

ECuity III Project

Working Paper #1

Cut-point shift and index shift in self-reported health

By

Maarten Lindeboom *^a and Eddy van Doorslaer^b

March 2003

^a Department of Economics Free University of Amsterdam, Tinbergen Institute Amsterdam and Institute for the study of Labor (IZA) Bonn.

^b Department of Health Policy & Management, Erasmus University, 3000 DR Rotterdam, The Netherlands.

*Corresponding author. Tel. +31-20-4446033, Fax: +31-20-4446005, E-mail: mlindeboom@feweb.vu.nl. This paper derives from the project “The dynamics of income, health and inequality over the life cycle” (known as the ECuity III Project), which is funded in part by the European Community’s Fifth Framework Programme (contract QLK6-CT-2002-02297). We are grateful to the EC for financial support, to Statistics Canada for access to the NPHS data. We furthermore acknowledge the insightful comments of Andrew Jones on an earlier version of this paper.

Abstract

There is a concern that ordered responses on health questions may differ across populations or even across subgroups of a population. This reporting heterogeneity may invalidate group comparisons and measures of health inequality. This paper proposes a test for differential reporting in ordered response models which allows distinguishing between cut-point shift and index shift. The method is illustrated using Canadian National Population Health Survey data and the McMaster Health Utility Index (HUI) is used as a more objective health measure than the simple 5-point scale of self-assessed health. We find clear evidence of cut-point shifting for age and gender, but not for income, education or language.

JEL Classification: D30; D31; I10; I12

Keywords: hierarchical ordered probit; health measurement; cut point shift; index shift; Canada;

1. Introduction

One of the most commonly employed indicators of overall individual health in general population surveys is the simple question “How is your health in general?”, with response categories ranging from “very good” or “excellent” to “poor” or “very poor”. While this ordinal variable is known to be a very good predictor variable of other outcomes, like subsequent use of medical care or mortality, there is some concern that its comparability across populations – or population subgroups – is problematic because of a problem which has been termed ‘state-dependent reporting bias’ (Kerkhofs and Lindeboom, 1995), ‘scale of reference bias’ (Groot, 2000), response category cut-point shift’’ (Sadana *et al*, 2000, Murray *et al*, 2001), ‘reporting heterogeneity’ (Shmueli, 2002) or differential item functioning’’ (Hays *et al*, 2000). Basically, it occurs if sub-groups of a population use systematically different threshold levels for SAH, despite having the same level of ‘true’ health. These differences may be influenced by, among other things, age, sex, education, language and personal experience of illness. Basically, it means that different groups appear to ‘speak different languages’ and use different reference points when they are responding to the same question.

The problem may be particularly great for comparisons across cultural groups with different norm and expectations. Murray *et al* (2001), for example, quote some evidence from Australian national health surveys, aboriginals in Australia reporting better self-assessed health than the general Australian population while all other indicators (such as mortality) show this subgroup to be at a serious disadvantage with respect to health. Similarly, Sen (2002) recently quoted the well-known evidence from India again, where Kerala, the state with the highest life expectancy, consistently also shows the highest rates of morbidity. He concludes that: “There is a strong need for scrutinising statistics on self reported illness in a social context by taking note of levels of education, availability of medical facilities and public information on illness and remedy”. In our view, the best way to do this is to formalise the problem of heterogeneous reporting behaviour and to see if we can come up with tests that helps us in judging the reliability of subjective health information. This is exactly what we do in this paper.

We present a framework for individual reporting behaviour that enables us to formally test whether variations in responses to health questions reflect true health differences or reporting behaviour. Our health-reporting model also allows us to distinguish between different types of reporting heterogeneity: cut-point shift and index shift. Index shift refers to the situation where the individual response behaviour does not affect the shape of the SAH distribution, but rather its

location. Cut-point shift refers to the situation where the shape of the distribution of the SAH changes. Stated differently, in the case of an index shift the location of the reporting thresholds changes, but their relative position remains unaltered. In the case of a cut-point shift the relative positions of the reporting thresholds change.

Heterogeneous reporting behaviour may have important implications for the measurement and explanation of inequalities in health by, for instance, income or education. If, given the same level of ‘true’ health, the assessment of reported health differs systematically by socio-economic status, this will bias the measured degree of socio-economic inequality in health. The distinction between cut-point shift and index shift becomes relevant when one aims to unravel and quantify the contributions of various determinants of health to measured inequality. We will touch upon this in our closing section and conclude that estimates of our reporting model can be used to obtain a reliable estimates of (the decomposition of) the inequality in health.

As will be clear intuitively, a valid benchmark that can be used to validate the responses on SAH is essential to our approach. The WHO’s Evidence for Health Policy group (Tandon et al, 2002) has proposed the use of vignettes, whereby respondents are asked to not only rate their own health state (or its dimensions) but to evaluate also some health state descriptions of fictitious individuals. The health state evaluations of these fictitious individuals are then used to estimate individual specific reporting thresholds (cut-points). An obvious alternative is to use a more objective indicator of health and to condition on this measure while making comparisons across different socio-economic groups. This is the route that we will follow in this paper. The vignette approach has its advantages, but it is very demanding in terms of data collection and very often not a realistic option for researchers wanting to analyse existing surveys. Moreover, as we will discuss in more detail at the end of section 2, most likely the implementation of the vignette approach also requires a more objective source of information about the respondents’ health where one needs to condition on.

