Inflated Responses in Measures of Self-Assessed Health[1]

William H. Greene

Stern Business School, New York University, USA

Mark N. Harris[2]

Curtin Business School, Curtin University, Australia

Bruce Hollingsworth

Division of Health Research, Lancaster University, UK

Abstract/Summary

This paper focuses on the self-reported responses given to survey questions of the form 'In general how would you rate your health?' with typical response items being on a scale ranging from poor to excellent. Usually, the overwhelming majority of responses fall in either the middle category or the one immediately to the "right" of this (for example, *good* and *very good*). However, based on a wide range of other medical indicators, such favorable responses appear to paint an overly rosy picture of true health. The hypothesis here is that these "middle" responses have been, in some sense, inflated. That is, for whatever reason, a significant number of responders inaccurately report into these categories. Our results do indeed suggest that such inflation is present in these categories. Adjusted responses to these questions could lead to significant changes in policy, and should be reflected upon when analysing and interpreting these scales.

Keywords: Self-assessed health, inflated outcomes, mis-reporting, ordered probit, panel data.
JEL codes: C3, C5, I1.

---

[1] We are grateful to the Australian Research Council who helped fund this research. We are also grateful to Michael Shields, David Johnson, Nigel Rice, John Wildman and Andrew Jones for useful comments and suggestions, and also to seminar participants at the National Institute for Labour Studies (University of Adelaide), Curtin University, University of Western Australia, the University of Queensland, and participants at the 2012 Annual Health Econometrics Workshop. This paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the author and should not be attributed to either FaHCSIA or the Melbourne Institute.
[2] Corresponding author; mark.harris@curtin.edu.au; tel: +61 8 9266 9692

Introduction

The health sector is a critical part of the economy in developed countries, in most making up around 10% of GDP, and in some (for example, the USA) a great deal more. To determine if such resources are being utilised effectively, it is imperative for policy makers to fully understand the determinants of individuals' health levels so that expenditure can be more accurately targeted. In this regard, a very common health measure used in assessments of this nature is from survey data in the form of individual responses to questions of the form: *Overall, how would you rate your health?*. Here respondents typically tick one of 5 boxes ranging from *very bad* through to *average* to *very good* – this is commonly referred to as "the 5-point scale". Heavy reliance on these measures is based on the fact that they are relatively cheap and easy to collect and are included in a range of health and other surveys worldwide. Moreover, they have been shown to be useful - especially when more objective measures are not available; for example, they are good predictors of other health outcomes such as mortality (Bound, 1991, Burström et al, 2001, Mossey*et al.* 1982).

While there is some acceptance that there is potentially measurement error "at the margin" in such survey instruments (Currie and Madrian 1999; Crossley and Kennedy 2002), researchers and policy-makers typically take these responses as generally a true reflection of the individual's (and in aggregate, the population's) actual health status. However, a closer inspection of these *Likert*-scale responses, irrespective of the exact question wording, or response item labelling, reveal that the bulk of the observations (across survey, country and time), invariably correspond to the *good* and *very good* outcomes (or the middle response, and the one immediately "to the right" of this). An obvious question, therefore, is: do these favourable numbers really represent the true health of the nation when other, more objective measures of health, paint a much different picture?

Taking Australia as a typical example, and using the large, nationally representative, widely used and cited, panel data survey, the Household and Income Labour Dynamics of Australia (HILDA) – the data used in our empirical example – we find that around 75% of responses to the self-assessed health question fall in the *good* and *very good* categories.[3] However, from the Australian Diabetes, Obesity and Lifestyle Study (Dunstan *et al.*, 2002), for clinically measured conditions, we find that 60% of Australians are either mildly overweight or obese, and "Almost 1 in 4 Australians…has either diabetes or a condition of impaired glucose

[3]A figure remarkable stable over time.

metabolism" which is "associated with substantially increased immediate risk of heart disease as well as increased risk of diabetes in the future" (p.1). Half of the population have elevated cholesterol; over one in two Australians has at least one of the "deadly quartet" of health conditions; and less than half of the population undertakes enough physical activity sufficient for good health. Such statistics are clearly at odds with self-assessed health (SAH) measures that suggest Australia has a healthy population. Similar figures exhibiting an apparent disparity between the prevalence of clinically (or otherwise) measured conditions and self-reported health levels can be found across the developed world.

