DOES REPORTING HETEROGENEITY BIAS THE MEASUREMENT OF HEALTH?

by Shixuan Jian

(7888100)

Major Paper presented to the

Department of Economics of the University of Ottawa

in partial fulfillment of the requirements of the M.A. Degree

Supervisor: Professor Myra Yazbeck

Professor Paul Makdissi

ECO 6999

Ottawa, Ontario

August 2016

CATALOGUE

INTRODUCTION

Measuring health is key. Any researcher who wants to provide reliable evaluations or policy recommendations should use sound health measurements. In applied health economics, researchers use various measurement methods to assess true level of health. Nevertheless, health status is multidimensional and complex, which pose a real challenge for a health measure to capture the true health situation in all dimensions.

Generally speaking, health status measurements consist of four categories: subjective health measures, generic health measures, vignette-based health measures and objective health measures *(Ziebarth, 2010).* Based on the diagnosis by physicians, such as weight and height, objective health measures seem to have the least probability of suffering from measurement errors and reporting biases. Though objective health measures proxy the actual level of health most closely, they are expensive to collect. Besides, they fail to capture the multidimensional nature of health *(Molina, 2016)*, and may transmit misleading information to researchers and increase the risk of biased conclusions. For example, if an individual has a severe problem walking, it is not wise to translate this into bad health status, the same person may have excellent performance in other dimensions.

Subjective health measures refer to measures that are entirely based on self-reported responses, such as health satisfaction and self-assessed health (SAH) *(Ziebarth, 2010)*. Compared to the objective health measures, subjective health measures are cheap, easy to collect, equally predictive and able to cover multiple health dimensions. Self-reported health is a frequently used subjective health measure to analyse particular concerns about health, such as health inequality. Unlike objective health measures, self-reported health measures consist of numerous questions like "How much pain did you feel in last week", it measures perceived health status in categorical responses. Earlier studies

widely employ self-reported health to analyse health (*Bago d'Uva et al., 2008; Dowd and Todd, 2011; Hirve et al., 2014*).

Nevertheless, self-reported health seems to be inevitably suffering from reporting heterogeneity as the assessment of health is entirely subjective *(Bago d'Uva et al., 2008),* which implies respondents may report different health levels given the same latent health status. Reporting heterogeneity of self-reported health refers to the variation that relates to individual characteristics in a systematic way. If the variations of health assessments are not systematic, it is usually taken to be random *(Shmueli, 2003; Bago D'Uva et al., 2008),* which implies the variation need not be the concern because measurements of health would still be valid. However, if there are systematic differences in reporting behaviours and we still treat them as random errors, the conclusions may be invalid.

To address the issue brought by the measurement error of self-reported health, researchers usually choose one of the two approaches. One approach is conditioning on a kind of generic health measures which are single indexes that are aggregated by responses of self-reported health questions from different health domains. Generic health measures are quasi-objective and comprehensive, Canadian HUI-3, Germany SOEP and Finnish 15D are examples of these measures. An alternative is using anchoring vignettes to calibrate self-reported health. Vignettes are hypothetical descriptions of fictitious individuals; they are exogenous. However, the use of vignettes relies on two fundamental assumptions. With the assumption of *response consistency* and another assumption of *vignette equivalence*, researchers could calibrate self-reported health by using the responses of vignettes. *Response consistency* requires respondents answer the vignettes in the same way as they rate their health, and *vignette equivalence* defines all respondents evaluate each vignette in each domain as the same fixed health level in the same unidimensional scale regardless of random error *(Hirve et al., 2014)*.

Since generic health measures are based on subjective health status, they are presumably objective and easy to collect using general surveys. However, this approach may introduce noise because generic health measures generated by self-reported health questions (Baker et al., 2004). *Shmuli (2003)* reports different reporting heterogeneity properties by using different health measures; SF36, a generic health measure, is sensitive to age, sex, economic status, ethnic origin and religiosity, while the objective health measure, the number of chronic conditions, only shows significant age-related reporting difference. Besides, if self-reported health is intimately related to real health and it contains information on socio-economic related variation, conditioning on generic health measures would eliminate the effect of socio-economic related variation *(Bago d'Uva et al., 2008).*

Because of the disadvantages of conditioning on generic health measures, employing anchoring vignettes is becoming popular among researchers. *Bago d'Uva et al., (2008)* propose a hierarchical ordered probit (HOPIT) model in their paper 'Does reporting heterogeneity bias the measurement of health disparities?' to identify reporting behaviour and to purge reporting bias in three developing countries: Indonesia, India, and China. This paper finds significantly and consistently regional, income-related, and age-related reporting heterogeneity across all countries, but less consistently by sex and education.

This paper follows *Bago d'Uva et al., (2008)*, but examine developed countries: the United States and Canada. Compared to developing countries, developed countries have higher education levels and income level. Besides, developed countries and developing countries have very different age structure. Previous works have shown that reporting behaviours by socioeconomic variables vary with contexts. Hence, the findings of *Bago d'Uva et al., (2008)* may not be consistent with developed countries. Also, Canada has a universal health care systems, but the U.S. has a private system, which may enable us to explore different reporting behaviour patterns across these countries. This paper uses American and Canadian data from *WHO Multi-Country Survey Study on Health and Responsiveness 2000-2001 (WHO-MCS)* to identify whether reporting behaviour is

heterogeneous across different socio-economic characteristics (income, education, gender and age) in the domains of *mobility*, *cognition*, *pain*, *self-care*, *usual activities*, and *affect*. If so, estimating the magnitude of reporting heterogeneity and then purging the bias from measures of health disparities. Following *Bago d'Uva et al., (2008)*, our concern is the comparability of self-reported health across socioeconomic and demographic groups within a country and the true health effects by socioeconomic groups. To our knowledge, this is the first paper identify and compare health reporting behaviours in the U.S. and Canada using anchoring vignettes. Subsequent sections are the literature review, econometric model, data, results, and conclusion.

LITERATURE REVIEW

As mentioned earlier, self-reported health may be prone to measurement error. The variation in the responses of self-reported health should not be our concern if measurement error is not systematic. However, substantial evidence show that measurement error of self-reported health is rarely non-random and relates to socioeconomic and demographic variables. For example, *Currie and Madrian (1999)* indicate the main problem with self-reported health is that it is easily influenced by education, income, employment, and health insurance status. Nevertheless, previous work finds contradicting conclusions about reporting heterogeneity by socioeconomic and demographic variables under different contexts. What follows demonstrates different findings by two different identifying methods: conditioning on generic health measures and anchoring vignette, and shows why using anchoring vignettes are better than conditioning on generic health measures.

Use objective health indicators

Even though anchoring vignettes provide an alternative to identifying heterogeneous reporting behaviour, data collecting would be demanding. Objective and generic health measures are sometimes easier to apply compared to anchoring vignettes, for researchers who want to test reporting heterogeneity, this strategy uses quasi-objective health indicators, such as self-reported clinical health *(Etilé and Milcent, 2006)*, or the synthetic health index based on self-reported health, such as *MacMaster Health Utility Index (Lindeboom and van Doorslaer, 2004)*. However, results vary across countries. *Van Doorslaer and Gertham (2003)* and *Lindeboom and van Doorslaer (2004)* both report reporting heterogeneity by age and gender, but not by income and education. Nevertheless, *Kerkhofs and Lindeboom (1995)* find substantial evidence of education-related reporting heterogeneity, *Etilé and Milcent (2006)* report significant income-related reporting heterogeneity in France.

Van Doorslaer and Gerdtham (2003) use Swedish data to figure out whether self-reported health would affect individuals' subsequent survival probability, besides they also examine whether this effect differed between different socio-economic groups. They employ self-assessed health (SAH) as the subjective health measure and choose survival probability as the objective health measure to be conditioned on. The result shows that SAH has a significantly predictive power of subsequent mortality risk, female, younger, and have lower blood pressure decrease mortality risk. The relationship between SAH and mortality risk varies across demographic characteristics such as age, sex, and hypertensive status but not by socioeconomic characteristics such as income, education or functional limitation status. To be specific, respondents who are males, older, and have higher blood pressures incline to report better health compared to females, younger and have lower blood pressures. Nevertheless, *Doorslaer and Gerdtham (2003)* only demonstrate how the predictive power differs by different variables, which implies the different effect of SAH on mortality risk could be heterogeneous reporting behaviour or something else. In another word, the sources of variation in the effect of self-reported health on mortality risk were vague; this poses a threat to the significance of reporting heterogeneity.

Using a different health measure (HUI-3), the findings of *Lindeboom and van Doorslaer (2004)* is partially consistent with that of *Doorslaer and Gerdtham (2003)*, where reporting heterogeneity is significant for age and sex but not for income and education. *Lindeboom and van Doorslaer (2004)* propose to condition on a generic health measure to test reporting homogeneity. Unlike *Doorslaer and Gerdtham (2003)* using an objective health measure, *Lindeboom and van Doorslaer (2004)* choose a generic health measurement: *McMaster Health Utility Index Mark 3(HUI-3)* to proxy objective health measurement. The results accept the null hypothesis of homogeneity for language, income, and education. Nevertheless, for age and gender, reporting heterogeneity is significant: female and older respondents incline to report better health compared to male and younger respondents given the same objective health status, this is partially consistent with the results of *Van Doorslaer and Gerdtham (2003)* which indicates male and older respondents are more

likely to report better health. Though the results indicate significant heterogeneous reporting behaviour by age and gender, it is possible that *HUI-3* is the source of variation (*Lindeboom and van Doorslaer, 2004*). Because *HUI-3* is generated by responses to self-reported health, if a certain level of health domain differs by age or gender, the results would indicate significant reporting heterogeneity in this domain. Besides, unlike *Doorslaer and Gerdtham (2003)*, *Lindeboom and van Doorslaer (2004)* only test for reporting heterogeneity of self-reported health without indicating health effect by demographic groups.

Though *Lindeboom and van Doorslaer (2004)* and *Doorslaer and Gerdtham (2003)* both reject reporting heterogeneity by income, *Etilé and Milcent (2006)* report income-related reporting heterogeneity but reporting homogeneity by age, sex, and education of self-assessed general health (SAH) in France. They use clinical health as the objective health measure, and they test income-related reporting behaviour in two steps: the first method identifies whether reporting heterogeneity exists, and the second one assesses the magnitude of reporting heterogeneity based on a classification of respondents. The results show that socio-economic variables have a homogeneous effect on the cut-off points, but region and income had a heterogeneous effect on the cut-points where those with higher income and live in Western France are more likely to report better health. As for the health effects, SAH and income are positively correlated, gender seems to have no effect on health but divorced respondents prone to report better health status. Compared to *Lindeboom and Van Doorslaer (2004)*, the results are quite different where *Etilé and Milcent (2006)* identifies income-related reporting heterogeneity, but *Lindeboom and Van Doorslaer (2004)* suggest no income-related reporting heterogeneity. Since both France and Canada have a universal health care system, the reason to explain the contradicting results may be different countries. Besides, this paper indicates how SAH is affected by different socioeconomic and demographic groups. However, the results of this paper strongly rely on the assumption of putatively objective health indicators. Similar to *Doorslaer and Gerdtham (2003),* this paper also employs SAH as the self-reported

health measure. Though the SAH is naturally multidimensional, if it subjects to errors, it may also introduce errors into the results.

Kerkhofs and Lindeboom (1995) finds significant reporting heterogeneity for education. They propose the ordered probit model to estimate the impact of labour state dependent reporting errors by comparing the subjective health measure with an objective measure. The objective health measure use *Hopkins Symptom Checklist (HSCL)* to construct and the subjective health measure is a work-related measure which is controlled by different labour market states (employed, unemployed, early retired and disabled). The model allows the thresholds depend on labour market states and other exogenous variables: gender, age, education, marital status, and region. The results show that most other exogenous variables seem to have weak evidence of contributing to misreporting except education. More educated respondents prefer to report a lower health level. This paper also mentions if the objective health measure is subject to the same type of systematic reporting errors, then the magnitude of reporting errors would be larger than that of the actual reporting errors. Since this paper uses Dutch data, the evidence of education-related reporting bias may not be significant in other countries.