The approach adopted in this paper relates to Kerkhofs & Lindeboom (1995), who focussed on differential reporting behaviour across labour market status¹. In this paper we focus on differential response to SAH by background characteristics. We exploit the fact that the Canadian *National Population Health Survey 1994-95* is one of the few general population health surveys which contains both the simple SAH question and one of the well-known generic measures of health utility. This is the *McMaster Health Utility Index Mark III*, which is used, *inter alia*, for computing healthy life

¹ There is an issue in the retirement literature that subjective health measures may depend on the retirement status of the respondent. Individuals out of work may overstate health problems in order to justify inactivity or because receipt of social security benefits requires ill health. Kerkhofs and Lindeboom (1995) formulate a health-reporting model that allows for differential reporting across labour market status (state dependent

expectancies in Canada (Feeny *et al.*, 1995, Furlong *et al.*, 1998). While this measure also relies on self-reporting, one advantage is that respondents are only required to classify themselves on eight health attributes and the overall individual health utility score on a scale of zero to one is derived using weights which are derived from a different valuation survey on a different sample of individuals. As such, it represents a more valid and reliable general health measure than the single SAH question.

Ideally, one would prefer to use generic measures like the HUI to measure health differences across different socio-economic groups. However, generic health measures like the HUI are only rarely available in—and if so, often differ across—general socioeconomic household surveys, which makes comparisons across different populations impossible. By contrast, the general SAH question is very widely used. Performing tests for reporting heterogeneity in SAH using surveys with more abundant health information (such as the NPHS) provides information on the likelihood and size of the reporting heterogeneity problems in other surveys with simpler health measures².

The paper is organised as follows. Section 2 presents the theoretical model of reporting behaviour which underlies our testing strategy. We use a Canadian dataset is (described in section 3) to illustrate the reporting model and the testing procedures and report on this in section 4. We conclude in section 5. This section also contains a discussion on how the reporting model could be used for the measurement and decomposition in the inequality in health.

2. A model of reporting behaviour

The reported subjective health measure is denoted by H^s and is the respondent's answer to a question like e.g. “How good is your health in general?” with replies ranging from excellent, very good, good, fair, to poor. It is assumed that these responses are generated by a corresponding latent true health variable H^* . Rather than to one single measure, H^* could refer to a set of latent health measures covering the different domains of individual health. For ease of exposition we will restrict ourselves to a single health index. The reporting model relates the unobserved latent treat to the individual's responses to health questions. Of relevance for the issue of cross-population comparability is that the relationship

reporting behaviour). They found that Dutch recipients of Disability Insurance benefits strongly overstated their health problems.

² Simpler measures are, for instance, ADL measures, or whether or not the respondent suffers from a range of chronic conditions. On its own, each of these measures may not sufficiently well describe the general health status of an individual and it will be difficult to use the different measures simultaneously to describe health inequalities. Our method (see section 2) can deal with this ‘dimensionality’ problem in a natural way, as we condition on the objective health information and use it to proxy an underlying latent ‘true’ health variable.

between H^* and H^S may not be constant across populations and may even vary within populations. For instance, an individual with a university degree and a true health level H^* may state that his health is ‘fair’, whereas an otherwise identical but lower educated respondent may report to be in excellent health. Or stated more generally, the response labels may have a different meaning in different subgroups of the population.

The introduction of a true health concept H^* is essential for the issue of cross-population comparability. Any unconditional comparison of H^S (i.e. not controlling for H^*) across different groups of the population may in this case reflect both differences in health conditions as well as differences in reporting behaviour. It is therefore important to note that in practical situations one does not observe H^* , but instead has to rely on a range of more objective measures derived from tests on the presence of various health problems, diseases and health-related impediments in performing a range of daily activities. These could be used to proxy the (different domains of) true unobserved health (H^*).

The above results in the following relations of the health reporting model:

$$H^S = f_1(H^*, X_1, \mathbf{e}_1; \mathbf{b}_1) \quad (1a)$$

$$H^* = f_2(H^o, X_2, \mathbf{e}_2; \mathbf{b}_2) \quad (1b)$$

The variables \mathbf{e}_1 and \mathbf{e}_2 are random variables, $f_1(\cdot)$ describes the reporting behaviour and $f_2(\cdot)$ the relationship between the true health and its determinants. Consequently, the effect of X_1 (as embodied in the \mathbf{b}_1 ’s) could be viewed as pure reporting behaviour, whereas X_2 corrects for the dissimilarity between H^o and H^* . Note that f_2 will be an identity in the (ideal) case where H^* is adequately captured by H^o . H^* is by definition unobserved. The difficulty lies in the assessment of the relative importance of X_1 (reporting behaviour) versus X_2 (true health effects or heterogeneity in health) in the determination of H^S . It will depend crucially on how the empirical model is implemented.