The aim of this paper is to examine whether there is indeed inflating of these outcomes in SAH measures. To test this hypothesis, a new econometric model is proposed, based on an ordered probit framework, which explicitly allows for inflation in these two outcomes. The results suggest that estimated (prior) probabilities of inaccurate reporting are non-trivial (at just under 10%); moreover this rises to over 12% based on the posterior probability (conditional on being in one of these two categories). Of the overall (marginal) estimated probability of being in these two categories (40% for *good*; 38%, *very good*) we estimate that this is comprised of 5 and 4 percentage points, respectively, arising from simply inaccurate reporting. In other words, for some individuals their responses may not reflect true *or* self-reported health, but this inaccuracy may be down simply to a 'box-ticking' strategy.

Overall therefore, these findings suggest that a strong reliance should perhaps not be placed on these typically highly favoured responses in health, and related, surveys without questioning the potential reasons for such responses, which are discussed further in the conclusions.

The current paper thus offers a significant contribution to the health economics' literature by identifying the potential for inflation in measures of self-assessed health, and also in suggesting a strategy and econometric approach to quantify the extent of such. However, this approach (and model) is likely to have a wide-ranging appeal across a host of health-related survey instruments that are based on similar *Likert*-scales and therefore similarly likely to be affected by such a "box-ticking" strategy. Indeed, the use and analysis of such scales is omniscient not just across the health sciences, but also right across the social sciences (and wider); and (as we show here) if such inflation is present in the data and not accounted for, biased and inconsistent parameter estimates may result, yielding potentially erroneous advice for policy-makers.

Background

We focus in this paper on the issue of mis-reporting in SAH. However, the validity and accuracy of survey data in general have long been of concern to researchers because of various potential errors in measurement. The early literature in this area focussed on broad potential general sources of error in reported responses: *acquiescence*—that responses such as *True, Yes,* and *Agree* are preferred; *evasiveness*—responses such as *Indifferent* and *Uncertain* are preferred, and; *extreme-response styles*— such that respondents tend to choose responses on higher or lower regions of rating scales(Cronbach, 1950).Cognitive theories went on to postulate that there might be further *primacy* and *recency* effects, respectively, when response items are presented visually and orally (Krosnick and Alwin, 1987). Primacy effects are present when respondents exhibit a tendency to select the response items presented to them first, as a result of them having spent more time in cognitive processing of these items. Recency effects, on the other hand, explain how respondents might select the response items read out to them last, because processing of earlier response items are quickly terminated by processing of later items. On a related issue, Wildt and Mazis (1978) found that both the label and the location of a response item on the response scale may have an effect on respondents.

While this earlier literature provided broader categories of mis-reporting, more recent studies have examined how these potential issues could have manifested in various other contexts. For example, while *acquiescence* is a possible source of mis-reporting, social desirability may similarly be a factor influencing responses. Adams *et al.* (2005) looked at the relationship between self-reported physical activity and the desire for social desirability. Using two measures of physical activity—one objectively measured and the other self-reported—as well as information on personality traits for social desirability, they found that over-reporting of physical activity (the difference between the objective and self-reported measures) was significantly associated with social desirability.

Mis-reporting, possibly to make oneself appear more "socially-acceptable", has also been found in other studies. Ezzati*et al.* (2006), for example, examined the difference between measured and self-reported height and weight, and found that women under-reported their weight but men did not; also, younger men over-reported their height more than women of the same age did. Hebert *et al.* (2002) also found that self-reports of diet were influenced by both the tendency to keep with cultural norms and the desire to obtain a positive response in

testing situations. The issue of *evasiveness* has also been further explored. Böckenholt*et al.* (2009) considered the method of randomized responses, which is used to try and obtain honest answers to sensitive issues on the assumption that the randomization eliminates response bias. Respondents are told that the surveyor has no way of knowing if they are answering questions on sensitive issues or not, since they may be randomly allocated to have answered other questions instead. On aggregate, however, it is possible to estimate percentage of positive responses to the sensitive questions. They find that some respondents may not follow the randomization scheme because they distrust it; others may follow it until sensitive questions are asked, which shows that the method of randomization may not have corrected response bias as much as hoped for.

On the issue of *extreme-response* reporting style, Arce-Ferrer (2006) found that participants who are less familiar with rating scales may have higher tendency to report on extreme ends of the scale. Language and culture may affect desirability of demonstrating high levels of language precision, thereby affecting the tendency for extreme reporting and centre-scale reporting.