Four articles mentioned above test reporting heterogeneity conditioning on objective health measures, all confirm reporting heterogeneity by exogenous demographic or socioeconomic variables. However, the results are mixed. Two reasons may be able to explain why researchers report mixed results relating the same socio-economic variables: one is different countries, another one is different objective health measures. It is not hard to understand that respondents from different countries show various patterns of reporting behaviour because individuals are largely affected by the cultures. However, the results may also be dependent on different generic health measures. *Ziebarth (2010)* finds the magnitude of the concentration index varies between the various generic and object health measures, which lends support to the idea that results may vary according to generic health measures and so make the results not reliable.

The crucial problem of conditioning on objective or quasi-objective health measures is the use of generic health measures which are synthetic indexes of multi-dimensional self-reported health. If self-reported health suffers from reporting heterogeneity, so would generic health measure. The risk of reporting heterogeneity makes generic health measures no longer the valid bias-free and objective health measures to become the fundament of this approach *(Etilé and Milcent, 2006)*. Another issue is that this approach may disguise socio-economic related variation in self-reported health. If self-reported health contains information of true health, conditioning on quasi-objective health measure will cause a lost on this part of the information *(Lindeboom and van Doorslaer, 2004)*. Given these two disadvantages above, an alternative using extra information provided by vignettes offers another approach to be explored. *Lindeboom and van Doorslaer (2004)* and *Etilé and Milcent (2006)* both mention anchoring vignettes as an alternative to address the potential risk that is brought by generic health measures. Anchoring vignettes may be more data-demanding and expensive, but they avoid identifying reporting behaviour from variation in self-reported health that cannot be explained by quasi-objective health measures and the high risk of measurement errors.

Use anchoring vignettes

An alternative separating reporting behaviour from true health disparities is using anchoring vignettes. Vignettes are a set of health descriptions of fictitious individuals, they are exogenous and represents given health level. In this approach, respondents rate the vignettes along with own health into categories which normally range from 'severe' to 'none'. Thresholds of health categories depend on individual characteristics, such as age, gender, education, and income. The evaluation of vignettes is used to estimate thresholds. The estimated coefficients reflect direction and magnitude of shifts on thresholds. There are two types of variations: index shift and cut-points shift; the former refers to a parallel shift of the thresholds, which implies the reporting behaviour affect the thresholds by the same magnitude, the latter relates to the situation where reporting behaviour affects cut-off points differently *(Lindeboom and van Doorslaer, 2004)*. The estimated cut-off points are then used to calculate health equations and adjust the assessments of respondents' own health.

The application of anchoring vignettes requires two fundamental assumptions: vignette equivalence and response consistency. Vignette equivalence requires all respondents interpret a given vignette in the same fixed health level apart from random errors *(Hirve et al., 2014)*, in another word, vignettes are independent of respondents' individual characteristics *(Molina, 2016)*. This assumption ensures the respondents differ only in the thresholds other than their interpretation of hypothetical scenarios. If respondents relate individual characteristics to the understanding of vignettes, we cannot take the differences in self-reported health as the impact of random errors, so we cannot identify reporting behaviour. Response consistency requires individuals rate vignettes in the same scales as they assess their own health, which implies the ratings of vignettes and own health share the same thresholds *(Hirve et al., 2014)*. This assumption will be violated if respondents have different standards for themselves and others, then the cut-off points used to rate vignettes and self-reported health are different, and we cannot evaluate the actual health effect.

Earlier studies test the validity of the assumptions but found mixed results. *Bago d'Uva et al. (2011)* reject the validity of both hypotheses. However, *Bago d'Uva et al. (2011)* only test one health domain: mobility, so that the analysis is incomplete and the results may be different if we test other domains. Following the method proposed by *Bago d'Uva et al. (2011)*, *Molina (2016)* verify the validity of vignette equivalence and find the assumption is not always valid. Nevertheless, *Rice et al. (2011)* analyses the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness, the results do not contradict the assumption of vignette equivalence, and it lends supports to the use of vignette methodology. Also, *Hirve et al. (2014)* find vignette equivalence is valid in most domains, but response consistency is not valid. Though there is evidence of invalid assumption of response consistency, authors suggest the assumption of the test used to verify the validity of response consistency is not valid, which implies the result of invalid response consistency is based on an invalid assumption. Besides, revising contents of vignettes may affect the validity of assumptions. *Datta Gupta et al. (2010)* indicate that response consistency is vital for health policy suggestions because relaxing this assumption would cause a significant change in cross-country health rankings, but they do not test response consistency with a formal test. However, *van Soest et al. (2011)* propose a formal test about the premise of response consistency, so does *King et al. (2004)*, they check the validity of response consistency, both results confirm the validity of response consistency.

Summing up the results of earlier tests on the validity of assumptions, we tend to take these two assumptions as valid. Some researchers, such as *Bago d'Uva et al., (2008)* and *Dowd and Todd (2011),* do not test two assumptions before starting their analysis. However, some researchers choose to test assumptions before applying vignettes, for example, *Hirve et al., (2014)* and *Molina (2016)* both test the validity of two assumptions before applying vignettes. Even though they both find evidence of invalid assumptions, *Molina (2016)* adjust the model to allow the violation of assumption to ensure the validity of the conclusions. In contrast to *Molina (2016)*, *Hirve et al., (2014)* suggest the invalidity

of the test of response consistency and do nothing with the approach that follows *Tandon et al. (2003)* and *Bago d'Uva et al. (2011).*

Bago d'Uva et al. (2008) employ anchoring vignette to identify reporting heterogeneity of self-reported health in three developing countries: Indonesia, India, and China. The baseline model is an ordered probit model; the one-step hierarchical ordered probit (HOPIT) model is used to identify true health effect and purge reporting bias from health effect. Unlike *Hirve et al. (2014)*, they take two assumptions underlying HOPIT model as indisputable. The results reject the null hypothesis of parallel cut-point shift which implied covariates affect all cut-points by different magnitude. Besides, the results reject the null hypothesis of reporting homogeneity by any socio-economic variable (age, sex, region, education, and income) in most cases, and the evidence of reporting heterogeneity is stronger by income. In general, younger, male (except Indonesia) and urban respondents with higher education level (excluding China) tend to report better health in all three countries. After correcting reporting heterogeneity, health disparities by education slightly reduces in Indonesia and India but health disparities by income increase in all three countries. Nevertheless, this paper proposed that the underlying assumptions may not be valid for low-income and low-educated countries, which implies that vignette method may be more feasible for countries with higher income and education levels. This is why we are looking at high-income countries. Unlike previously mentioned papers, this paper employs hierarchical ordered probit to identify and purge reporting heterogeneity, and it also measured the effect of health inequality in developing countries where greater differences exist in the understanding of health compared to developed countries.

Hirve et al. (2014) employ HOPIT model to evaluate reporting heterogeneity of self-reported health among more aged people in India by using the data comes from the *WHO Study on global AGEing and adult health (SAGE)*. They test the validity of two assumptions and find both assumptions valid. Four health domains including mobility, vigorous activity, cognition and learning while socio-economic variables include gender, age, socio-economic status and education level. The results reject reporting homogeneity

of self-reported health by socioeconomic status and education in mobility and cognition: younger, male with higher socioeconomic status incline to report better ability moving around, while respondents with higher socioeconomic status and higher education level are more likely to rate lower levels of cognition. Nevertheless, reporting heterogeneity is not significant in the domain of vigorous activity and learning though the pattern is similar to that of mobility. After adjustment, female, older respondents with lower socioeconomic status are less likely to report better health. Unlike *Bago d'Uva et al. (2008)* and *Dowd and Todd (2011)*, this paper does not test whether reporting behaviour is a parallel shift. However, this article examines the validity of two assumptions which is required by HOPIT model, and the results do not discredit the overall use of vignette approach. Moreover, this article finds proof that reporting behaviours may not be uniform across different domains within a country.

Molina (2016) finds reporting heterogeneity by gender and education in Indonesia, the United States, England and China. Unlike the same HOPIT model proposed by *Bago d'Uva et al. (2008)*, thresholds vary across individuals but employ a nonnegative exponential function instead of linear function. This difference is due to the finding that the assumption of vignette equivalence is not always valid, this adjustment allows for the violation of assumption so to ensure the validity of conclusions. This study uses six health domains (mobility, pain, cognition, affect, sleep and breathing), three socio-demographic variables (age, gender, education), and three health vignettes. The data of the U.S. and England respectively comes from *the U.S. Health and Retirement Study (HRS)* and *the English Longitudinal Study of Aging (ELSA)*. The results show significant reporting heterogeneity by education in all four countries, where more educated respondents tend to report worse health categories. Reporting heterogeneity across gender is important, males are more likely to report better health than females. After adjustment, the education disparities increase after correction, which implies ignoring reporting behaviour would underestimate differences in health. However, gender gaps are narrowed after adjustment but with varying degrees across countries. The adjustment eliminates significant gender differences in almost all domains of England and China, but it only

narrows gender gaps in half domains of Indonesia and the U.S. The paper uses two developing countries and two developed countries to contrast, which is the first attempt. Besides, the results are interesting because they report similar reporting behaviour by gender of a developing country and a developed country. Similar to *Hirve et al., (2014)* and *Dowd and Todd (2011)*, this article also explores reporting behaviours of aged people. Especially, the data used in this paper is same as that used in *Dowd and Todd (2011),* so do the conclusions about reporting heterogeneity by education and gender. Nevertheless, as the author mentions, the results may not be consistent with younger respondents because this article examines aged populations.

Dowd and Todd (2011) use anchoring vignettes to test and adjust reporting differences by age, gender, race, and education in the U.S. for six health domains (pain, sleep, mobility, memory, shortness of breath and depression). Similar to *Hirve et al. (2014)* and *Molina (2016)*, they concern on the aged population. Consistent with *Molina (2016)*, the data comes from *the U.S. Health and Retirement Study (HRS).* Following *Bago d'Uva et al. (2008)*, this paper uses both generalised ordered probit model and HOPIT model. The results reject the null hypothesis of the parallel shift in all domains when all covariates are jointly considered and are separately considered. Also, the results reject the null hypothesis of homogeneous reporting in every health domain considering the joint socio-economic status and decline homogenous reporting in most domains by gender, race, and education. In general, both the Black and Hispanics are optimistic about relatively minor health problems but are pessimistic at higher levels of severity. Males tend to report better health in most domains except for memory and pain, less-educated and older respondents incline to rate better health. Besides, the magnitude of the heterogeneity varies from different health domains. Unlike the results of *Bago d'Uva et al., (2008)* or *Molina (2016)*, the adjustment has a mixed impact on the magnitude of the estimated coefficients for a certain covariate across health domains. Compared to *Bago d'Uva et al. (2008)*, this paper does not include income and region as covariates but focus on race and education. In developing countries, the difference of understanding about health by region and income may be greater, so when *Dowd and Todd (2011)* test reporting

heterogeneity in the U.S., region and income are not as important as that of developing countries. Compared to *Hirve et al. (2014)*, the results are somewhat consistent where reporting heterogeneity by age, gender, and education is significant. As I mentioned earlier, this paper concerns on the aged population, which implies the conclusions may not be consistent with other age groups in the U.S.

Four articles mentioned above all employ HOPIT model and anchoring vignettes to analyse within country differences by demographic and socioeconomic variables. The results support significant reporting heterogeneity by different socioeconomic and demographic variables, but by various direction and magnitude. Besides, only *Bago d'Uva et al. (2008)* and *Dowd and Todd (2011)* test whether the reporting behaviour affect thresholds by the same magnitude, both results reject parallel shift which implies reporting behaviour affect cut-off points differently. Nevertheless, *Lindeboom and van Doorslaer (2004)* reject reporting homogeneity by language, income, and education but find parallel shifts other than cut-points shift. In the case of index shift, *Lindeboom and van Doorslaer (2004)* find it is impossible to distinguish reporting behaviour from true health effects without the help of external information because they employ the approach by conditioning on objective health measures, which is a fundamental identification problem. By using anchoring vignettes, we can separate health effect from reporting behaviour.