We will often observe H^S as an ordered categorical variable. The essence of differential reporting, cut point shift or state-dependent reporting behaviour is that X_1 will not only affect the mean of the index function, but will, *in addition*, have a direct influence on the cut-points corresponding to the different response categories. X_2 is included to measure the dissimilarities between H^o and H^* and therefore its role is expected to be different. More specifically:

$$H^* = f(H^o; \alpha) + X_2' \beta_2 + \varepsilon_2 \quad (2)$$

The health response is now defined as:

$$H^S = i \Leftrightarrow c_{i-1} < H^* \leq c_i, \quad i = 1, \dots, n. \quad (3a)$$

Where n is the number of response categories. The cut-points c_i are allowed to vary with different values of X_1 :

$$c_i = g_i(X_1; \beta_i), \quad i = 1, \dots, n-1, c_0 = -\infty, c_n = \infty. \quad (3b)$$

In this model there may be differential response (H^S) for people with identical levels of true health. $g_i(\cdot)$ is left unspecified, but in its most flexible form each cut-point is affected by exogenous variables in a different way (\mathbf{b}_{li} , $i = 1, \dots, n-1$). Combining (2) and (3) we obtain:

$$H^S = i \Leftrightarrow g_{i-1}(X_1; \beta_{i-1}) - X_2' \beta_2 < f(H^o; \alpha) + \varepsilon_2 \leq g_i(X_1; \beta_i) - X_2' \beta_2 \quad (4a)$$

Now the role of X_1 and X_2 and the identification of their effects becomes clearer. The variables X_2 are part of the index function (cf. eq. 2) and determine the true level of health (H^*), whereas X_1 variables affect the thresholds and are allowed to have differential effects across thresholds. A variable that belongs in X_2 , but that was mistakenly included in X_1 will shift all threshold variables in the same way. This can easily be tested by imposing the restriction that the \mathbf{b}_{li} are equal for all i . If this restriction is rejected then the X_1 variable is (at least to some extent) responsible for cut-point shift (or state dependent reporting errors). This simple test will be important because our primary aim is to determine whether variations in responses reflect true health differences or reporting behaviour.

Our empirical strategy will be to estimate separate ordered response models for groups in our sample stratified according to age, sex, education, income and language. More specifically:

$$H^S_k = i \Leftrightarrow \mathbf{d}_{i-1}^k < f(H^o; \alpha^k) + \varepsilon_2^k \leq \mathbf{d}_i^k \quad (4b)$$

Note that this specification is more flexible than the usual hierarchical ordered response models like e.g. the HOPIT model. In these models one usually assumes a common vector α and imposes a functional form on the function g_i . For instance, Kerkhofs & Lindeboom (1995), Groot (2000) and Tandon *et al* (2002) take $g_i(X_1; \beta_i) = X_1' \beta_i$. In our specification of (4'), each subgroup k is allowed to differ *both* with respect to cut-points (\mathbf{d}_i 's) *and* with respect to index effects (α 's)³.

A test for differential response behaviour across different subgroups, i.e. different \mathbf{d}_i 's and/or different α 's, can be based on straightforward likelihood ratio tests. The tests compare the sum of the likelihood values for two subgroups k and k' with the likelihood values obtained using the total group (i.e. with equal \mathbf{d}_i 's and α 's imposed). In cases where equality of response behaviour for two groups is rejected, we can additionally test whether this is due to index shifts (different α 's) or to cut-point shifts

³ The separate δ 's in our specification do not impose the linear additive functional form and allow for interaction effects of all elements of X_1 .

(different d_i 's). The simplest way in which this additional test can be implemented is to estimate (4') subject to either the restriction that $d_i^k = d_i^{k'}$ or $\alpha^k = \alpha^{k'}$, $\forall i$, and to compare the likelihood value of the restricted model (L^R) with the sum of the likelihood values of the unrestricted models (L^U). A test is based on $-2 \ln(L^R / L^U)$ which is χ^2 distributed with degrees of freedom equal to the number of restricted parameters. Below we will illustrate these testing procedures with data from the Canadian *National Population Health Survey 1994-95*. Before we do this we would like to make four remarks.

Firstly, the core of the model lies in equation (1) and the equations that relate this model to the actual responses on the self-assessed health question (equations 3a and 4). We feel that this is the most natural way to implement the reporting behaviour. Given the model structure, the testing procedure remains identical, regardless of the estimation method used and as such it is immaterial whether one uses hierarchical probit, interval regression or other methods to estimate the parameters in (4a) or (4b).

Secondly, the question remains which variables should be included in the threshold equations and represent reporting heterogeneity (X_1) and which variables should be in the health index representing true health effects (X_2)⁴. X_1 and X_2 are only identified when they are related in a non-linear way (one can see this directly from equation 4a). This means that in the case that X_1 affects all thresholds in a similar way (index shift), identification crucially relies on the functional form of $g_i(X_1; \beta_{1i})$ and that without additional information, it is not possible to identify whether a variable should be in X_1 or X_2 . This has consequences for the interpretation of β_2 as in this case it may reflect both pure health effects (a variable is rightfully included in X_2) and index-shift reporting behaviour (a variable belongs to X_1 but is included in X_2). It may be good to add that the distinction between index shift and true health effects is of less relevance for the measurement and explanation of health inequalities across different sub-populations. We return to this at the end of this paper.