These general issues of mis-reporting in the context of survey data are still of critical importance in health surveys, and one area which has received particular attention in the health economics literature in terms of how to deal with these issues empirically is general mis-reporting across SAH categories (see, for example, Jones *et al.*, 2010). This has usually taken the form of applying Generalised Ordered Probit (GOP) models (first suggested by Pudney and Shields, 2000).[4] Kerkoffs and Lindeboom (1995) and Jones and Schurer (2011) both suggest variants of the GOP model. In such models the boundary parameters embedded in the ordered choice model are functions of observed personal characteristics (although we note that the strict approach of Pudney and Shields, 2000, is just one of many ways in which heterogeneity can be introduced into the boundary parameters; see Greene and Hensher, 2010). Jones and Schurer (2011) and Carro and Trafferi (2012) both develop elaborate models of heterogeneity involving a correlated random effect in the broader health satisfaction index model and a conventional fixed effect in the inherent boundaries of the model. These studies examine heterogeneity in broad terms, rather than as a symptom of 'mis-reporting.'

[4] We use the GOP model interchangeably with the very similar Hierarchical Ordered Probit (HOPIT) model (see, for example, Greene and Hensher, 2010).

Many studies have attempted to test more narrowly for the presence of mis-reporting in health by comparing to more objective measures of health. Baker *et al.* (2004) compared self-reported presence of medical conditions against medical records to test for the presence of measurement errors, and found that measurement error was associated with absence from the labour market: being in the labour market, for example, decreases the chances of false positive reporting of migraines by 48 percent. Butler *et al.* (1987) also find that while correlation was high between self-reported measures of arthritis and a simulated clinical measure of it, work status affected measurement error significantly such that individuals who were not working were more likely than those who were to have measurement errors in their self-reports of arthritis. Similarly, Klesges *et al.* (1995) compared distribution of self-reports of smoking against an objective measure of smoking exposure, and found that heavier smokers, Caucasians and people with less education tended to display digit preferencing (a bias towards integers).

Kerkoffs and Lindeboom (1995) in a Dutch panel study on retirement and aging approach the mis-reporting issue from an institutional perspective. They reason that due to the availability of certain benefits, individuals in a specific few groups – employed, unemployed, disabled and early retired – will have different incentives to mis-report their health status. Their analysis of subjective reported health employs a companion, objective measure of health based on an inventory of numerous mental and physical symptoms. The objective measure enters the ordered choice model with other covariates while the different status groups, with a different set of individual attributes such as education, introduce heterogeneity into the thresholds that determine the cell probabilities. The motivation behind their specification resembles ours. However, they treat the mis-reporting issue more generically than we do – indeed, the specification is consistent with a more expansive definition of heterogeneity across the four groups.

Researchers have also recently experimented with biomarkers as objective indicators of health. Dowd and Zajacova (2010) tested whether respondents with higher levels of education also had healthier levels of biomarkers for the same level of self-assessed health. They found that among respondents of the same level of SAH, those who were better educated had healthier levels of biomarkers compared to those who were less educated. This was true also for biomarkers that were not regularly tested by physicians, and suggested that differences in health expectations rather than health knowledge may be the more probable explanation.

Clearly, a superficially obvious means of accounting for mis-reporting is to substitute objective measures of health for self-assessed (subjective) ones. However, these are also not free from reporting error and moreover, are not always available (Bound *et al.*, 1991). In fact, as Disney *et al.* (2006) point out, there may in fact be a loss of information about the "true" relationship between a more subjective measure of health and behaviour (replacing an error in variables problem with a similar problem, just with a proxy variable). Thus our focus remains on self-assessed measures, and in particular SAH, as these are most frequently used and available in practice.

As touched upon above, various econometric models have been employed in the health economics (and other) literature(s) to account for mis-reporting. The usual ordered probit estimates a latent health index, as well as the cut-off points beyond which the latent health translates into the observed SAH responses, and is commonly used as the base model (Lindeboom and van Doorslaer, 2004). Several authors have used the Pudney and Shields' (2000) approach to extend the ordered probit model by allowing the cut-off points to be determined by observable characteristics. As the cut-off points have an influence on SAH independently of (true) latent health, the model accounts for mis-reporting by these observable characteristics.