Summing up results of empirical studies, reporting heterogeneity is confirmed existed in self-reported health. Results of papers mentioned above suggest socioeconomic and demographic characteristics have the evident effect on misreporting behaviour, though different researchers report different related individual characteristics. *van Doorslaer and Gerdtham (2003)* and *Lindeboom and Doorslaer (2004)* both report age-related and sex-related misreporting but not income-related or education-related misreporting. Nevertheless, the results of *Etilé and Milcent (2006)* and *Bago d'Uva et al. (2008)* both support reporting heterogeneity by income. Besides, education-related or gender-related misreporting is confirmed by most works (*Bage d'Uva et al., 2008; Peracchi & Rossetti, 2012; Hirve et al., 2014; Dowd & Todd, 2011; Molina, 2016*). Nevertheless, no one has

ever compared reporting behaviours of different developed countries before. In this paper, we examine reporting heterogeneity by age, gender, education, and income for the U.S. and Canada. Following *Bago d'Uva et al. (2008)*, firstly estimate an ordered probit model as the baseline model, then use HOPIT model to identify reporting behaviour and true health effect and compare the reporting behaviours and health effects across countries.

ECONOMETRIC MODEL

Since self-reported data consists of individual categorical responses which base on the health-related questions, it is viewed as categorical data. Categorical data is discrete and observed. Nevertheless, the true health level is unobserved. What we interested in is how people present the unobserved true health into observed self-reported health category, we take the self-reported health data as the result of a mapping between latent health and response categories (*Bago d'Uva et al., 2008*), then the reporting homogeneity could be translated into constant mapping across respondents. However, respondents might choose different response categories out of different preferences, socio-economic status or genders, regardless of the fact that they share the same latent health level, the reporting heterogeneity implies different mappings.

The most popular model employed by researchers to analyse categorical data is ordered regression models, such as ordered probit or logit *(Jones et al., 2013)*. In this paper, we opt for ordered probit to analyse reporting behaviour because the specification of ordered probit has been generalised to deal with reporting heterogeneity *(Jones et al., 2013)*.

Ordered probit: homogenous reporting behaviour

This paper starts the analysis by estimating a standard ordered probit as a baseline model. Let $h_i, i = 1, \dots, N,$ be a self-assessed health measure for individual i. It is categorical and observed and is generated by an unobserved latent health measure h_i^*, such that:

$$h_i^* = X_i\beta + \varepsilon_i,\ \varepsilon_i|X_i \sim N(0,1) \qquad (1)$$

where X_i is a vector of covariates, the socio-demographic variables that may contribute to reporting heterogeneity. As the latent variable is unobserved and the observed counterpart is categorical, the scale and the location are not identified. This is why the

variance of the error term is normalised to 1, and the constant is set to zero. As for the observed self-assessed health measure, it is generated by mapping to latent health measure in the following way:

$$h_i = k \Leftrightarrow \mu^{k-1} < h_i^* \leq \mu^k \tag{2}$$

where $k = 1, \dots, K$, $\mu^0 < \mu^1 < \cdots < \mu^{K-1} < \mu^K$ and $\mu^0 = -\infty,\ \mu^K = +\infty$.

Maximum likelihood method allows us to estimate parameters β and $\mu^0, \dots, \mu^K$. When estimating an ordered probit, homogeneous reporting is assumed (i.e., constant cut-off points). Given the information in equation (1) and (2), the probabilities of each respondent i choose each health response category k is denoted P_{ik} and can be written as follows:

$$\begin{aligned} P_{ik} = P(h_i = k) &= P(\mu^{k-1} < h_i^* \leq \mu^k) \\ &= \phi(\mu^k - X_i\beta) - \phi(\mu^{k-1} - X_i\beta) \end{aligned} \tag{3}$$

The associated log-likelihood is:

$$lnL = \sum_{i=1}^{N} \sum_{k=1}^{K} h_{ik} ln P_{ik} \tag{4}$$

where $h_{ik} = 1$ if $h_i = k$ and $h_{ik} = 0$ if $h_{ik} \neq k$. Maximise the log-likelihood function to get estimated parameters β along with cut-off points $\mu^0, \dots, \mu^K$.

Homogenous reporting behaviour corresponds to constant mappings between observed self-reported health measure and unobserved latent health measure, with constant cut-off points $\mu^0, \dots, \mu^K$ across respondents, which implies the ordered probit model hinges on the assumption of the parallel shift. Nevertheless, if cut-off points vary across individuals, the assumption of homogeneity would no longer hold and the estimation results for β and $\mu^0, \dots, \mu^K$ would be biased. If so, the estimation results would reflect both health disparities and the reporting heterogeneity (*Dowd and Todd, 2011*). A generalised ordered probit relaxes the assumption of parallel shifts by allowing the cut-off points depend on covariates. By excluding the effect of covariates from the equation

(1), it is possible to identify cut-points and express the cut-points as a function of covariates. Normalising one cut-point to a constant, the other cut-points parameters would reflect how the covariates shift the thresholds compared to the baseline threshold. If the impact of covariates is the same on all cut-points, we could say that there is a parallel shift. If not, it is not appropriate to interpret the different impact of covariates into reporting heterogeneity because the relationship between covariates and health level may vary by different health levels *(Hernandez-Quevedo et al., 2004)*. In another word, the heterogeneous effect could be due to the heterogeneity in the latent index itself.

We will estimate a standard ordered probit as the baseline model by imposing the assumption of reporting homogeneity. The results of ordered probit will be used later for comparison with a more flexible specification that does not impose the assumption of reporting homogeneity.

Hierarchical ordered probit: heterogeneous reporting behaviour

Vignettes are hypothetical descriptions of several fictive scenarios; they represent fixed latent health levels, respondents are required to rate vignettes as if they were rating their own health. With the assumption of response consistency, if systematic variation across individuals exists in self-reported health, it can also be reflected by answers of vignettes. With the assumption of vignette equivalence, respondents differ only in thresholds, which implies that any individual variation in vignettes ratings would be due to reporting heterogeneity. Therefore, employing the anchoring vignette approach could identify reporting behaviour and then purge reporting heterogeneity from true health effect *(King et al., 2004)*.

The general idea of hierarchical ordered probit (HOPIT) is to use the ratings of vignettes to estimate cut-off points, then using the estimated thresholds and self-reported

health to evaluate true health effect. HOPIT is specified in two parts: one reflecting reporting behaviour and another reflect true health effect.

Reporting behaviour. This component employs generalised ordered probit which includes covariates in cut-off points but not in latent vignette health level. Because the assumption of vignette equivalence, systematic variation in vignette ratings is attributed to reporting heterogeneity as fixed health level depicted by vignette j is the same across respondents. Let $h_{ij}^{v^*}$ and h_{ij}^{v} respectively denote latent health level of vignette and observed vignette ratings of individual i of vignette j, reporting behaviour is modelled as follows:

$$h_{ij}^{v^*} = \alpha_j + \varepsilon_{ij}^{v},\ \varepsilon_{ij}^{v} \sim N(0,1) \tag{6}$$

$$h_{ij}^{v} = k \Leftrightarrow \mu_i^{k-1} \leq h_{ij}^{v^*} < \mu_i^{k}, \qquad k = 0, \dots, K\ , \tag{7}$$

where α_j is the same across individuals according to the assumption of vignette equivalence, $\alpha_1 = 0$ with the normalisation, the error term ε_{ij}^{v} is randomly distributed and obeys standard normal distribution, and $E[h_{ij}^{v^*}]$ depends on the corresponding vignette solely.

The cut-points μ_i^k relies on covariates X_i but do not vary across vignettes j due to the assumption of response consistency; X_i is a vector of covariates that does not contain constant, so γ_0^k are intercepts in respective cut-off points:

$$\mu_i^k = \gamma_0^k + X_i\gamma^k, \quad k = 0, \dots, K \tag{8}$$

$$\mu_i^0 < \mu_i^1 < \cdots < \mu_i^{K-1} < \mu_i^K \text{and } \mu^0 = -\infty, \mu^K = \infty$$

The probability of individual i rating vignette j as category k is denoted by P_{ijk}^{v}:

$$P_{ijk}^{v} = \Pr\left(h_{ij}^{v} = k \middle| X_i\right) = \phi\left(\mu_i^k - \alpha_j\right) - \phi(\mu_i^{k-1} - \alpha_j)$$

$$= \phi\left(\gamma_0^k + X_i\gamma^k - \alpha_j\right) - \phi(\gamma_0^{k-1} + X_i\gamma^{k-1} - \alpha_j) \tag{9}$$

The coefficients of thresholds by each covariate in each domain would reflect the direction and magnitude of variation. We use the estimated coefficients to test parallel shift in thresholds.

Health equation. After defining the reporting behaviour, health equation then can be expressed in a similarly ordered probit, let $h_i^{s^*}$ and h_i respectively indicates latent health measure and observed self-reported health, then the latent health level is specified as follows:

$$h_i^{s^*} = \beta_0 + X_i\beta + \varepsilon_i^s,\ \varepsilon_i^s|X_i \sim N(0, \sigma^2) \tag{10}$$

where X_i is a vector of covariates. The observed self-reported health level is then determined by

$$h_i^s = k \Leftrightarrow \mu_i^{k-1} \le h_i^{s^*} < \mu_i^k,\ k = 0, \dots, K, \tag{11}$$

$$\mu_i^0 < \mu_i^1 < \cdots < \mu_i^{K-1} < \mu_i^K \text{ and } \mu^0 = -\infty, \mu^K = \infty$$

Due to the assumption of response consistency, the cut-points are defined as same as specified in (8).

The probability of individual i rating self-reported health as category k is denoted as P_{ik}^s:

$$P_{ik}^s = \Pr(h_i^s = k) = \Pr(\mu_i^{k-1} - \beta_0 - X_i\beta \le \varepsilon_i^s < \mu_i^k - \beta_0 - X_i\beta)$$

$$= \phi\left(\frac{\mu_i^k - \beta_0 - X_i\beta}{\sigma}\right) - \phi\left(\frac{\mu_i^{k-1} - \beta_0 - X_i\beta}{\sigma}\right) \tag{12}$$

The associated log-likelihood is:

$$lnL = \sum_{i=1}^n \sum_{k=1}^K h_{ik}^s\, lnP_{ik}^s + \sum_{i=1}^n \sum_{j=1}^J \sum_{k=1}^K h_{ijk}^v\, lnP_{ijk}^v \tag{13}$$

where $h_{ik}^s = 1$ if $h_i^s = k$, $h_{ik}^s = 0$ if $h_i^s \neq k$; $h_{ijk}^v = 1$ if $h_{ij}^v = k$, $h_{ijk}^v = 0$ if $h_{ij}^v \neq k$.

We use the estimated coefficients of covariates in health equation compared to that from ordered probit model to assess the true health effect.

DATA

The data came from the WHO Multi-Country Survey Study on Health and Responsiveness 2000-2001 (WHO-MCS). The goal of this study is the development of "valid, reliable, and comparable instruments to describe individual health measure and health system responsiveness on a core set of domains" *(Üstün et al., 2003)*.

The data we use come from face-to-face household surveys which investigated adults aged 18 years or above. A respondent was randomly selected from the eligible individuals in each household and was asked to rate their health measure along with vignette in each of six domains (*mobility*, *cognition*, *pain, self-care*, *usual activities* and *affect*). The survey also set a set of vignettes for each health domain, the description of vignettes is in Appendix B. Each respondent is asked to rate vignettes of two health domains.