Thirdly, the response equation (4a) with individual specific thresholds can also be derived from a ordered response model with heteroskedastic error terms, regardless of whether there exist variations in cut-points across different socio-economic groups (see for instance van Doorslaer and Jones (2002)). This only matters for the interpretation of the cut-point functions $g_i(X_1; \beta_{1i})$ and it does not influence our testing procedure to distinguish between cut-point shift and index shift.

Fourthly, essential to our approach is a valid benchmark: the objective health measure H^P . Tandon et al (2002) propose to use vignettes to make the self-report responses comparable across different populations. The vignettes are included in the WHO Multi-Country Study and are designed

⁴ Similarly, it does not mean that variables in X_1 do not have an effect on true health (i.e. that they should not be included in X_2).

to capture the domain of mobility. Respondents are faced with different fictitious individuals who are mobility constraint in varied degrees. The descriptions of these individuals range from a perfectly healthy athlete to an almost fully paralysed individual. The respondents are asked to classify these vignettes on the same five-point response scale as the SAH measure. Differences in the response behaviour across different sub-populations are then used to identify variations in the cut-points. More specifically, Tandon et al (2002) model the responses to vignettes k , Y^k , as:

$$Y^s = i \Leftrightarrow c_{i-1} < Y^* \leq c_i, i = 1, \dots, n. \quad (5)$$

With the response thresholds as $c_i = g_i(X_1; \beta_{1i}) = X_1' \beta_{1i}$ and the index function as $Y^* = X_2' \beta_2 + \varepsilon, \varepsilon \sim N(0,1)$.

A first remark, regarding this model is that, as with our model, it is not clear which variables belong in the thresholds (X_1) and which variables in the index function (X_2). For this one has to rely on the testing procedure that we proposed above to distinguish between cut-point shift and index shifts. Furthermore, as a second remark, the rating of a specific health situation will most likely depend on the health status of the respondent. It is very likely that the view of very fit and healthy individuals on specific health states will differ from the views of individuals with severe physical impairments. If this is the case, then one should add the respondent's health status as a conditioning variable in X_2 . We then have, in essence, returned to our model summarised in (4a) and all remarks listed above will remain to hold⁵.

3. Data and variable definitions

The data used in this paper are taken from the first wave (held in 1994-1995) of the Canadian *National Population Health Survey (NPHS)* (cf. Tambay and Catlin, 1995). For our analyses we have included respondents aged between 20 and 70 years and have excluded cases with incomplete or inconsistent information on the relevant socio-demographic and health variables. The remaining sample size consists of 13699 observations.

The two key variables for this study are self-assessed health (SAH) and the more objective health status as measured by the Health Utility Index (HUI). In terms of the notation of the previous section, H^s equals SAH and H^o equals HUI. In the *NPHS*, respondents were asked: "In general, how would you say your health is?" The response categories were excellent, very good, good, fair, and

poor. Because only a relatively small fraction of the sample is in poor health, we combine the poor and fair category of the SAH variable. Also, each respondent was assigned a Health Utility Index score based on the responses to the questions of the eight-attribute Health Utility Index Mark III health status classification system. The Health Utility Index is a generic health status index, developed at McMaster University that measures both quantitative and qualitative aspects of health (Torrance *et al*, 1995, 1996; Feeny *et al*, 1995). It provides a description of an individual's overall functional health, based on eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. The Health Utility Index assigns a single numerical value, between zero and one, for all possible combinations of levels of these eight self-reported health attributes. A score of one indicates perfect health. The Health Utility Index also embodies the views of society concerning health status, inasmuch as preferences about various health states are elicited from a representative sample of individuals. More specifically, the HUI embodies information from vignettes, such as the ones proposed by Tandon *et al* (2002). This makes the HUI an almost ideal candidate for the objective health variable where we need to condition on (H^0).

Total income before taxes and deductions is measured in the NPHS as a categorical variable with 11 response categories. Education is measured as the highest level of general or higher education completed (5 levels). Table A1 provides some descriptive statistics of the relevant (but dichotomized) variables in our analysis. Reporting behaviour may differ across socio-economic status, but also across different cultural or ethnic groups. The Canadian NPHS can to some extent be used to test for cross-cultural differences in reporting behaviour by using language as a proxy for groups with different cultural backgrounds. We distinguish four language groups: those speaking English only, French only, those speaking both languages and those speaking another language only.

Table A2 provides cross-tabulations of mean HUI values by level of SAH and background characteristics. The tabulations show that, on average, the male, the younger, and the higher education and income respondents report higher health utility values, but conditional on reported level of SAH, the differences are much smaller. Only at lower levels of health, some differences remain, with particularly the female, older, and lower income respondents reporting lower mean HUI values than their counterparts. In the next section, we explain how we have formally tested for reporting heterogeneity.

4. Illustration: empirical implementation and results

⁵ One exception is the remark on the distinction between index shift and true health differences and the interpretation of β_2 . In the vignette model all variation can be ascribed to differential reporting behaviour.