The use of vignettes has also been suggested as fertile ground for new research. Bago d'Uva *et al.* (2006) provide an example of how vignettes (questions asked to respondents about what level of SAH they think a person, under hypothetical scenarios, is in) can be used to model mis-reporting. Individual characteristics are assumed to affect the cut-off points equally in both the vignette model and the model for respondent health. Assuming response consistency (individuals classify their health the same way as they classify the hypothetical cases) and vignette consistency (that vignettes are perceived by all respondents on the same uni-dimensional scale) the effect of individual characteristics on the cut-off points can be identified. Jones, Rice and Rabone (2012) also examine the use of vignettes as a specific treatment for cross country heterogeneity. However, although a complimentary approach to the one adopted in the current paper, analysis using vignettes is not pursued here: for one, as with almost all other similar surveys, vignettes are not available in the data we have to hand.

The issue we consider here bears some superficial connection to the familiar problem of ordinary measurement error in the dependent variable in a regression. However, in practical terms, the fact that our observed response is discrete makes nearly all of the received results

on this subject irrelevant. There is a sparse literature on measurement error of sorts in discrete response, including Hausman, Scott-Morton and Abrevaya (1995) who studied misclassification in binary choices, and Winkelmann (1996) who examined underreporting of counts. Arguably, our inflated response data are not actually mismeasured. The individual is reporting their preferred answer to the survey question; in light of this, we use the terms *inaccurate* reporting and *mis-reporting*, interchangeably from here on in, although believe that the former is probably appropriate. Indeed, even the term inaccurate might be "incorrect" if the individual responds to the question as accurately as they see fit in describing their health at *that particular point in time*. However, of key importance here, is whether these responses accurately reflect "true" health, as identified by more objective measures? Therefore, at issue in this study is how these answers should be interpreted. We find that a behavioural interpretation couched in terms of a latent class model, as we shall see below, is an appropriate way to proceed.

What motivation would individuals have to mis-report SAH? Kerkoffs and Lindeboom (1995) surmised that certain institutional features provided an incentive to inaccurately report SAH. However, the treatment considered in this paper is more focused on the health outcome, itself: a simple comparison of the distribution of the SAH responses against more clinically measured health outcomes, clearly points to these "middle", and "to right of middle" outcomes being over-inflated. Therefore an obvious question is why should this be the case? That is, what are some of the potential motivations for inaccurate reporting here?

Firstly, digit/item preferencing: this issue has been addressed by, for example, Fry and Harris (2005), with regard to inflationary expectations and student course evaluations. The latter relates strongly to the current paper, as it relates to students "ticking a box" on a *Likert*-scale for satisfaction levels. Without paying too much attention to the question at hand, respondents avoid the extreme responses and opt for the defensible option of somewhere "in the middle (or average)" or just a bit better than average.

Secondly, adaptation: individual's valuation of their own health status changes over time – even if their objective illness levels remain the same. This is often reflected by higher valuations the longer an individual has a certain condition as they "adapt" to the condition in terms of lifestyle changes. However, it is hard to generalise that adaptation causes inaccurate reporting as individuals may move in and out of illness, and the adaptation processes will be

different for different individuals. So, we list this as a potential motivation noting it is potentially difficult to measure (Hauck and Hollingsworth, 2011).

Thirdly, as noted above, there is a significant amount of marketing and related literature relating to respondents wanting to please the interviewer and to avoid giving what might be viewed as a socially unacceptable answer (see, for example, Worcester and Burns, 1975). The issue of not having a neutral, or mid-point, on a *Likert*-scale was considered by Garland (1991) who found that in doing-so, minimised the social desirability bias by respondents wishing to "please" the interviewer. Moreover, there is also a significant amount of literature regarding the number of scale steps in the response answers and respondents' (over-)use of the midpoint category. Matell and Jacoby (1972) for instance, find that with three and five point scale formats, about 20% of respondents choose the mid-point, whereas this falls to 7% when these were increased to seven and above. There is also evidence that grammatically balanced *Likert*-scales are unbalanced in their interpretations: 'tend to disagree' is (often) not directly opposite to 'tend to agree' (Worcester and Burns, 1975).

Finally, there is a simple cost-of-time argument: individuals who value their time more highly are likely to pay less detailed attention to the question at hand, and answer "quickly and easily" by opting for responses in the middle of *Likert*-scales. Research typically finds that the opportunity cost of time is positively related to income, employment and wages (for example, Prochaska and Schrimper, 1973, and Mormorstein *et al.*, 1992) such that we would expect to see this reflected in our empirical findings with regard to our inaccurate reporting equation.

