.002</td>
<td>−0.013</td>
<td>−0.012</td>
</tr>
<tr>
<td></td>
<td>(−8.27)</td>
<td>(−4.69)</td>
<td>(−0.49)</td>
<td>(0.38)</td>
<td>(−7.69)</td>
<td>(−5.33)</td>
</tr>
<tr>
<td>Education (2)</td>
<td>0.341</td>
<td>0.464</td>
<td>0.002</td>
<td>−0.444</td>
<td>0.325</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>(3.91)</td>
<td>(3.88)</td>
<td>(0.01)</td>
<td>(−1.14)</td>
<td>(4.05)</td>
<td>(2.53)</td>
</tr>
<tr>
<td>Education (3)</td>
<td>0.279</td>
<td>0.464</td>
<td>0.373</td>
<td>0.652</td>
<td>0.317</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>(3.29)</td>
<td>(4.13)</td>
<td>(1.33)</td>
<td>(1.67)</td>
<td>(4.11)</td>
<td>(4.76)</td>
</tr>
<tr>
<td>Education (4)</td>
<td>0.445</td>
<td>0.500</td>
<td>0.351</td>
<td>0.428</td>
<td>0.412</td>
<td>0.423</td>
</tr>
<tr>
<td></td>
<td>(4.72)</td>
<td>(3.83)</td>
<td>(1.27)</td>
<td>(1.13)</td>
<td>(4.88)</td>
<td>(3.86)</td>
</tr>
<tr>
<td>Income</td>
<td>0.074</td>
<td>−0.033</td>
<td>0.045</td>
<td>0.072</td>
<td>0.058</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>(1.39)</td>
<td>(−0.48)</td>
<td>(3.90)</td>
<td>(4.49)</td>
<td>(4.68)</td>
<td>(4.51)</td>
</tr>
<tr>
<td>Country</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>−0.953</td>
<td>−1.270</td>
</tr>
</tbody>
</table>

Note: t-ratios in parentheses; Italic indicates significance at 5%. n/a = not applicable. Education 1 = less than primary (reference category), Education 2 = primary, Education 3 = secondary and Education 4 = high school or above.
two countries. For example, we found that on average and after controlling for age, urban/rural status, education and income, before adjustments, being female increases the probability of reporting the bottom three health categories by 1.8 and 9.3% in Vietnam and France, respectively. After adjustments, being female increases the probability by 1.0 and 3.3%, respectively. Of particular interest is the finding that, when adjusted for reporting heterogeneity, sex is no longer significantly associated with self-rated feelings of sadness, lowness or depression in France. Whether this heterogeneity stems from symptom “over-reporting” among females or “under-reporting” by males is not possible to determine in the present study.

Bago d’Uva et al. (2008b) explore reporting heterogeneity across six dimensions of health including experiences of distress, sadness or worry within three Asian countries (China, India and Indonesia) and find evidence of reporting heterogeneity across all socio-demographic characteristics for all countries and health domains. They also reject parallel cut-point shift in all but one case and conclude socio-demographic characteristics shift the thresholds by different magnitudes and in different directions. Our results for our tests of homogeneity and parallel cut-point shift for Vietnam are largely consistent with those of Bago d’Uva et al. (2008b) for China and the health domain that examines experiences of distress, sadness or worry. However, the effect of adjusting for reporting heterogeneity on the association between education and income (before and after adjustments) and experiences of distress, sadness or worry are not in agreement. Using data from the Survey of Health, Ageing and Retirement in Europe (SHARE) covering eight countries, Bago d’Uva et al. (2008a) examine differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. The authors explore six health domains including feelings of sadness, lowness or depression and find that reporting differences generally lead to an underestimation of socio-economic inequalities in health.

Gender differences in depression are well documented (Angst et al., 2002; Nolen-Hoeksema, 1990; Ustun et al., 2004; WHO International Consortium in Psychiatric Epidemiology, 2000; Wittchen and Jacobi, 2005), and a variety of substantive reasons for these differences have been presented, including: adverse experiences in childhood, depression and anxiety disorders in childhood and adolescence, sociocultural roles with related adverse experiences, and psychological attributes related to the vulnerability to life events and coping skills (Piccinelli and Wilkinson, 2000). However, genetic and biological factors do not appear to explain gender differences. In the absence of clear biological markers, the classification of depression will continue to rely on subjective responses to questions about symptoms of depression and the use of arbitrary decision rules to define “cases”. In addition to substantive reasons for sex differences in depression, it appears that reporting heterogeneity (male–female differences in health expectancies) may also contribute to observed gender differences in the measured prevalence of depression.

Study limitations

First, as a measure of mental health status we use self-reported feelings of sadness, lowness or depression obtained through the use of a single question. Such a measure may not be a good proxy for depressive episodes or more general measures of mental health status. As such, our variable for feelings of sadness, lowness or depression cannot represent a diagnosis of depression. However, in classifying depression, similar questions are used in standard psychiatric interviews as devices for screening respondents into or away from subsequent probes. Second, the number of individuals surveyed in Vietnam is substantially larger than the number surveyed in France, leading to imbalances in statistical power and the ability to detect reliable effects that favour tests done in Vietnam. However, given the substantially larger (and statistically significant) effect sizes found in France, this limitation does not undermine our general conclusions.

Unanswered questions and future research

The results of this study illustrate the advantages of using anchoring vignettes to identify differences in demographic and socio-economic characteristics in the reporting of health. World Health Survey data are a rich source of information. In addition to studying heterogeneity in the reporting of depressive episodes, these data can be used to examine several other dimensions of health including mobility, pain, cognition, self-care and personal relationships. Further exploitation of the World Health Survey – for which 70 datasets exist – can shed additional light on the benefits of the World Health Survey data for understanding the determinants of self-reported health across a large number of low-, middle- and high-income countries.