Empirical implementation

We apply the procedure discussed in section 2 on the Canadian NPHS. Estimation of model (4') requires us to be more specific about the functional form of $f(H^o; \alpha^k)$ and the distribution of ε_2^k . We take $f(H^o; \alpha^k)$ as a quadratic function of the generic overall health status index (the Health Utility index, HUI). ε_2^k is taken to have a standard normal distribution. Note that our tests for differential response and the distinction between cut-point and index shift do not depend on these assumptions.

The strategy proposed in section 2 is to stratify the sample into (more) homogenous subgroups and to estimate ordered response models for each group. We stratify the sample according to language, age, gender, income and education. For each group, we obtain estimates of α^k (index function) and d_i^k (cut-points). We then aggregate over one of the stratifiers and estimate a model where the stratification variable is included in the index function. In this model it is assumed that there is a common effect of $f(H^o; \alpha^k)$ and that there are common cut-points.

For example, we observe four different language (or cultural) groups in the sample: French only (Fr), English only (En), French *and* English (Bo) and Other (Ot). We observe an indicator for objective health H^o and a response to a question concerning general health (H^S). The latter variable is an ordered categorical variable, with 4 response categories⁶. It is hypothesized that the different cultural backgrounds may lead to differential response behaviour, even if individuals have identical health (H^o). So, according to the differential response model (4'):

$$H^S_k = i \Leftrightarrow d_{i-1}^k < f(H^o; \alpha^k) + \varepsilon_2^k \leq d_i^k \quad (4')$$

For each k (Fr, En, Bo, Ot) we obtain a separate set of estimates of 8 (4*2) α^k 's and 12 (4*3) d_i^k 's. The null hypothesis of homogeneity is that there is no differential response with respect to the *index* as far as H^o is concerned (i.e. the translation between the objective health and the subjective health is identical across the groups) and the *cut-off points* (i.e. the response categories have the same meaning for all k = Fr, En, Bo, Ot groups):

$$H^S_k = i \Leftrightarrow d_{i-1} < f(H^o; \alpha) + \beta_1 Fr + \beta_2 En + \beta_3 Bo + \beta_4 Ot + \varepsilon_2 \leq d_i \quad (4'')$$

In this specification, the language variables are still included in the index, but this is not required⁷. Our test for differential response behaviour is based on the comparison of the likelihood value of the restricted model with 8 parameters (4''), with the unrestricted model with 20 parameters

⁶ The original variable contains 5 response categories ranging from excellent to poor. In the empirical analyses we have merged the response categories "fair" and "poor".

(4'). If model (4'') is rejected (i.e., if homogeneous response behaviour is rejected) a model could be estimated in which the index function is as in (4'''), but where differential (i.e. language specific) cut-off points are allowed. A comparison of the likelihood value of this model with (4'') reveals whether the differential response is due to differential cut-points or to an index shift.

This procedure was applied for five stratifying variables: four language groups (French, English, French and English and Other), two gender groups (male, female), two age groups (below and over 45), two educational groups (high and low), and two income groups (high and low). Low income is defined as an income less than Can \$ 24,748. A low education is defined as the highest level of educational attainment being secondary school or less.

Results

The likelihood test results are presented in Tables 1-5. We start by testing for homogeneity in response behaviour by language. Language is the closest we could get with this sample to differences in cultural backgrounds in Canada, and it has been argued that response behaviour is likely to differ most with respect to culture. We distinguish four language groups: people who speak English only, French only, both of these or another language. We do test this separately for 16 groups distinguished by 4 dichotomizations of income, education, age and gender. It can be seen from Table 1 that homogeneity is rejected in only 4 cases, three groups of males and one of females. Where homogeneity was rejected, we also tested to see whether the rejection was due to index shift or cut-point shift. In all of the male groups, it was only due to index shift. Only in the group of old females with low education and income, there was clear evidence of cut-point shift.

Next we aggregated across language groups and tested for response shift by income, while controlling for language in the index function only. The test results in Table 2 show that in one group only (young males with low education), homogeneity across income was rejected, and it was due entirely to index shift. Table 3 presents the same results for education. It is striking to see that for none of the four demographic groups we find any evidence of response shifting by education. Apparently, the relationship between HUI and SAH is similar for those with high and low education.

⁷ Of course also a normalisation is required.

All in all, for the two indicators of socioeconomic status, we find little or no evidence of response shift.

The picture is quite different for age and gender, as can be seen from Tables 4 and 5. We find a clear rejection of the response homogeneity hypothesis for both of these demographic characteristics. Given similar HUI levels, males and females, and young and old, do *not* report similar levels of SAH. In particular, younger and female respondents seem to rate their health levels lower than older and male respondents with a similar HUI. This is in line with the findings of Van Doorslaer & Gerdtham (2003) for Swedish adults.

<include tables 1a-5b around here>

5. Conclusions and discussion

Differential health reporting by subgroups of the population presents a potentially serious problem to the validity of comparisons between -- and the measurement of inequalities across -- subgroups. Any tests for such differential reporting inevitably have to condition on some other, preferably more objective, measure of health. For many purposes, the collection of additional data using vignette-type questions as proposed by WHO (...) will prove very data demanding and not feasible for general-purpose surveys. In such cases, one will have to rely on the use of other, more objective, health measures collected in the survey.