The use of these scales is evident in almost all health surveys. They are easy to collect, but this cannot lead to the assumption that they are easy to make use of. Health economists continue to use SAH measures in empirical analyses. Yet, there is convincing evidence that such SAH responses have been "inflated" and a number of possible reasons for this have been identified. This phenomenon is heavily related to the existing literature on measurement error and mis- and inaccurate, reporting. Such a specific form of such inaccurate reporting has not, to date, been addressed in either the health or econometrics literature, so a new approach is proposed here to test this hypothesis.

Empirical Approach

The existing literature (for example, Contoyannis*et al.*, 2004) provides an excellent starting point for both the techniques and appropriate variables to use in developing a model of an individual's SAH. As almost all measures of SAH are elicited from survey responses on a *Likert*-type response scale, invariably ordered probability models (logits and probits) form the basis of most empirical analyses, as the data are both discrete and ordered (see Greene and Hensher, 2010, for a summary of ordered choice modelling). The ordered probit (OP) model is usually justified on the basis of an underlying latent variable, y^* which is a linear (in unknown parameters, β_y) function of: observed characteristics, x_y; a (standard normal) disturbance term, ε_y; and its relationship to certain boundary parameters, μ. Thus with

$$y^* = x_y'\beta_y + \varepsilon_y, \tag{1}$$

and where the mapping between the latent and observed components is assumed to be given by

$$y = \begin{cases} 0 & if \quad y^* \le \mu_0, \\ 1 & if \quad \mu_0 < y^* \le \mu_1, \\ 2 & if \quad \mu_1 < y^* \le \mu_2, \\ 3 & if \quad \mu_2 < y^* \le \mu_3, \\ 4 & if \quad \mu_3 < y^*. \end{cases} \tag{2}$$

The usual OP (Greene and Hensher, 2010) probabilities result:

$$\Pr(y) = \begin{cases} 0 & = & \Phi(-x_y'\beta_y), \\ 1 & = & \Phi(\mu_1 - x_y'\beta_y) - \Phi(-x_y'\beta_y), \\ 2 & = & \Phi(\mu_2 - x_y'\beta_y) - \Phi(\mu_1 - x_y'\beta_y), \\ 3 & = & \Phi(\mu_3 - x_y'\beta_y) - \Phi(\mu_2 - x_y'\beta_y), \\ 4 & = & 1 - \Phi(\mu_3 - x_y'\beta_y), \end{cases} \tag{3}$$

with $\Phi(.)$ representing the standard normal distribution function, and with the normalisation that $\mu_0 = 0$.[5]

So here the latent variable y^* represents an individual's underlying health level, and "observed" health is how the individual actually responds to the appropriate survey question. In the survey data we use: $y = 0$ indicates *Poor*; $y = 1$, *Fair*; $y = 2$, *Good*; $y = 3$, *Very good*; and $y = 4$, *Excellent*. This set-up is akin to a Grossman health production function, whereby an individual's health outcomes are determined by a range of health inputs (x_y).

[5] The extension to more than five outcomes is direct.

However, our key hypothesis is that, possibly for the reasons noted above, the outcomes corresponding to *my health is good* and *my health is very good*, are an over-representation of a population's true health status.(In general, however, we would expect inflation in the outcomes corresponding to y = 2 and 3 on a five-point *Likert* scale, no matter the response item labels.) We refer to these two choice outcomes as the "middletons". The OP framework, as it stands above, cannot accommodate this phenomenon, or moreover, test this hypothesis.

Consider another latent variable, r^*, which represents an individual's propensity to report accurately/inaccurately. Let this latent variable be a function of a set of covariates, x_r, with unknown weights β_r, and a (standard normal) disturbance term, ε_r. Again, assuming linearity, we write

$$r^* = x_r'\beta_r + \varepsilon_r. \tag{4}$$

When this index reaches a critical level (normalised to zero), the individual will accordingly report accurately (r = 1). The probability that an individual will report accurately is therefore a probit probability of the form

$$\Pr(accurate) = \Phi(x_r'\beta_r), \tag{5}$$

and, by symmetry, 1 minus this, for inaccurate reporting probabilities (r = 0). For individuals who report accurately (r^* > 0, r = 1) they choose freely from the full choice set (here, $j = 0, \cdots, 4$); this choice will accordingly be determined by the standard ordered probit equations given above.