We use the WHO-MCS data for the United States and Canada. After dropping missing values of self-reported health, vignette ratings, and the socio-economic variables, the resulting data set contains 848 for the U.S. and 698 for Canada. Table AI in Appendix A documents the number of observations lost due to item non-response, it is clear that health variables (own health and vignettes) contribute mostly to the loss of observations.

Health Variables

Six health domains include *mobility*, *cognition*, *pain*, *self-care*, *usual activities* and *affect*. Own health measure of respondents was obtained from following questions:

- Overall in the last 30 days, how much difficulty did you have with moving around? (mobility)
- Overall in the last 30 days, how much difficulty did you have with self-care, such as washing or dressing yourself? (self-care)

- Overall in the last 30 days, how much difficulty did you have with work or household activities? (usual)
- Overall in the last 30 days, how much pain or discomfort did you have? (pain)
- Overall in the last 30 days, how much distress, sadness or worry did you experience? (affect)
- Overall in the last 30 days, how much difficulty did you have with concentrating or remembering things? (cognition)

Five response categories are: “none”, “mild”, “moderate”, “severe”, and “extreme/cannot do” respectively corresponds to 1, 2, 3, 4 and 5.

A random subsample of respondents was presented with a set of vignettes in each health domain. The vignettes presented hypothetical cases of different health measure; each respondent was asked to rate a set of vignettes in two domains and half respondents rated vignettes in mobility while roughly one-quarter of the samples rated vignettes in other domains. Table AII in Appendix A presents the frequencies of own health and vignettes by domain and country.

Before applying health variables in the model, we reverse the response categories corresponding number, which means “none”, “mild”, “moderate”, “severe”, and “extreme/cannot do” now respectively corresponds to 5, 4, 3, 2 and 1.

Socio-economic variables

Ratings of health measure may differ due to different socio-economic status, if so, then measurement of health would be biased. Frequently, health measurement varies directly with the socioeconomic characteristics. Generally speaking, physical functioning would decrease with age increase; then the reporting behaviour might differ across different age

levels. Similarly, male and female have different preferences which may affect their judgment of health measure. Education is another important variable affecting reporting behaviour. Among all, income seems to have rather a direct effect on responses. Earlier researchers found both supports and disagreements of income-related reporting heterogeneity. In conclusion, age, gender, education, and income are considered relating socio-demographic variables in this paper.

After deciding age, gender, education and income as sociodemographic variables, a precise definition for each variable is necessary. Age is denoted by four categories: 18-29 years as the reference group, 30-44 (AGE3044), 45-59 (AGE4559), and more than 60 (AGE60). Gender is represented by a dummy variable: FEMALE, it equals to 1 if the respondent is a female. Education expresses in the number of years of completed school. Income, however, is divided into two groups: lower income group as the reference group of those who in the first, second and third income quintile, and higher income group is for those who are in the fourth and fifth income quintile.[1] We denote INC to represent income, if the respondent is in the higher income group, then the value of INC equals to 1. Table I presents descriptive statistics for the covariates by country.

Table I. Descriptive statistics of covariates

	United States		Canada	
Variables	Mean	Std. dev.	Mean	Std. dev.
AGE3044	0.281	0.450	0.298	0.458
AGE4559	0.294	0.456	0.268	0.443
AGE60	0.343	0.475	0.199	0.400
FEMALE	0.462	0.499	0.536	0.499
EDUC	13.903	3.218	14.102	3.616
INC	0.300	0.458	0.364	0.481
N	848		698	

[1] Our analysis decision is to treat the variable as an ordered categorical variable by dichotomising the income into higher and lower income groups. Details about income variable (aboutu6) are presented in Table AIII in Appendix.

From Table 1, the age structures are slightly different for two countries. Respondents who are older than 30 years but younger and 45 years and those who are older than 45 years but younger than 60 years both consist of nearly 30% of the whole sample in two countries. However, those who are older than 60 years makes up 34.3 percent of American sample while the corresponding percentage in Canada is less than 20. For gender, over half the Canadian sample (53.6%) is women while the percentage of female respondents in the U.S. is 46.2%. Nevertheless, education level in two countries are similar with the mean value of years spent in school around 14 for both the U.S. and Canada. Besides, 36.4% of respondents of Canada are in the higher income group while the number of percentage is 30% for the U.S. Since we test reporting heterogeneity by covariate in the following content, the difference in the descriptive statistics may help us understand and explain the various patterns of reporting behaviours in different countries.

RESULTS

For each country, I estimate the ordered probit model (Equation (1)-(4)) and HOPIT model (Equation (6)-(13)) for each of the six health domains. Both the index function and the cut-points employ the same covariates: AGE3044, AGE4559, AGE60, FEMALE, EDUC and INC. The health function in the vignette component of the HOPIT model includes only dummies representing the respective vignettes in each domain. I firstly present results of a test for homogeneous reporting behaviour, then turn to estimated magnitudes of reporting heterogeneity and end with the true health effects that after the adjustments of reporting heterogeneity.

Test of reporting homogeneity

Table 2 presents the results of tests of homogeneous reporting and parallel shifts on the thresholds.

Table II. Log-likelihood ratio tests of homogeneity and parallel cut-point shift: *p*-values

	Homogeneity					Parallel shift
	All	Age	Female	Educ	income	All
U.S.						
Mobility	0.064	0.207	0.003	0.720	0.818	0.054
Cognition	0.000	0.082	0.147	0.000	0.169	0.004
Pain	0.000	0.000	0.000	0.001	0.005	0.000
Self-care	0.000	0.000	0.002	0.014	0.151	0.000
Usual	0.000	0.000	0.073	0.000	0.494	0.013
Affect	0.048	0.891	0.002	0.661	0.086	0.159
Canada						
Mobility	0.025	0.063	0.522	0.036	0.158	0.103
Cognition	0.002	0.221	0.614	0.001	0.058	0.091
Pain	0.001	0.326	0.329	0.004	0.025	0.033
Self-care	0.004	0.004	0.671	0.365	0.214	0.143
Usual	0.000	0.010	0.982	0.000	0.289	0.108
Affect	0.036	0.154	0.937	0.174	0.041	0.013

For reporting homogeneity, the first column indicates the p-values of reporting homogeneity when age, gender, education, and income are jointly considered. The other columns present p-values of the possibility of reporting homogeneity by individual characteristics. Results show that the null hypothesis of reporting homogeneity when all covariates are jointly considered is rejected (5% or less) in all domains for both countries except the *mobility* for the U.S. Homogeneous reporting by age, however, is rejected in three domains for the U.S. and two domains for Canada. Homogeneity by gender is not rejected in all domains of Canada but is rejected in almost all domains of the U.S. (except for *cognition*). However, homogeneity by education is rejected in four domains respectively for the U.S. and Canada. The evidence of reporting heterogeneity by income seems to be weak because the null hypothesis of homogeneity is not rejected in five domains for the U.S. and four domains for Canada.

The last column presents the probability of parallel shift of cut-points when all covariates are jointly considered. The parallel shift is rejected in most domains of the U.S. (except *affect* and *mobility*), which implies reporting behaviour is stronger at some levels of health then the covariates shift the cut-points by different magnitude. However, parallel shift on thresholds is only rejected in two domains for Canada. Parallel shift indicates covariates shift the cut-points up or down by the same magnitude, which implies that the relative positions of thresholds remain unaltered.

These results confirm the existence of reporting heterogeneity, but the reporting differences are not uniform for two countries. While reporting heterogeneity is stronger at some levels of health across all health domains in the U.S., it shifts the thresholds by the same magnitude in most domains of Canada. Though the results indicate many uniform shifts in cut-points for Canada, there is an evidence of reporting heterogeneity for some health domains. This situation supports the need for an HOPIT model that accounts for reporting heterogeneity in self-reported health. Besides, the conclusion about Canada is

partly consistent with *Lindeboom and van Doorslaer (2004)* which reports little evidence of reporting heterogeneity by income and those significant reporting differences show the parallel shifts in Canada. Similarly, the results of the U.S. are consistent with *Dowd and Todd (2011)* which finds significant reporting heterogeneity across most health domain in the U.S.

Reporting behaviour

The response categories for the questions: 'How much difficulty/pain/distress...' range from 'severe/cannot do' to 'none' which correspond to 1 to 5. Therefore, the positive coefficients across all cut-points of a certain covariate indicate higher standards of health ratings, in other words, lower probabilities of reporting better health level. Accordingly, negative coefficients of threshold suggest higher probabilities of reporting better health levels because respondents tend to lower their standards. Consistent with homogeneity tests, there are significant effects of different covariates on cut-points for two countries.

Table 3 presents the cut-points coefficients for age groups. According to the table, both the U.S. and Canada show mild evidence of reporting heterogeneity by age. In the U.S., the older respondents tend to report heterogeneously but in different directions across domains. In the domain of *pain*, significant coefficients across all age groups are all positive and are all in lower cut-points, implying individuals tend to exaggerate pain or discomfort when facing extreme, serve and moderate problems about pain. Nevertheless, the coefficients of uppermost cut-points in *self-care* are negative while that in the domain of *usual activities* are positive; this means older American tend to rate minor problem of *self-care* less severe but that of usual activities more severe. Though all significant coefficients are positive and are most from upper cut-points in Canada, only Canadian

Table III. Estimated coefficients of age in the cut-points

		U.S.				Canada			
		ctpt1	ctpt2	ctpt3	ctpt4	ctpt1	ctpt2	ctpt3	ctpt4
Mobility	AGE3044	0.001	-0.083	-0.099	-0.082	-0.039	0.086	0.058	0.002
		(0.995)	(0.547)	(0.484)	(0.591)	(0.763)	(0.425)	(0.607)	(0.989)
	AGE4559	0.156	0.075	0.007	-0.067	0.253	**0.250**	0.196	-0.055
		(0.359)	(0.582)	(0.962)	(0.660)	(0.057)	(0.027)	(0.099)	(0.677)
	AGE60	0.107	-0.114	-0.088	**-0.286**	-0.221	0.148	0.176	-0.146
		(0.517)	(0.390)	(0.519)	(0.050)	(0.168)	(0.252)	(0.183)	(0.315)
Cognition	AGE3044	-0.095	-0.220	-0.108	-0.103	-0.160	0.052	**0.285**	0.139
		(0.632)	(0.154)	(0.494)	(0.581)	(0.300)	(0.662)	(0.013)	(0.320)
	AGE4559	-0.141	-0.189	0.138	0.193	-0.193	0.141	**0.265**	0.252
		(0.483)	(0.220)	(0.385)	(0.306)	(0.231)	(0.234)	(0.024)	(0.082)
	AGE60	-0.366	-0.260	-0.080	0.000	0.065	0.044	0.218	0.147
		(0.063)	(0.082)	(0.607)	(0.999)	(0.682)	(0.715)	(0.075)	(0.324)
Pain	AGE3044	0.419	**0.630**	0.260	0.236	-0.009	0.012	0.068	0.071
		(0.081)	(0.000)	(0.102)	(0.238)	(0.955)	(0.923)	(0.604)	(0.737)
	AGE4559	**0.503**	**0.661**	0.067	0.103	0.211	0.225	**0.282**	0.224
		(0.034)	(0.000)	(0.668)	(0.595)	(0.165)	(0.060)	(0.032)	(0.289)
	AGE60	0.271	**0.481**	0.031	-0.158	0.008	0.254	0.242	0.018
		(0.250)	(0.004)	(0.838)	(0.403)	(0.961)	(0.058)	(0.101)	(0.937)
Self-care	AGE3044	0.239	0.076	0.151	0.011	-0.125	0.066	0.210	0.195
		(0.324)	(0.666)	(0.389)	(0.967)	(0.473)	(0.607)	(0.128)	(0.295)
	AGE4559	0.146	0.040	-0.093	**-0.678**	**0.435**	**0.307**	**0.439**	0.262
		(0.544)	(0.817)	(0.583)	(0.006)	(0.008)	(0.015)	(0.001)	(0.147)
	AGE60	-0.010	-0.033	-0.075	**-0.915**	0.086	0.151	**0.397**	0.079
		(0.966)	(0.847)	(0.652)	(0.000)	(0.658)	(0.306)	(0.011)	(0.696)
Usual	AGE3044	-0.265	-0.218	-0.128	0.211	0.180	0.033	0.175	-0.199
		(0.188)	(0.162)	(0.423)	(0.307)	(0.231)	(0.771)	(0.141)	(0.233)
	AGE4559	-0.146	0.208	0.137	**0.462**	0.096	**0.356**	**0.323**	-0.254
		(0.470)	(0.182)	(0.389)	(0.027)	(0.549)	(0.003)	(0.009)	(0.136)
	AGE60	-0.352	0.043	0.272	**0.441**	0.169	0.185	0.230	0.022
		(0.075)	(0.777)	(0.081)	(0.028)	(0.282)	(0.115)	(0.063)	(0.902)
Affect	AGE3044	-0.039	0.162	-0.136	-0.173	-0.059	0.032	0.097	0.053
		(0.846)	(0.372)	(0.497)	(0.470)	(0.672)	(0.797)	(0.446)	(0.735)
	AGE4559	-0.140	0.111	-0.039	0.006	0.131	-0.057	-0.227	-0.125
		(0.484)	(0.542)	(0.844)	(0.981)	(0.378)	(0.671)	(0.094)	(0.449)
	AGE60	0.010	0.055	-0.054	0.046	-0.348	-0.038	-0.069	-0.221
		(0.958)	(0.757)	(0.787)	(0.848)	(0.066)	(0.809)	(0.675)	(0.247)