In this paper we show how differential reporting by certain characteristics can be tested for in the context of an ordered response model. In addition, we show how cut-point shift and index shift can be tested separately using simple likelihood ratio tests. We discuss how our approach relates to the vignette approach and conclude that our testing procedure to distinguish between cut-point shift and index shift is also of use for the vignette method. We also argue that the responses to the vignettes will most likely depend on the health status of the individual respondent and that one therefore, as in our method, needs to condition on other objective health measures.

We illustrate our model and testing procedure by testing whether in Canada the relationship between the ordered responses to the self-assessed health (SAH) question is prone to index or cut-point shift. For language, income and education we find very few violations of the homogeneous response hypothesis, and in the few cases where it is violated, it appears almost invariably due to index shift rather than cut-point shift. The results are very different for age and gender: males and females respond differently to the SAH question, and so do the young and the older respondents. Female and older respondents appear to be milder in their self-assessments than their male and younger counterparts. Moreover, the differences are reflected both in index and in cut-point shifts. As a result, any differences between these groups in SAH tend to understate the ‘true’ differences insofar as these are mirrored by the differences in the Health Utility Index.

These Canadian results are of course sample specific and hold most likely at the most for the adult Canadian population. Future research should involve other countries to see if these results persist across countries. The tools provided in this paper are very powerful for the measurement and the decomposition of the inequality in health. In the case of index shifts estimates of health inequalities or differences between socioeconomic groups do not suffer from reporting heterogeneity as long as demographic differences are accounted for (for instance via standardisation on the relevant variables). In the case of cut-point shift this standardisation is less straightforward because this relates to the SAH measure in a non-linear way and therefore one has to rely on other ways to control for reporting heterogeneity. One way to do this is to exploit the structure of the reporting model. Model estimates can be used to generate an estimate of the ‘true’ health (H^*) via equation (2). In this measure variations in cut-points are eliminated, as by construction of the model these are in the response category thresholds (c_i). Next, one could use the index H^* to construct income related health inequalities and the decomposition of these inequalities. An illustration of this would go beyond the scope of the current paper and we therefore leave this for future research.

Tables

Table 1 Test for differential response by language group

	Males		Young				Old			
			Education high		Education low		Education high		Education low	
	Income high	Income low	Income high	Income low	Income high	Income low	Income high	Income low		
log lik French speaking	-33.71	-31.89	-61.68	-123.41	-73.49	-81.88	-40.00	-108.48		
log lik English speaking	-600.24	-277.09	-449.53	-490.61	-942.05	-602.32	-454.03	-559.29		
log lik Both languages	-142.05	-60.33	-84.20	-104.04	-286.11	-145.22	-93.15	-85.41		
log lik Other language	-175.68	-115.25	-97.86	-138.11	-235.92	-171.42	-78.20	-110.21		
Log lik sum (unrestricted model, Λ^U)	-951.69	-484.56	-693.28	-856.17	-1537.58	-1000.84	-665.38	-863.38		
Log likelihood restricted model (Λ^R)	-957.34	-495.72	-696.79	-861.65	-1548.93	-1012.02	-672.32	-870.89		
χ^2 test for overall shift $-2*(\Lambda^R-\Lambda^U)$	11.30	22.31	7.02	10.96	22.71	22.37	13.88	15.02		
P-value (12 degrees of freedom)	0.50	0.03	0.86	0.53	0.03	0.03	0.31	0.24		
Log lik semi-restr model (cut-pnts free, Λ^{CP})	n.a.	-490.80	n.a.	n.a.	-1546.89	-1008.52	n.a.	n.a.		
χ^2 test for cut-point shift $-2*(\Lambda^R-\Lambda^{CP})$	n.a.	9.82	n.a.	n.a.	4.08	7.01	n.a.	n.a.		
P-value (9 d.f.)	n.a.	0.37	n.a.	n.a.	0.91	0.636	n.a.	n.a.		
χ^2 test for index shift $-2*(\Lambda^{CP}-\Lambda^U)$	n.a.	12.49	n.a.	n.a.	18.63	15.36	n.a.	n.a.		
P-value (6 d.f.)	n.a.	0.05	n.a.	n.a.	0.00	0.02	n.a.	n.a.		
	Females		Young				Old			
			Education high		Education low		Education high		Education low	
	Income high	Income low	Income high	Income low	Income high	Income low	Income high	Income low		
log lik French speaking	-63.65	-40.52	-59.65	-206.82	-112.21	-114.42	-53.00	-179.85		
log lik English speaking	-704.39	-386.92	-446.10	-755.92	-1016.34	-961.91	-351.37	-748.52		
log lik Both languages	-150.87	-64.94	-69.15	-162.07	-362.63	-252.63	-66.11	-135.06		
log lik Other language	-132.94	-110.11	-113.91	-201.85	-214.36	-229.30	-40.83	-140.71		
Log lik sum (unrestricted model, Λ^U)	-1051.85	-602.49	-688.80	-1326.65	-1705.55	-1558.26	-511.31	-1204.14		
Log likelihood restricted model (Λ^R)	-1056.05	-607.99	-695.15	-1334.47	-1709.37	-1566.56	-519.20	-1217.68		
χ^2 test for overall shift $-2*(\Lambda^R-\Lambda^U)$	8.40	10.99	12.70	15.64	7.66	16.62	15.78	27.07		
P-value	0.75	0.53	0.39	0.21	0.81	0.16	0.20	0.01		
Log lik semi-restr model (Cut-pnts free, Λ^{CP})	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	-1206.17		
χ^2 test for cut-point shift $-2*(\Lambda^R-\Lambda^{CP})$	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	23.01		
P-value (9 d.f.)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.006		
χ^2 test for index shift $-2*(\Lambda^{CP}-\Lambda^U)$	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	4.06		
P-value (6 d.f.)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.67		