On the other hand, for individuals who report inaccurately (r = 0), the inflation-hypothesis states that they are faced with the binary choice of *Good versus Very good* SAH, only. Let this choice be dictated by a further latent variable m^*, determined by an equation of the form

$$m^* = x_m'\beta_m + \varepsilon_m, \tag{6}$$

where x_m are covariates with unknown weights β_m, and ε_m a (standard normal) disturbance term. When this index reaches a threshold value, again normalised to zero, the inaccurate reporter will choose outcome y = 3 (*Very good*) with probability $\Phi(x_m'\beta_m)$, and outcome y = 2 (*Good*), with probability 1 - $\Phi(x_m'\beta_m)$.

Under independence of all of the stochastic elements of the system (we re-visit this assumption below), the joint probabilities of inaccurate reporting and *Good* and *Very good* outcomes, will therefore be

$$\begin{aligned}\Pr(\textit{inaccurate}, \textit{good}) &= \Phi(-x_r'\beta_r)\Phi(-x_m'\beta_m),\\ \Pr(\textit{inaccurate}, \textit{very good}) &= \Phi(-x_r'\beta_r)\Phi(x_m'\beta_m).\end{aligned} \tag{7}$$

And, for those accurate reporters and all choice probabilities will be:

$$\Pr(\textit{accurate}, y) = \begin{cases} 0 = \Phi(x_r'\beta_r)[\Phi(-x_y'\beta_y)], \\ 1 = \Phi(x_r'\beta_r)[\Phi(\mu_1 - x_y'\beta_y) - \Phi(-x_y'\beta_y)], \\ 2 = \Phi(x_r'\beta_r)[\Phi(\mu_2 - x_y'\beta_y) - \Phi(\mu_1 - x_y'\beta_y)], \\ 3 = \Phi(x_r'\beta_r)[\Phi(\mu_3 - x_y'\beta_y) - \Phi(\mu_2 - x_y'\beta_y)], \\ 4 = \Phi(x_r'\beta_r)[1 - \Phi(\mu_3 - x_y'\beta_y)]. \end{cases} \tag{8}$$

Marginal probabilities for the full choice set are simply the sum of the two components such that:

$$\Pr(y) = \begin{cases} 0 = \Phi(x_r'\beta_r)[\Phi(-x_y'\beta_y)], \\ 1 = \Phi(x_r'\beta_r)[\Phi(\mu_1 - x_y'\beta_y) - \Phi(-x_y'\beta_y)], \\ 2 = \Phi(x_r'\beta_r)[\Phi(\mu_2 - x_y'\beta_y) - \Phi(\mu_1 - x_y'\beta_y)] + \Phi(-x_r'\beta_r)\Phi(-x_m'\beta_m), \\ 3 = \Phi(x_r'\beta_r)[\Phi(\mu_3 - x_y'\beta_y) - \Phi(\mu_2 - x_y'\beta_y)] + \Phi(-x_r'\beta_r)\Phi(x_m'\beta_m), \\ 4 = \Phi(x_r'\beta_r)[1 - \Phi(\mu_3 - x_y'\beta_y)]. \end{cases} \tag{9}$$

This model now has the attributes that can test our hypothesis: the SAH categories of *Good* and *Very good* get this additional boost from the inaccurate reporters. Moreover, the extent to which probabilities of inaccurate reporting diverge from zero is a reflection of the strength by which we can accept or refute our hypothesis.

Although this approach does not explicitly allow for "deflation", to the extent that the probability of accurate reporting is less than 1, the probabilities of all of j = 0, 1 and 4, will all be lower than they would have been otherwise (all other things equal). A measure of from what categories has this inflation into the middletons come from, would appear to be the marginal probabilities of equation (9) "purged" of any mis-reporting effects: to this extent one could consider the outcome probabilities, conditional on being an inaccurate reporter, relative to marginal probabilities and/or sample proportions.[6]

[6] An anonymous referee has suggested that a minority of respondent may inflate into categories other than the hypothesised middletons. Clearly we cannot rule this possibility out, although, in aggregate, it appears at odds with the empirical distribution of SAH measures. For this reason we would expect that such inflation, if present, to be a relatively minor magnitude such it would not adversely affect our findings.