Note: *p*-values in parentheses. Bold indicates significance at 5%. Ctpt indicates cut-point, ctpt1 is the lowest cut-point determining probability of extreme difficulty/pain/distress, ctpt4 is the highest cut-point determining probability of no difficulty/pain/distress.

aged 45 to 59 years have significantly positive and larger shift in cut-points, which indicated that they consistently rate a given health problem across all domains (except

affect) more severe. However, we find no evidence showing significant reporting heterogeneity for the oldest age group.

Table IV. Estimated coefficients of female and education in the cut-points

FEMALE	U.S.				Canada			
	ctpt1	ctpt2	ctpt3	ctpt4	ctpt1	ctpt2	ctpt3	ctpt4
Mobility	**0.319**	-0.01	0.008	-0.083	0.131	0.053	-0.062	-0.065
	(0.000)	(0.889)	(0.910)	(0.274)	(0.181)	(0.510)	(0.464)	(0.484)
Cognition	**-0.208**	-0.071	0.057	0.07	-0.019	0.115	0.05	-0.012
	(0.035)	(0.326)	(0.438)	(0.431)	(0.865)	(0.160)	(0.535)	(0.904)
Pain	0.152	**0.325**	**0.387**	**0.385**	-0.151	-0.091	-0.142	0.012
	(0.124)	(0.000)	(0.000)	(0.000)	(0.142)	(0.273)	(0.124)	(0.934)
Self-care	-0.019	0.043	**0.256**	**0.382**	-0.012	-0.063	-0.135	-0.018
	(0.874)	(0.613)	(0.002)	(0.000)	(0.918)	(0.470)	(0.158)	(0.888)
Usual	**-0.253**	-0.128	0.019	0.071	0.013	0.044	0.042	0.035
	(0.014)	(0.092)	(0.816)	(0.509)	(0.899)	(0.584)	(0.619)	(0.769)
Affect	**0.418**	0.15	0.072	-0.056	-0.009	0.06	0.032	-0.036
	(0.000)	(0.108)	(0.488)	(0.661)	(0.932)	(0.531)	(0.746)	(0.765)
EDUC	U.S.				Canada			
	ctpt1	ctpt2	ctpt3	ctpt4	ctpt1	ctpt2	ctpt3	ctpt4
Mobility	-0.012	-0.012	-0.01	-0.001	0.131	0.053	-0.062	-0.065
	(0.397)	(0.266)	(0.377)	(0.958)	(0.181)	(0.510)	(0.464)	(0.484)
Cognition	**-0.068**	**-0.031**	-0.023	0.018	**-0.043**	**-0.036**	**-0.043**	-0.025
	(0.000)	(0.018)	(0.075)	(0.240)	(0.006)	(0.003)	(0.000)	(0.095)
Pain	-0.007	0.011	**0.034**	**0.058**	-0.035	**-0.045**	-0.014	0.03
	(0.629)	(0.326)	(0.004)	(0.000)	(0.055	(0.001)	(0.364)	(0.202)
Self-care	-0.027	0.026	**0.026**	**0.038**	-0.004	-0.012	-0.018	0.024
	(0.145)	(0.055)	(0.043)	(0.012)	(0.830)	(0.427)	(0.250)	(0.264)
Usual	**-0.062**	**-0.056**	**-0.058**	**-0.043**	**-0.034**	**-0.032**	**-0.047**	**-0.054**
	(0.001)	(0.000)	(0.000)	(0.035)	(0.020)	(0.005)	(0.000)	(0.004)
Affect	0.008	-0.006	0.005	0.029	-0.019	0.013	-0.012	-0.01
	(0.676)	(0.712)	(0.794)	(0.207)	(0.209)	(0.323)	(0.362)	(0.519)

Note: *p*-values in parentheses. Bold indicates significance at 5%. Ctpt indicates cut-point, ctpt1 is the lowest cut-point determining probability of extreme difficulty/pain/distress, ctpt4 is the highest cut-point determining probability of no difficulty/pain/distress.

Table 4 presents cut-points coefficients for gender and education. Unlike the results for age, Table 4 indicates different reporting behaviours between countries. Reporting heterogeneity by gender is rejected in Canada across all domains. However, the coefficients of cut-points are significant at least one cut-points for all health domains in the U.S. In the domains of *pain* and *self-care*, significantly positive coefficients of upper thresholds indicate larger movements compared to lower thresholds, which indicates that female American have lower probabilities reporting better health facing minor problems of pain or self-care. For reporting heterogeneity by education, better educated Canadian are inclined to rate a given vignette into a better category due to the significant and negative coefficients across all the thresholds in the domain of *cognition* and *usual activities*, and the American respondents show the same pattern in the same domains. However, in the domain of *pain* and *self-care*, better-educated American have a higher standard of health for higher cut-points.

Table V. Estimated coefficients of income in the cut-points

	U.S.				Canada			
	ctpt1	ctpt2	ctpt3	ctpt4	ctpt1	ctpt2	ctpt3	ctpt4
Mobility	0.051	-0.081	-0.023	-0.013	0.124	0.09	0.095	**0.237**
	(0.627)	(0.336)	(0.792)	(0.889)	(0.236)	(0.306)	(0.309)	(0.022)
Cognition	**-0.227**	-0.163	-0.107	-0.033	0.221	-0.03	0.047	**0.235**
	(0.043)	(0.053)	(0.214)	(0.758)	(0.064)	(0.740)	(0.614)	(0.042)
Pain	-0.211	-0.09	0.044	**0.314**	-0.083	**0.214**	**0.230**	**0.385**
	(0.067)	(0.280)	(0.610)	(0.004)	(0.493)	(0.024)	(0.029)	(0.025)
Self-care	0.085	0.139	0.087	**0.276**	0.012	**0.212**	0.073	-0.104
	(0.516)	(0.145)	(0.362)	(0.022)	(0.923)	(0.037)	(0.509)	(0.485)
Usual	-0.129	-0.033	-0.001	0.172	0.123	-0.046	-0.023	0.201
	(0.264)	(0.707)	(0.995)	(0.181)	(0.275)	(0.607)	(0.813)	(0.126)
Affect	0.107	0.169	0.224	**0.402**	-0.062	0.031	0.045	**0.371**
	(0.389)	(0.132)	(0.079)	(0.009)	(0.587)	(0.760)	(0.670)	(0.004)

Note: *p*-values in parentheses. Bold indicates significance at 5%. Ctpt indicates cut-point, ctpt1 is the lowest cut-point determining probability of extreme difficulty/pain/distress, ctpt4 is the highest cut-point determining probability of no difficulty/pain/distress.

Table 5 presents coefficients of cut-points by income groups. From the table, Canada shows significant reporting heterogeneity by income because there is a significant effect on at least on cut-point for almost all domains except for *usual activities*. The significant coefficients are all positive, which indicates that respondents with higher income level have lower probabilities of reporting very good health status. In specific, the coefficients are positive across almost all cut-points (except *ctpt1*) in the domain of *pain*. Also, in Canada, most significant coefficients are for the uppermost cut-points, implying individuals with higher income have lower probabilities evaluating a vignette corresponding to no difficulty/pain/distress. Similarly, higher income groups in the U.S. also have higher standards for uppermost cut-points in the domain of *pain*, *self-care* and *affect*. An exception is the domain of *cognition*, where the significant coefficient is negative, in another word, American with higher income has lower standards of having the extreme problem concentrating or remembering things.

According to the results of coefficients of cut-points by individual characteristics, the reporting heterogeneity exists in both countries. However, the reporting behaviours are not uniform between countries, for example, the results show no evidence of reporting heterogeneity by gender across all domains in Canada but strong evidence of reporting heterogeneity by gender in the U.S. across all domains. Also, the reporting behaviour is not uniform across all domains within a country, for example, better-educated American have a higher standard regarding problems of pain but a lower health standard for usual activities. In general, male,

However, since most coefficients of cut-points are not significant, it is difficult to evaluate directly the reporting behaviours based on cut-points. To present the reporting behaviour intuitively, we estimate the effect of reporting behaviour by each covariate across all domains. By using the parameters we obtained in the reporting model (6) to (9), we estimate the probabilities respectively for a reference respondent and the same reference respondent but with a certain individual characteristic from a top quintile, rating a vignette as ‘none’. The ratio of these two probabilities measures the relative magnitude

of the reporting effect. For the reference respondent, we use a male individual from the youngest age group (18 to 29 years), never go to school, and with the lower income level, then the corresponding respondent is the same respondent but now with a covariate of the top category. Take an example, to assess the effect of reporting effect by age, firstly calculating the probability that reporting the best health level of the reference respondent in each domain, then calculating the same probability of the same individual but who are older than 60 years. The ratio of these two probabilities represents the 'relative magnitude of the reporting effect' *(Bago d'Uva et al., 2008)* by age. If a ratio of a domain is smaller than one, it indicates that the oldest respondents have lower probabilities rating a given vignette as very good health compared to the youngest respondents holding other characteristics constant, which implies the older have a higher standard of health. Therefore, a ratio larger than one implies a respondent from the oldest age group has higher probability reporting no difficulty/pain/distress compared to a reference respondent. Only when a ratio equals to one representing both the older and younger respondents have the same probabilities evaluating a given vignette as very good health, so either a ratio smaller or larger than one is a sign of reporting heterogeneity. Besides, the difference between the ratios and one represent the effect of reporting behaviour by a covariate. We do the calculations for all covariates, change in turn age from oldest to youngest, gender from male to female, education lever from the lowest to the highest[2] and income from lowest to highest. Results are presented in Figure 1.

In the U.S., there are either no difference reporting by age or the oldest respondents have higher probabilities reporting very good health. The greatest differences are in the domain of *mobility*, *pain* and *usual activities* with the oldest group having probabilities of reporting no difficulty/pain/distress that are all 6% more than the youngest group. However, the results of Canada indicate that respondents who are in the oldest group are less likely to report very good health in the domain of *cognition*, *pain* and *usual activities*

[2] Because education is not a categorical data, we denote the lowest education level as never been to school and denote the highest education as the number of years spent in school is 14.

Figure 1. Relative probabilities of reporting very good health by socio-economic group

with the greatest difference is in the *pain* (5%). For other domains, Canadian respondents who are older than 60 years show the same reporting behaviour as American.