Table 2 Test for differential response by income

Grouped over language (but controlled for in index)	Males				Females			
	Young		Old		Young		Old	
	Educ high	Educ low	Educ high	Educ low	Educ high	Educ low	Educ high	Educ low
log lik High income earners	-957,34	-696,79	-1548,93	-672,32	-1056,05	-695,15	-1709,37	-519,20
log lik Low income earners	-495,72	-861,65	-1012,02	-870,89	-607,99	-1334,47	-1566,56	-1217,68
Log lik sum (unrestricted model, Λ^U)	-1453,05	-1558,44	-2560,95	-1543,21	-1664,04	-2029,62	-3275,94	-1736,88
Log likelihood restricted model (Λ^R)	-1459,44	-1563,34	-2570,69	-1544,78	-1668,38	-2035,27	-3278,79	-1738,98
$-2*(\Lambda^R-\Lambda^U)$	12,78	9,80	19,47	3,13	8,68	11,29	5,70	4,21
P-value (7 degrees of freedom)	0,08	0,20	0,01	0,87	0,28	0,13	0,58	0,76
Log lik semi-restricted model (Cut-pnts free, Λ^{CP})	n.a.	n.a.	-2570,60	n.a.	n.a.	n.a.	n.a.	n.a.
χ^2 test for cut-point shift $-2*(\Lambda^R-\Lambda^{CP})$	n.a.	n.a.	0,18	n.a.	n.a.	n.a.	n.a.	n.a.
P-value (3 d.f.)	n.a.	n.a.	0,98	n.a.	n.a.	n.a.	n.a.	n.a.
χ^2 test for index shift $-2*(\Lambda^{CP}-\Lambda^U)$	n.a.	n.a.	19,29	n.a.	n.a.	n.a.	n.a.	n.a.
P-value (5 d.f.)	n.a.	n.a.	0,00	n.a.	n.a.	n.a.	n.a.	n.a.

Table 3 Test for differential response by education

Grouped over language, income (but controlled for in index)	Males		Females	
	Young	Old	Young	Old
	log lik High educated	-1459,44	-2570,69	-1668,38
log lik Low educated	-1563,34	-1544,78	-2035,27	-1738,98
Log lik sum (unrestricted model, Λ^U)	-3022,78	-4115,46	-3703,64	-5017,77
Log likelihood restricted model (Λ^R)	-3025,74	-4120,15	-3709,71	-5023,48
$-2*(\Lambda^R-\Lambda^U)$	5,93	9,37	12,14	11,42
P-value (8 degrees of freedom)	0,65	0,31	0,15	0,18

Table 4 Test for differential response by gender

	Young	Old
log lik Males	-4120,15	-3025,74
log lik Females	-5023,48	-3709,71
Log lik sum (unrestricted model, Λ^U)	-9143,63	-6735,45
Log likelihood restricted model (Λ^R)	-9159,17	-6744,06
$-2*(\Lambda^R - \Lambda^U)$	31,09	17,21
P-value (9 degrees of freedom)	0,00	0,05
Log lik semi-restricted model (Cut-pnts free, Λ^{CP})	-9154,60	-6743,49
χ^2 test for cut-point shift $-2*(\Lambda^R - \Lambda^{CP})$	9,15	1,13
P-value (3 d.f.)	0,03	0,77
χ^2 test for index shift $-2*(\Lambda^{CP} - \Lambda^U)$	21,94	16,08
P-value (7 d.f.)	0,00	0,02

Table 5 Test for differential response by age

	Males	Females
log lik Old group	-3025,74	-3709,71
log lik Young group	-4120,15	-5023,48
Log lik sum (unrestricted model, Λ^U)	-7145,89	-8733,19
Log likelihood restricted model (Λ^R)	-7185,18	-8748,75
$-2*(\Lambda^R - \Lambda^U)$	78,57	31,12
P-value (9 degrees of freedom)	0,00	0,00
Log lik semi-restricted model (Cut-pnts free, Λ^{CP})	-7166,00	-8740,54
χ^2 test for cut-point shift $-2*(\Lambda^R - \Lambda^{CP})$	38,35	16,42
P-value (3 d.f.)	0,00	0,00
χ^2 test for index shift $-2*(\Lambda^{CP} - \Lambda^U)$	40,22	14,70
P-value (7 d.f.)	0,00	0,04