We note that it would also be possible to simultaneously entertain inflation models of inaccurate reporting in conjunction with the more usual approach of GOP models for reporting heterogeneity "at the margin". Here we would simply let the inherent boundary parameters of the model be a function of observed covariates. However, unfortunately in our case, there would be a paucity of identifying variables for the boundary equations.[7]

The latent variables r^* and y^* and also r^* and m^* though, relate to the same individuals (although not so the m^* and y^*, as they relate to distinct sub-groups of the population: the accurate, and inaccurate, reporters). Therefore, it is likely that the unobservables across these equations will be related, with respective correlation coefficients ρ_{ry} and ρ_{rm}. This refinement now yields marginal probabilities that are functions of bivariate normal distributions, such that

$$\Pr(y)=\begin{cases} 0 &= \Phi_2(x_r'\beta_r,-x_y'\beta_y;-\rho_{ry}),\\ 1 &= \Phi_2(x_r'\beta_r,\mu_1-x_y'\beta_y;-\rho_{ry})-\Phi_2(x_r'\beta_r,-x_y'\beta_y;-\rho_{ry}),\\ 2 &= [\Phi_2(x_r'\beta_r,\mu_2-x_y'\beta_y;-\rho_{ry})-\Phi_2(x_r'\beta_r,\mu_1-x_y'\beta_y;-\rho_{ry})]+\Phi_2(-x_r'\beta_r,-x_m'\beta_m;\rho_{rm}),\\ 3 &= [\Phi_2(x_r'\beta_r,\mu_3-x_y'\beta_y;-\rho_{ry})-\Phi_2(x_r'\beta_r,\mu_2-x_y'\beta_y;-\rho_{ry})]+\Phi_2(-x_r'\beta_r,x_m'\beta_m;-\rho_{rm}),\\ 4 &= \Phi_2(x_r'\beta_r,\mu_3-x_y'\beta_y;-\rho_{ry})]. \end{cases} \quad (10)$$

Once the form of the probabilities and which outcome was chosen are both known, estimation is then undertaken using maximum likelihood techniques (see, for example, Greene, 2012). The contribution to the log-likelihood function for an individual would be

$$l(\theta)=\sum_{j=0}^{4} d_{ij}\ln\left[\Pr(y_i=j\mid x_i)\right],\ d_{ij}=1[y_i=j], \quad (11)$$

where θ contains all of the parameters of the model to be estimated.

Finally, as is common with survey data on SAH, the data we have in hand are *panel data*: that is, we have repeated observations on individuals over time. It is possible to condition on individual unobserved heterogeneity (see, for example, Mátyás and Sevestre, 2008): that is, there will clearly by unobserved heterogeneity driving, in part, all our three latent equations.

To account for this, we augment each *r*, *m* and *y* equation with an unobserved random effect ($\alpha_{r,m,y}$); the latent equations now become

[7] Again this point was raised by an anonymous referee. We note here that the identification issues noted could possibly be resolved with use of anchoring vignettes, and we leave this as an interesting line of future research.

$$\begin{pmatrix} r* \\ m* \\ y* \end{pmatrix} = \begin{pmatrix} x_r'\beta_r + \varepsilon_r \\ x_m'\beta_m + \varepsilon_m \\ x_y'\beta_y + \varepsilon_y \end{pmatrix} + \begin{pmatrix} \alpha_r \\ \alpha_m \\ \alpha_y \end{pmatrix}. \tag{12}$$

We allow for correlations across the *r* and *y*, and *r* and *m* equations, but not across the *m* and *y* ones, for reasons already noted, such that[8]

$$\begin{pmatrix} \alpha_r \\ \alpha_m \\ \alpha_y \end{pmatrix} \sim MVN(0,\Omega); \quad \Omega = \begin{pmatrix} \sigma_r^2 & \sigma_{rm} & \sigma_{ry} \\ \sigma_{rm} & \sigma_m^2 & 0 \\ \sigma_{ry} & 0 & \sigma_y^2 \end{pmatrix}. \tag{13}$$

The presence of the random effects in the likelihood function significantly complicates estimation. The method chosen here to integrate these unobserved effects out of the likelihood function is to use maximum simulated likelihood techniques, using three Halton-draw sequences each of length 100. The simulated log likelihood function is

$$\begin{aligned} l_s\left(\theta^*\right) &= \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} P_{itr}^*, \\ P_{itr}^* &= f\left(\theta, d_{ijt}, x_{r,m,y}, \beta_{r,m,y}, \mu, \Omega\right), \end{aligned} \tag{14}$$

where P_{itr}^* is the probability corresponding to the observed choice (for individual *i*, in time period *t*) conditional on all unobserved effects in the model.