For reporting behaviour by gender, American females are less likely to report excellent health than males in most domains. The effect is largest for *pain*, with a difference of 17% relative probabilities. In Canada, females have equal probabilities reporting excellent health compared to males in the domain of *cognition* and *self-care* but lower probabilities reporting very good health in the domain of *pain* and *usual* and higher probabilities reporting very good health in the domain of *mobility* and *affect*. Nevertheless, those significant reporting differences for *mobility*, *pain* and *affect* are all only 1%, which is unimportant. In general, gender seems to have no effect on health reporting behaviour for Canada but female American respondents have lower probabilities rating a given vignette as no difficulty/pain/distress. The result of the U.S. is consistent with that of *Dowd and Todd (2011)* and *Molina (2016)* which both find men are generally more optimistic compared to women, but the result of Canada seems to contradict to that of *Lindeboom and van Doorslaer (2004)* which indicates that female Canadian are more likely to report better health, this difference may be the different approaches used to identify reporting behaviours.

As shown in the results of both the U.S. and Canada, better-educated respondents are less likely to report very good health for most domains; the greatest differences are both in the domain of *pain* where the better-educated respondents have respectively 38% and 37% smaller probabilities reporting excellent health compared to less-educated respondents for the U.S. and Canada. On the contrary, better-educated American have 17% larger probabilities reporting very good health in the domain of *usual activities* and Canadian with higher education level has respectively 10% and 5% larger probabilities reporting very good health in the domain of *cognition* and *affect*. For reporting behaviour by education, the results are consistent with *Bago d'Uva et al. (2008)* which finds evidence to support that better-educated respondents of countries with higher education levels have higher health expectations. Because both the U.S. and Canada are developed countries, they have higher education levels which may enable respondents better understand the health status described in a given vignette.

According to the figure, the results of both the U.S. and Canada indicate that people with higher income are less likely to report very good health in all domains except the *mobility* for the U.S.; the largest differences for both countries are in the domain of *pain*, with the relative probabilities reaching 13% and 35% respectively for the U.S. and Canada. In general, respondents with higher income are less likely to report better health across all domains for both Canada and the U.S.

The results of relative probabilities of reporting very good health by the socioeconomic groups are mostly consistent with the results of estimated coefficients of socioeconomic variables in the cut-points. In general, male, old, and less-educated respondents with lower income are more likely to evaluate a given vignette as very good health.

True health effect

With the existence of reporting heterogeneity, the coefficients estimated of the standard ordered probit model (1) contain both heterogeneous reporting behaviour and true health effect. Therefore, any analysis based on these coefficients would be invalid and biased. Employing anchoring vignette approach, we separate reporting heterogeneity from true health effect, so the coefficients identified by HOPIT model reflect true health effect.

Table 6 presents estimated coefficients of age groups before and after adjustment. For both the U.S. and Canada, the results of the ordered probit and HOPIT both show significantly negative relationship between age and almost all health domains (except *cognition* and *affect*) among the middle-aged (45-59 years) and the oldest (more than 60 years) respondents. Consistent with the reporting behaviour that old respondents are more likely to report better health in two countries., the adjustment increases the magnitude of significant coefficients in most domains (*mobility*, *pain*, *self-care* and *usual activities*), which indicates that the association between age and health is underestimated without considering reporting heterogeneity.

Table VI. Estimated coefficients of age groups before and after adjustment

		U.S.		Canada	
		Before	After	Before	After
Mobility	AGE3044	**-0.384**	**-0.504**	0.052	0.096
		(0.037)	(0.037)	(0.724)	(0.733)
	AGE4559	**-0.699**	**-0.805**	**-0.584**	**-1.029**
		(0.000)	(0.001)	(0.000)	(0.000)
	AGE60	**-0.926**	**-1.226**	**-0.557**	**-1.065**
		(0.000)	(0.000)	(0.000)	(0.000)
Cognition	AGE3044	-0.027	-0.147	0.178	**0.378**
		(0.861)	(0.517)	(0.156)	(0.047)
	AGE4559	-0.047	0.094	-0.151	0.062
		(0.758)	(0.679)	(0.227)	(0.746)
	AGE60	-0.245	-0.307	-0.099	0.048
		(0.102)	(0.163)	(0.474)	(0.814)
Pain	AGE3044	**-0.378**	-0.136	0.086	0.202
		(0.019)	(0.570)	(0.475)	(0.443)
	AGE4559	**-0.730**	**-0.640**	**-0.484**	**-0.596**
		(0.000)	(0.007)	(0.000)	(0.020)
	AGE60	**-0.810**	**-0.887**	**-0.555**	**-0.823**
		(0.000)	(0.000)	(0.000)	(0.003)
Self-care	AGE3044	-0.365	-0.567	-0.293	-0.757
		(0.259)	(0.315)	(0.268)	(0.381)
	AGE4559	**-0.799**	**-1.900**	**-0.823**	**-2.350**
		(0.010)	(0.001)	(0.001)	(0.007)
	AGE60	**-0.989**	**-2.388**	**-0.700**	**-2.121**
		(0.001)	(0.000)	(0.007)	(0.018)
Usual	AGE3044	**-0.453**	-0.707	-0.016	-0.169
		(0.034)	(0.112)	(0.918)	(0.667)
	AGE4559	**-0.684**	**-0.876**	**-0.582**	**-1.516**
		(0.001)	(0.046)	(0.000)	(0.000)
	AGE60	**-1.027**	**-1.540**	**-0.582**	**-1.350**
		(0.000)	(0.000)	(0.000)	(0.001)
Affect	AGE3044	0.072	-0.018	0.136	0.267
		(0.622)	(0.952)	(0.232)	(0.182)
	AGE4559	0.087	0.126	-0.223	**-0.453**
		(0.550)	(0.668)	(0.055)	(0.027)
	AGE60	**0.335**	0.556	0.205	0.133
		(0.019)	(0.055)	(0.116)	(0.577)

Note: *p*-values in parentheses. Bold indicates significance at 5%.

Table VII. Estimated coefficients of gender and education groups before and after adjustment

FEMALE	U.S.		Canada	
	Before	After	Before	After
Mobility	0.032	-0.021	0.093	0.114
	(0.695)	(0.844)	(0.336)	(0.545)
Cognition	-0.010	0.038	**0.234**	**0.282**
	(0.898)	(0.726)	(0.009)	(0.037)
Pain	-0.037	**0.335**	0.005	-0.031
	(0.629)	(0.002)	(0.957)	(0.859)
Self-care	0.172	**0.603**	0.106	0.309
	(0.129)	(0.003)	(0.464)	(0.517)
Usual	-0.063	-0.092	-0.030	-0.035
	(0.481)	(0.624)	(0.768)	(0.892)
Affect	**-0.165**	-0.246	0.111	0.161
	(0.027)	(0.109)	(0.181)	(0.278)
EDUC	U.S.		Canada	
	Before	After	Before	After
Mobility	**0.040**	**0.040**	0.017	0.019
	(0.003)	(0.023)	(0.233)	(0.492)
Cognition	**0.049**	**0.051**	0.012	-0.017
	(0.000)	(0.005)	(0.388)	(0.408)
Pain	**0.034**	**0.080**	0.003	0.017
	(0.006)	(0.000)	(0.828)	(0.542)
Self-care	**0.049**	**0.112**	0.018	0.079
	(0.008)	(0.001)	(0.403)	(0.284)
Usual	**0.032**	0.011	**0.039**	0.043
	(0.029)	(0.740)	(0.010)	(0.271)
Affect	**0.036**	**0.074**	-0.003	-0.014
	(0.003)	(0.004)	(0.797)	(0.508)

Note: *p*-values in parentheses. Bold indicates significance at 5%.

Table 7 presents estimated coefficients of gender and education groups before and after adjustment. The health effect by gender and education is not important across almost all health domains in Canada. However, the results of ordered probit show significantly positive relationship between education and all health domains in the U.S. After adjustment, all significant coefficients of gender and education increase, this is consistent with reporting behaviour by gender and education in the U.S. Because females

are less likely to report better health and so do the better-educated respondents, coefficients of female and education would increase after purging reporting bias.

Table VIII. Estimated coefficients of income groups before and after adjustment

	U.S.		Canada	
	Before	After	Before	After
Mobility	**0.293**	**-0.923**	**0.222**	**0.593**
	(0.003)	(0.006)	(0.044)	(0.005)
Cognition	0.085	0.038	0.012	0.196
	(0.349)	(0.772)	(0.906)	(0.197)
Pain	0.123	**0.336**	0.160	**0.596**
	(0.164)	(0.008)	(0.091)	(0.003)
Self-care	**0.471**	**1.028**	**0.424**	**1.291**
	(0.002)	(0.000)	(0.019)	(0.036)
Usual	**0.465**	**1.012**	0.099	0.380
	(0.000)	(0.000)	(0.389)	(0.194)
Affect	0.159	**0.578**	0.132	**0.396**
	(0.067)	(0.002)	(0.151)	(0.015)

Note: *p*-values in parentheses. Bold indicates significance at 5%.

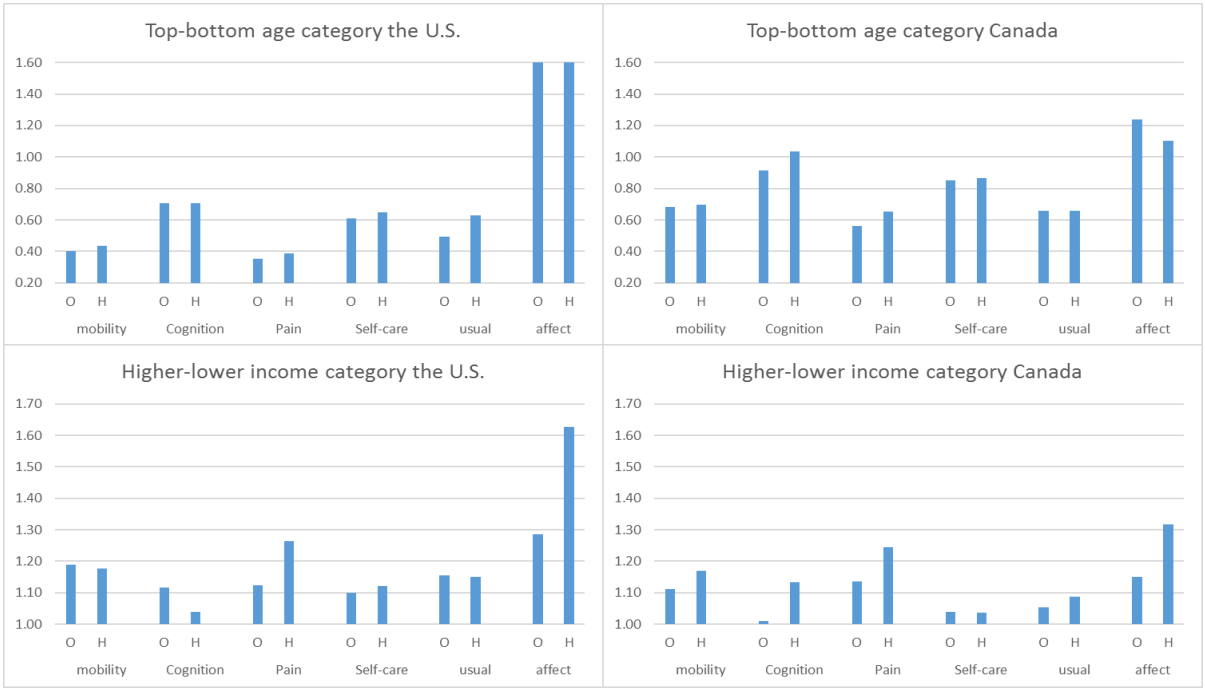
Table 8 presents estimated coefficients of income groups before and after adjustment. According to the ordered probit results, higher income has significantly positive impact on three health domains (*mobility*, *self-care* and *usual activities*). After the adjustment, the magnitude of the effect of income increases tremendously in the domain of *pain*, *self-care*, *usual activities* and *affect* for the U.S., which is consistent with reporting behaviour shown in Figure 1. Similarly, the correction tremendously rise income gradients in four of six domains (*mobility*, *pain*, *self-care* and *affect*) for Canada.