References

- Cutler, D., Richardson, E., 1997. Measuring the health of the United States Population. *Brookings Papers on Economic Activity, Microeconomics*, 217-271.
- Feeny, D., Furlong, W., Boyle, M., Torrance, G., 1995. Multi-attribute health status classification systems: Health Utilities Index. *Pharmacoeconomics* 7 (6), 490-502.
- Furlong, W., Feeny, D., Torrance, G.W., Goldsmith, C., DePauw, S., Zhu, Z., Denton, M., and Boyle, M., 1998. *Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report*, McMaster University, Centre for Health Economics and Policy Analysis Working Paper No. 98-11.
- Gravelle H. (2010), 'Measuring Income Related Inequality in Health and health Care: the partial Concentration Index with Direct and Indirect Standardisation, Working Paper university of York.
- Groot, W., 2000. Adaptation and scale of reference bias in self-assessments of quality of life. *Journal of Health Economics* 19(3), 403-420.
- Hays, RD, LS Morales and SP Reise (2000). Item response theory and health outcomes measurement in the 21st Century, *Medical Care*, 38 (9), Supplem II, II28- II42
- Humphries, K and E van Doorslaer, 2000, Income-related inequalities in health in Canada, *Social Science and Medicine*, 50, 663-671
- Jones, A.M., 2000. Health Econometrics, in A.J.Culyer and J.P.Newhouse (eds.) *Handbook of Health Economics*. Elsevier, Amsterdam.
- Kerkhofs, M., Lindeboom, M., 1995, Subjective health measures and state dependent reporting errors. *Health Economics* 4, 221-235.
- Murray, CJL, A Tandon, J Salomon, CD Mathers and R Sadana, 2001, *Cross-population comparability of evidence for health policy*, GPE Discussion Paper Nr 46, WHO/EIP, Geneva.
- Sadana, R, CD Mathers, AD Lopez, CJL Murray and K Iburg, 2000, *Comparative analysis of more than 50 household surveys on health status*, GPE Discussion Paper No 15, EIP/GPE/EBD, World Health organisation, Geneva.
- Sen, A, 2002, Health: perception versus observation, *British Medical Journal*, 324, 660-1
- Shmueli, A, 2002, Reporting heterogeneity in the measurement of health and health-related quality of life, *Pharmacoeconomics*, 20 (6), 405-412
- Tambay J-L., Catlin, G., 1995. Sample Design of the National Population Health Survey. *Health Reports* 7(1), 29-38.
- Tandon, A, CJL Murray, JA Salomon and G King, 2002, *Statistical models for enhancing cross-population comparability*, Global Programme for Evidence on Health Policy Discussion Paper Nr. 42, WHO, Geneva.
- Torrance, G.W., Furlong, W., Feeny, D., Boyle, M., 1995. Multi-attribute preference functions: Health Utilities Index. *Pharmacoeconomics* 7(6), 503-520.
- Torrance, G.W., Feeny, D., Furlong W.J., Barr, R.D., Zhang, Y., Wang, Q., 1996. Multiattribute Utility Function for a Comprehensive Health Status Classification System. *Medical Care* 34(7), 702-722.
- Van Doorslaer, E., Wagstaff, A., Bleichrodt, H., *et al*, 1997. Income-related inequalities in health: some international comparisons. *Journal of Health Economics* 16, 93-112.
- Van Doorslaer, E and AM Jones, 2003, Inequalities in self-reported health: validation of a new approach to measurement, *Journal of Health Economics*, 22, 61-87

Van Doorslaer, E. & U. Gerdtham (2003), 'Does inequality in self-assessed health predict inequality in survival by income? Evidence from Swedish data', forthcoming in *Social Science & medicine*.

Wagstaff, A., van Doorslaer, E., 1994. Measuring inequalities in health in the presence of multiple-category morbidity indicators. *Health Economics* 3, 281-291.

Appendix

Table A1: Descriptive statistics

Variables	Mean	Std. Dev
Health Utility Index	0.893	0.130
SAH excellent	0.237	0.425
SAH very good	0.384	0.486
SAH good	0.265	0.441
SAH fair	0.091	0.287
SAH poor	0.024	0.152
speaks French only	0.088	0.283
speaks both English and French	0.147	0.354
speaks other language	0.147	0.354
Female	0.550	0.498
Low education	0.532	0.499
Young (<35y)	0.486	0.500
Low income (<Can\$ 24748 per eq adult)	0.497	0.500

Table A2: Mean HUI values by level of SAH and background characteristics

		Self-assessed health				
By:		Poor or fair	Good	Very good	Excellent	Total
sex	male	0.736	0.882	0.928	0.946	0.900
	female	0.715	0.871	0.920	0.944	0.888
age	old	0.714	0.864	0.914	0.934	0.869
	young	0.754	0.893	0.931	0.953	0.918
education	high	0.727	0.881	0.928	0.947	0.909
	low	0.722	0.872	0.919	0.942	0.879
income	high	0.743	0.882	0.926	0.945	0.909
	low	0.716	0.872	0.921	0.945	0.877
language	English	0.720	0.873	0.923	0.946	0.892
	French	0.735	0.882	0.930	0.941	0.893
	both	0.737	0.881	0.925	0.944	0.903
	other	0.722	0.880	0.923	0.945	0.891
Total		0.724	0.876	0.923	0.945	0.893
N		1567	3627	5265	3240	13699