Because the use of anchoring vignettes to identify systematic differences in reporting requires both response consistency and vignette equivalence, there is a need to test formally the validity of these assumptions (Bago d’Uva et al., 2008b). Van Soest et al. (2007) formally examine the assumption of response consistency...
using a self-reported measure of drinking behaviour along with an objective measure and set of anchoring vignettes in a sample of college students in Dublin and find that vignettes perform very well in correcting for differences in response scales. Similar tests of response consistency in other health domains are needed.

Given our findings of heterogeneity in the reporting of feelings of sadness, lowness or depression in both France and Vietnam and the substantial impact reporting heterogeneity is shown to have on the association between sex and self-assessed health status, more research is warranted.

**References**


**Acknowledgements**

The authors thank the WHO for making these data available, Teresa Bago d’Uva for providing public access and support to hierarchical probit stata codes, Somrath Chatterji for providing statistical codes for the depression algorithm and Jeremiah Hurley, Kathy Georgiades, Kristin Cleverley, members of McMaster University’s Polinomics Group and anonymous reviewers for their comments.

**Declaration of interest statement**

The authors have no competing interests.
Appendix

Ordered and hierarchical ordered probit

Tandon et al. (2003) propose a modification of the standard probit model that incorporates information obtained through the use of vignettes.

Ordered probit model

The ordered probit model assumes there is a variable \( Y^*_i \) (e.g., mental health status) that is not directly observed (i.e., a latent variable) with mean for each individual \( i \), \( \mu_i \), and variance one. The mean of the unobserved variable is a function of individual characteristics (\( Z_i \)) such as sex, age, residence, education and income.

\[
Y^*_i \sim N(\mu_i, 1) \quad i = 1, \ldots, N
\]

\[\mu_i = Z'_i \beta\]

Let \( y_i \) be a self-reported ordinal measure of mental health status for each individual \( i \); let \( k \) represent response categories; and let \( \tau^k \) represent response cut-points. Thus, \( y_i \) is generated by the unobserved health status variable \( Y^*_i \) such that:

\[
y_i = k \text{ if } \tau^{k-1} \leq Y^*_i < \tau^k
\]

for \( k = 0, \ldots, K \)

\[
\tau^0 < \tau^1 < \cdots < \tau^{K-1} < \tau^K
\]

\( \tau^0 = -\infty \) and \( \tau^K = \infty \)

The parameters for \( \beta \) and the cut-points \( \tau^k \) can then be estimated. The key assumption of the ordered probit model that is of interest to us is the assumption that the same cut-points apply for each individual in the sample. Put another way, ordered probit models assume homogenous reporting. If it can be shown that the cut-points vary across some individual characteristics such as sex or residence, then the estimates of \( \beta \) will be shifted systematically. This is the case because the estimates of \( \beta \) will reflect not only health effects but also the effect of reporting (Tandon et al., 2003).

Hierarchical ordered probit

The approach proposed by Tandon et al. (2003) has two components. First, vignette information is used to model the cut-points as a function of covariates. Let \( Y^*_{ij} \) represent the unobserved health level of vignette \( j \) as perceived by individual \( i \).

\[
Y^*_{ij} \sim N(\mu_{ij}, 1) \quad i = 1, \ldots, N; j = 1, \ldots, V
\]

\[\mu_{ij} = J'_i \xi\]

where \( J_i \) is a vector of indicator variables for each \( V - 1 \) vignettes. Similar to the standard ordered probit model, \( y_{ij} \), the observed ordered response by individual \( i \) to vignette \( j \), is generated by the unobserved health status of vignette \( j \), \( Y^*_j \) such that:

\[
y_{ij} = k \text{ if } \tau^{k-1} \leq Y^*_j < \tau^k
\]

for \( k = 0, \ldots, K \)

\[
\tau^0 < \tau^1 < \cdots < \tau^{K-1} < \tau^K
\]

\( \tau^0 = -\infty \) and \( \tau^K = \infty \)

In contrast to the standard ordered probit model, the cut-points \( \tau^k \) are allowed to vary as a function of covariates:

\[
\tau^k = X_{ij} \gamma^k
\]

This modification makes it possible to test for reporting homogeneity across cut-points. Log-likelihood ratio tests or Wald tests of significance can be used to test for reporting homogeneity across a covariate (or a group of covariates) (Bago d’Uva et al., 2008b; Jones et al., 2007). This modification also makes it possible to test whether the effect across a covariate (or a group of covariates) is equal across all cut-points (i.e., parallel cut-point shift). If reporting heterogeneity is stronger at some levels of health than other, then cut-point shift is not parallel (Bago d’Uva et al., 2008b). Data from the random sub-sample of individuals that are asked
the vignette questions are used to estimate the first component of the hierarchical model.

Second, information from the individual’s self-report question is used to define the unobserved level of individual health $Y_i^{*}$ and the mechanism that links $Y_i^{*}$ to the observed ordinal variable, $y_i$. Formally,

$$Y_i^{*} \sim N\left(\mu_i, \sigma^2\right) \quad i = 1, \ldots, N$$

$$\mu_i = Z_i \beta$$

Let $y_i^*$ represent the observed ordinal responses on the self-report such that:

$$y_i^* = k \quad \text{if} \quad \tau_{i}^{k-1} \leq Y_i^{*} < \tau_{i}^{k}$$

for $k = 0, \ldots, K$

$$\tau_0 < \tau_1 < \cdots < \tau_{K-1} < \tau_K$$

$$\tau_0 = -\infty \quad \text{and} \quad \tau_K = \infty$$

Data from the full sample are used to estimate the second component of the hierarchical model.