The conclusion is that while both age and income have significant effects on health for two countries, gender and education only have the significant effect on health for the U.S. In specific, age is negatively correlated with health and income has the positive effect on

health. Besides, the true health effect is usually underestimated if ignoring reporting heterogeneity.

Again we quantify the effects of the adjustment for reporting heterogeneity. Because the true health effects are only significant by age and income for both countries, we calculate per domain the probability of being in the top category of health for the reference respondent and the same individual but in the oldest age group or with higher income level while holding other characteristics constant. Calculating the ratios of these two probabilities respectively for ordered probit model and HOPIT; the former reflects both reporting heterogeneity and true health effect, but the later indicates pure health effect. If the ratio is smaller than one, it means the effect of the covariate is negative, in another word, this covariate is negatively correlated with self-reported health. Besides, the difference between the ratios of ordered probit and HOPIT measures the extent of the bias of health effect. Figure 2 presents the relative probabilities.

Figure 2. Relative probabilities of reporting very good health by age and education



Note: O and H respectively refer to the ratio for ordered probit and HOPIT.

From the first row of the figure, it is evident that the age gradients for Canada are larger across almost all domains (except *affect*) compared to the U.S. The correction gives a significant rise to age gradients in most domains for both countries. The greatest difference for the U.S. is in the domain of *usual activities* with the percentage reaching 13, but for Canada, the largest change is for pain where the possibility for the oldest respondents reporting very good health increases 9%. Consistent with Table VI, the relationship between income and health are negative for almost all domains (except *affect*) in both countries.

Similar to age gradients, income gradients also increase across almost all domains (except *mobility* and *cognition* for the U.S.) after the adjustment for both countries. The greatest difference for the U.S. is in the domain of *affect* with the ratio increasing 35% after the correction and that for Canada is in the domain of *pain* where respondents with higher income have 11% larger possibility reporting very good health compared to individuals with lower income. Besides, all ratios are larger than 1, which indicates that income has a positive effect on health for both countries.

From the results, it is not hard to find out that both age and income have the significant effect on health; age and health are negatively correlated, but income has a positive relationship with health. Because higher age has a negative effect on health, the optimistic reporting behaviour of older respondents masks part of true health. Similarly, the pessimistic reporting behaviour of high-income groups decreases the positive health effect by income. After purging the reporting heterogeneity, both age and income gradients increase. In another word, both gradients are underestimated without considering reporting heterogeneity by age and income.

CONCLUSION AND DISCUSSION

This paper employs the approach of anchoring vignettes to test for reporting heterogeneity by socioeconomic groups in the self-reported health of adults in two developed countries (the U.S. and Canada). Using the WHO-MCS data which includes both the self-reported health and assessments of vignettes, we estimated HOPIT in two parts: firstly estimate the effect of reporting behaviour on health ratings, then estimate the relationship between socioeconomic variables and self-reported health and purge the effect of reporting heterogeneity off health effect. We use the results to test reporting heterogeneity and examine the impact of correction on health disparities by age and income.

The null hypothesis of reporting homogeneity when jointly considered all covariates is rejected in all health domains for two countries (except the *mobility* in the U.S.). The parallel shift of reporting cut-points is rejected in most domains of the U.S. but only two domains in Canada. We find strong evidence of reporting heterogeneity by socioeconomic variables in two countries, but the reporting behaviours are not uniform neither between countries nor across health domains, which is consistent respectively with *Bago d'Uva et al., (2008)* and *Hirve et al., (2014)*. In general, male, old, and less-educated respondents with lower income are more likely to report better health. The general result of reporting behaviour is consistent with earlier studies of developed countries that male *(Doorslaer and Gerdtham, 2003; Molina, 2016)*, older *(Doorslaer and Gerdtham, 2003; Lindeboom and van Doorslaer, 2004; Dowd and Todd, 2011)* and less-educated *(Kerkhofs and Lindeboom, 1995)* respondents are inclined to report health positively. As for the health effect, age has a negative effect on health, but income has a positive effect on health. After the correction for reporting heterogeneity, health disparities by age and income both increase significantly.

To my knowledge, this is the first study to test for reporting heterogeneity by socioeconomic groups in the U.S. and Canada using anchoring vignettes. In general, both reporting behaviour and health effect are similar in two countries despite they have different health care systems. Though I follow the method proposed by *Bago d'Uva et al. (2008)*, the results are somewhat different. The reporting behaviour is opposite in some cases across developed countries compared to that of developing countries. For example, older individuals tend to report better health in this paper while the same respondents from developing countries tend to report lower health levels. One reason might be the different context, where older people in developing countries have higher expectations of health. Another reason may be the different understanding of health, *Hirve et al. (2014)* suggests that different wordings of vignettes may have the different effect on health ratings. Future studies may give consideration to whether changes in wordings would affect health ratings. Besides, though our results are consistent with multiple studies, there are also contradictions. For example, *Lindeboom and van Doorslaer (2004)* indicate that females are more likely to report better health in Canada, but we find no evidence of reporting heterogeneity by gender in Canada. One reason might be different approaches used to identify reporting behaviour; another reason might be different data. Also, the relatively smaller number of observations in our paper may also contribute to the contradiction. Because the observations lost mostly due to non-response on own health and vignettes, further studies may need to increase the health domains to test whether larger number of observations change the results.

APPENDIX A

Table AI. Sample sizes and item non-response

	U.S.	Canada
Full Sample	1792	1594
Observations lost due to item non-response		
Own health domains	642	808
Health vignettes	629	816
All covariates	448	153
Final sample	848	698

Table AII. Frequencies of own health and vignettes by domain and country

	U.S.									Canada								
	Own	vig1	vig2	vig3	vig4	vig5	vig6	vig7	vig8	Own	vig1	vig2	vig3	vig4	vig5	vig6	vig7	vig8
Mobility																		
Extreme	0.24	1.46	0.73	0.98	2.93	10.73	90.00			1.29	0.29	0.00	0.87	7.83	13.91	86.67		
Severe	3.18	0.49	0.49	5.61	46.34	68.29	5.12			3.01	0.29	0.29	7.54	45.51	59.71	11.59		
Moderate	13.33	1.46	1.71	35.85	41.22	17.32	0.49			8.17	1.74	2.03	35.07	36.52	22.61	0.29		
Mild	24.53	2.93	5.37	50.00	8.29	2.44	0.73			16.62	4.06	4.93	46.67	8.41	2.32	0.58		
None	58.73	93.66	91.71	7.56	1.22	1.22	3.66			70.92	93.62	92.75	9.86	1.74	1.45	0.87		
N	848	410	410	410	410	410	410			*698*	*345*	*345*	*345*	*345*	*345*	*345*		
Cognition																		
Extreme	0.71	0.47	4.21	0.93	0.93	7.94	3.27	28.50	82.71	0.43	1.14	5.14	0.57	2.29	2.86	4.00	30.29	79.43
Severe	2.95	1.87	32.24	3.74	19.63	33.64	27.10	51.40	14.02	1.29	0.57	29.14	7.43	22.86	32.00	26.29	48.57	18.29
Moderate	13.33	1.40	34.58	25.70	40.65	45.33	49.07	14.95	1.40	10.89	5.14	39.43	31.43	38.86	42.29	48.00	18.86	1.14
Mild	39.27	4.21	21.96	48.13	35.98	9.81	18.22	3.74	0.47	29.23	4.00	19.43	44.00	30.29	22.29	18.86	2.29	0.00
None	43.75	92.06	7.01	21.50	2.80	3.27	2.34	1.40	1.40	58.17	89.14	6.86	16.57	5.71	0.57	2.86	0.00	1.14
N	848	214	214	214	214	214	214	214	214	*698*	*175*	*175*	*175*	*175*	*175*	*175*	*175*	*175*
Pain																		
Extreme	1.18	0.00	0.97	2.42	5.31	13.53	2.90	73.91		1.58	0.00	1.75	7.02	8.19	22.81	11.70	74.27	
Severe	4.95	0.48	10.63	37.68	38.65	55.07	32.85	13.53		6.45	2.34	18.13	26.32	49.71	56.14	38.01	21.64	
Moderate	15.21	12.08	57.49	34.78	39.61	21.74	35.27	3.86		11.03	14.62	51.46	44.44	40.35	18.71	32.75	2.34	
Mild	42.81	58.94	25.12	19.32	13.04	5.31	20.29	1.93		37.82	67.84	28.07	21.64	1.75	1.75	14.62	1.17	
None	35.85	28.50	5.80	5.80	3.38	4.35	8.70	6.76		43.12	15.20	0.58	0.58	0.00	0.58	2.92	0.58	
N	848	207	207	207	207	207	207	207		*698*	*171*	*171*	*171*	*171*	*171*	*171*	*171*	
Self-care																		
Extreme	0.00	0.00	0.00	0.51	2.53	8.08	3.54	71.21		0.72	0.00	0.59	3.53	4.71	15.88	7.65	60.00	
Severe	1.53	1.01	2.53	5.56	29.29	60.61	33.33	16.67		1.00	0.00	13.53	9.41	35.88	51.18	33.53	29.41	
Moderate	4.48	1.01	28.79	37.37	50.00	20.71	44.95	2.53		2.58	7.65	52.94	36.47	45.88	29.41	45.88	8.24	
Mild	6.84	2.02	56.57	44.44	11.11	3.03	11.11	1.01		3.15	8.82	28.24	41.18	12.94	1.18	11.76	1.18	
None	87.15	95.96	12.12	12.12	7.07	7.58	7.07	8.59		92.55	83.53	4.71	9.41	0.59	2.35	1.18	1.18	
N	848	198	198	198	198	198	198	198		*698*	*170*	*170*	*170*	*170*	*170*	*170*	*170*	
Usual																		
Extreme	0.94	1.01	5.05	1.01	2.53	11.62	2.02	13.13	*61.62*	2.01	0.00	2.87	1.15	4.02	14.94	5.75	12.07	48.28
Severe	2.95	2.02	29.29	7.58	23.23	52.53	19.19	65.15	*33.84*	3.44	0.57	17.82	13.79	28.74	58.05	31.03	68.97	41.38
Moderate	10.38	2.53	48.48	37.88	51.01	26.26	47.47	18.18	*2.53*	6.73	9.20	45.98	30.46	44.83	17.82	41.95	16.09	6.32
Mild	14.98	7.07	15.66	48.99	17.68	9.09	30.81	3.03	*0.00*	12.61	14.37	29.89	50.00	20.11	6.32	18.39	2.87	3.45
None	70.75	87.37	1.52	4.55	5.56	0.51	0.51	0.51	*2.02*	75.21	75.86	3.45	4.60	2.30	2.87	2.87	0.00	0.57
N	848	198	198	198	198	198	198	198	198	*698*	*174*	*174*	*174*	*174*	*174*	*174*	*174*	*174*
Affect																		
Extreme	2.36	0.49	0.00	5.34	3.40	43.20	78.64			2.29	0.00	0.00	10.80	3.98	41.48	78.98		
Severe	5.07	0.49	2.43	43.69	37.38	48.06	18.45			7.02	0.00	5.68	43.75	28.41	45.45	16.48		
Moderate	22.76	0.49	24.76	45.15	53.88	5.83	1.94			19.63	5.11	28.51	31.82	43.18	10.23	3.98		
Mild	38.92	7.28	68.93	5.83	4.85	1.94	0.00			35.39	12.50	59.09	13.07	22.16	1.70	0.00		
None	30.90	91.26	3.88	0.00	0.49	0.97	0.97			35.67	82.39	6.82	0.57	2.27	1.14	0.57		
N	848	206	206	206	206	206	206			*698*	*176*	*176*	*176*	*176*	*176*	*176*		

Table AIII. Codebook page for income variable

Name	Label/Description	Variable			Values	Value labels
		Type	Class	Scale		
aboutu6	income brackets	Numeric	Categorical	Interval	1	1st quintile
					2	2nd quintile
					3	3rd quintile
					4	4th quintile
					5	5th quintile
					7	Refuse
					8	DK

APPENDIX B. Vignette descriptions

MOBILITY

1 - [Paul] is an active athlete who runs long distance races of 20 kilometres twice a week and engages in soccer with no problems.

2 - [Mary] has no problems with moving around or using her hands, arms and legs. She jogs 4 kilometres twice a week without any problems.

3 - [Rob] is able to walk distances of up to 200 metres without any problems but feels breathless after walking one kilometre or climbing up more than one flight of stairs. He has no problems with day-today physical activities, such as carrying food from the market.

4 - [Margaret] feels chest pain and gets breathless after walking distances of up to 200 metres, but is able to do so without assistance. Bending and lifting objects such as groceries produces pain.

5 - [Louis] is able to move his arms and legs, but requires assistance in standing up from a chair or walking around the house. Any bending is painful and lifting is impossible.

6 - [David] is paralysed from the neck down. He is confined to bed and must be fed and bathed by somebody else.

COGNITION

1 - [Rob] can do complex mathematical problems in his mind. 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.

2 - [Sue] can only count money and bring back the correct change after shopping. Mental arithmetic is otherwise a problem. She can find her way around the neighbourhood and know where her own belongings are kept.

3 - [Henriette] can pay attention to the task at hand for periods of up to one hour, with occasional distractions and can quickly return to the task. She can remember names of

people she meets often, their addresses and important numbers, but occasionally has to remind herself of the names of distant relatives or acquaintances.

4 - [Helena] can remember details of events that have taken place or names of people she has met many years ago, She can do everyday calculations in her mind. During periods of anxiety lasting a few hours, she becomes confused and cannot think very clearly.

5 - [Tom] finds it difficult to concentrate on reading newspaper articles, or watching television programmes. He is forgetful and once a week or so, he misplaces important things, such as keys or money, and spends a considerable amount of time looking for them, but is able to find them eventually.

6 - [Julian] is easily distracted, and within 10 minutes of beginning a task, his attention shifts to something else happening around him. He can remember important facts when he tries, but several times a week finds that he has to struggle to recollect what people have said or events that have taken place recently.

7 - [Christian] is very forgetful and often loses his way around places which are not very familiar. He needs to be prompted about names of close relatives and loses important things such as keys and money, as he cannot recollect where they have been kept. He has to make notes to remind himself to do even very important tasks.

8 - [Peter] does not recognize even close relatives and cannot be trusted to leave the house unaccompanied for fear of getting lost. Even when prompted, he shows no recollection of events or recognition of relatives.

PAIN

1 - [Laura] has a headache once a month that is relieved one hour after taking a pill. During the headache she can carry on with her day to day affairs.

2 - [Phil] has pain in the hip that causes discomfort while going to sleep. The pain is there throughout the day but does not stop him from walking around.

3 - [Patricia] has a headache once a week that is relieved 3–4 hours after taking a pill. During the headache she has to lie down, and cannot do any other tasks.

4 - [Mark] has joint pains that are present almost all the time. They are at their worst in the first half of the day. Taking medication reduces the pain though it does not go away completely. The pain makes moving around, holding and lifting things, quite uncomfortable.

5 - [Jim] has back pain that makes changes in body position very uncomfortable. He is unable to stand or sit for more than half an hour. Medicines decrease the pain a little, but it is there all the time and interferes with his ability to carry out even day to day tasks.

6 - [Tom] has a toothache for about 10 minutes, several times a day. The pain is so intense that Tom finds it difficult to concentrate on work.

7 - [Steve] has excruciating pain in the neck radiating to the arms that is very minimally relieved by any medicines or other treatment. The pain is sharp at all times and often wakes him from sleep. It has necessitated complete confinement to the bed and often makes him think of ending his life.

SELF-CARE

1 - [Helena] keeps herself neat and tidy. She requires no assistance with cleanliness, dressing and eating.

2 - [Anne] takes twice as long as others to put on and take off clothes, but needs no help with this. She is able to bathe and groom herself, though that requires effort and leads to reducing the frequency of bathing to half as often as before. She has no problems with feeding.

3 - [Paul] has no problems with cleanliness, dressing and eating. However, he has to wear clothes with special fasteners as joint problems prevent him from buttoning and unbuttoning clothes.

4 - [Peter] can wash his face and comb his hair, but cannot wash his whole body without help. He needs assistance with putting clothes on over his head, but can put garments on the lower half of his body. He has no problems with feeding.

5 - [John] cannot wash, groom or dress himself without personal help. He has no problems with feeding.

6 - [Rachel] feels pain and discomfort while washing, and in combing her hair. As a result, she neglects her personal appearance. She needs assistance with putting on and taking off clothes. She has no problems with feeding.

7 - [Sue] requires the constant help of a person to wash and groom herself and has to be dressed and fed.

USUAL

1 - [John] is a teacher and goes to work regularly. He teaches the senior grades and takes classes for 6 hours each day. He prepares lessons and corrects exam papers. Students come to him for advice.

2 - [Dan] is a mason in a building firm. Three to four times per week, he is noticed to leave his bricklaying tasks incomplete. With help and supervision, he is able to use his skills to finish the walls of the buildings well.

3 - [Mathew] is a clerk in the local government office. He maintains ledgers with no errors and keeps them up to date. However, he ends up not doing any work for a day once every 2 weeks or so because of a migraine headache.

4 - [Maria] is an accountant in the local bank. She is regularly at work. However, she makes minor errors in the accounts and tends to postpone tasks. She delays producing account statements and is late on deadlines.

5 - [Carol] is a housewife who leaves most chores around the house half done. Even with domestic help she cannot complete important tasks in time, such as getting her son ready for school. Her husband has had to take over the cooking.

6 - [Doris] is a housewife and does most of the cooking and cleaning around the house. About once a week she leaves tasks half done. Her cooking has deteriorated and the house is not as clean as it used to be. She also takes about twice as long to do the chores.

7 - [Karen] is a teacher and has had to miss work for 2 weeks in the past month. Even now she feels tired and exhausted, and cannot stand for long periods in the classroom. Colleagues notice that she is making serious mistakes in correcting answer papers.

8 - [Jack] is a clerk at the local post office. He just sits around all day and cannot engage in any work. He cannot sort letters, manage the counter or interact with customers. His employers are considering replacing him.

AFFECT

1 - [Ken] remains happy and cheerful almost all the time. He is very enthusiastic and enjoys life.

2 - [Henriette] remains happy and cheerful most of the time, but once a week feels worried about things at work. She gets depressed once a month and loses interest but is able to come out of this mood within a few hours.

3 - [Jan] feels nervous and anxious. He is depressed nearly every day for 3–4 hours thinking negatively about the future, but feels better in the company of people or when doing something that really interests him.

4 - [Eva] feels worried all the time about things at work and home, and feels that they will go wrong. She gets depressed once a week for a day, thinking negatively about the future, but is able to come out of this mood within a few hours.

5 - [John] feels tense and on edge all the time. He is depressed nearly everyday and feels hopeless. He also has a low self esteem, is unable to enjoy life, and feels that he has become a burden.

6 - [Roberta] feels depressed all the time, weeps frequently and feels completely hopeless. She feels she has become a burden, feels it is better to be dead than alive, and often plans suicide.

REFERENCES

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

Bago d'Uva, T., O'Donnell, O., & van Doorslaer, E. (2008). Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology, 37*(6), 1375-1383.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics, 17*(3), 351-375.

Baker, M., Stabile, M., & Deri, C. (2004). What Do Self-Reported, Objective, Measures of Health Measure? *The Journal of Human Resources, 39*(4), 1067-1093.

Crossley, T. F., & Kennedy, S. (2002). The reliability of self-assessed health status. *Journal of Health Economics, 21*(4), 643-658.

Currie, & Madrian. (1999). Chapter 50 Health, health insurance and the labor market. *Handbook of Labor Economics, 3*, 3309-3416.

Datta Gupta, N., Kristensen, N., & Pozzoli, D. (2010). External validation of the use of vignettes in cross-country health studies. Economic Modelling, 27(4), 854-865.

Doiron, D., Fiebig, D. G., Johar, M., & Suziedelyte, A. (2015). Does self-assessed health measure health? *Applied Economics, 47*(2), 180-194.

Dowd, J. B., & Todd, M. (2011). Does self-reported health bias the measurement of health inequalities in U.S. adults? evidence using anchoring vignettes from the health and retirement study. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 66B*(4), 478-489.

Etilé, F., & Milcent, C. (2006). Income-related reporting heterogeneity in self-assessed health: Evidence from france. *Health Economics, 15*(9), 965-981.

Hernandez-Quevedo C, Jones AM, Rice N. 2004. Reporting bias and heterogeneity in self-assessed health. Evidence from the British Household Panel Survey. ECuity III Working Papers, York. 2004

Hirve, S., Verdes, E., Lele, P., Juvekar, S., Blomstedt, Y., Tollman, S., . . . Ng, N. (2014). Evaluating reporting heterogeneity in self-rated health among adults aged 50 years and above in india: An anchoring vignettes analytic approach. *Journal of Aging and Health, 26*(6), 1015-1031.

Idler, E., & Benyamini, Y. (1997). Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. *Journal of Health and Social Behavior, 38*(1), 21-37.

Jones, A., & O'Donnell, Owen. (n.d.). *Econometric analysis of health data*. Chichester: Wiley.

Jones, A. (n.d.). *Applied health economics* (2nd ed., Routledge advanced texts in economics and finance; 19). London; New York: Routledge.

Kerkhofs, M., & Lindeboom, M. (1995). Subjective health measures and state dependent reporting errors. *Health Economics, 4*(3), 221-235.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191-207.

Kreider, B. (1999). Latent Work Disability and Reporting Bias. *Journal of Human Resources*, *34*(4), 734-769.

Lindeboom, M., & van Doorslaer, E. (2004). Cut-point shift and index shift in self-reported health. *Journal of Health Economics, 23*(6), 1083-1099.

Marmot, M., Oldfield, Z., Clemens, S., Blake, M., Phelps, A., Nazroo, J., . . . Banks, J. (2014). English Longitudinal Study of Ageing: Waves 0-6, 1998–2013 [UK Data Archive]. Retrieved from http://dx.doi.org/ 10.5255/UKDA-SN-5050-8

Molina, T. (2016). Reporting heterogeneity and health disparities across gender and education levels: Evidence from four countries. *Demography, 53*(2), 295-323.

Parsons, D. O., & Bound, J. (1991). The health and earnings of rejected disability insurance applicants: Comment; reply. *The American Economic Review, 81*(5), 1419.

Peracchi, F., & Rossetti, C. (2012). Heterogeneity in health responses and anchoring vignettes. *Empirical Economics, 42*(2), 513-538.

Rice, N., Robone, S., & Smith, P. (2011). Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *The European Journal of Health Economics, 12*(2), 141-162.

Schneider, U., Pfarr, C., Schneider, B., & Ulrich, V. (2012). I feel good! gender differences and reporting heterogeneity in self-assessed health. *The European Journal of Health Economics, 13*(3), 251-265.

Shmueli, A. (2003). Socio-economic and demographic variation in health and in its measures: The issue of reporting heterogeneity. *Social Science and Medicine, 57*(1), 125-134.

Üstün TB et al. (2003). *Health Systems Performance Assessment: Debates, Methods and Empiricisms,* World Health Organization.

van Doorslaer, E., & Gerdtham, U. (2003). Does inequality in self-assessed health predict inequality in survival by income? Evidence from swedish data. *Social Science and Medicine, 57*(9), 1621-1629.

van Soest, Arthur, Delaney, Liam, Harmon, Colm, Kapteyn, Arie, & Smith, James P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. (Report). *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(3), 575.

Ziebarth, N. (2010). Measurement of health, health inequality, and reporting heterogeneity. *Social Science & Medicine, 71*(1), 116-124.