For the present, we assume that all covariates are independent of all stochastic elements of the system. Moreover, the very few instances of multiple observations per individual in our estimation sample (see footnote 5), means that it is not possible to follow the Mundlak (1978) approach to freely correlated covariates and unobserved effects. However, we do address some potential endogeneity concerns, which are discussed in the data sections below.

A behavioural interpretation couched in terms of a latent class setting (see, for example, Greene and Hensher, 2010, with regard to latent class, or finite-mixture, models and ordered choice models) appears to be appropriate here. That is, clearly the *Good* and *Very good* observations arising from both accurate and inaccurate reporters are observationally equivalent. Therefore we can only probabilistically allocate observations to each regime, or unobserved (latent) class, on the basis of the hypothesised model described above. Thus the

[8] We note that in the empirical application, due to the range of questions used in the analysis (see below), very few individuals were actually observed more than once; due to this, it was difficult to identify the correlations across the three sets of unobserved effects, and they were therefore set equal to zero. In a sense, the resulting empirical model could be thought more of as a random parameters (constants) one, as opposed to a more traditional panel data one.

latent class interpretation of our approach would be that there are two unobserved classes; these are identified on the basis of the *Good* and *Very good* responses only and moreover on the data chosen to enter this equation determining this split.[9] Thus, clearly, the choice of appropriate data to identify this equation will be extremely important, and we return to this in the follow section(s).

Following on with the latent class analogy, *ex post*, numerous quantities may be of interest to the researcher and/or policymaker. In terms of probabilities, we can consider various posterior, or conditional on the data, probabilities of "class" membership (see, again with particular reference to posterior probabilities in an ordered choice latent class set-up, Greene and Hensher, 2010). For example, conditional on the individual choosing one of the hypothesised inflated outcomes, we can estimate posterior probabilities that an individual is an inaccurate reporter. Or, given that the individual chooses say *Good* what is the posterior probability that he/she is jointly an inaccurate reporter and chooses this outcome. In this way, we can effectively (probabilistically) allocate individuals choosing one of these two outcomes, into either being accurate or inaccurate reporters; (Note that by construction, if the individual reports *poor, fair,* or *excellent,* they must be an accurate reporter.) We can also consider prior, that is unconditional on the observed choice: probabilities of inaccurate reporting (or "class" membership); overall probabilities of each outcome; or that probability split down into its component parts (as detailed above in the equations deriving the various probability components).

Finally, we can also consider partial effects of explanatory variables on the various probabilities noted above, corresponding to various parts of the overall model. For example, we surmise that partial effects on the conditional, posterior probability of inaccurate reporting will be of great importance to policymakers as this will help to identify those individuals more likely to inaccurately report. Partial effects purged of any mis-reporting/inaccurate reporting effects, especially with regard to important policy tools, on health outcomes are also extremely likely to be of interest to policy makers. That is, ignoring such inaccurate reporting, a variable might appear to have a (superficially) advantageous effect on health

[9]It is important to remember, as noted earlier, that these do not necessarily suggest anything about the accuracy of the responses. The latter responses may well be accurate reflections of the individuals' *self*-assessed health. However, we contend in this paper, for the reasons outlined above, that this may not be a particularly good representation of their *true* health. For some individuals, these responses may be neither an accurate representation of their true health or even their self-assessed health, but merely the outcome of a "box-ticking" strategy".

outcomes and therefore be an obvious focus area for policymakers. However, in allowing for such inaccurate reporting, if it is found that this variable is essentially just driving observed outcomes though inaccurate reporting then clearly this will no longer the case.

Data and Empirical Model Specification

The data used are the longitudinal survey of Household and Income Labour Dynamics of Australia (HILDA). Waves 1 – 8 are used (yielding an initial sample of over 100,000 observations). The HILDA survey is ideal for our purposes as it is a panel data set containing a host of information on SAH and other health measures, as well as numerous demographic variables. (See Wooden and Watson, 2007, for more information on the HILDA survey in general.) Indeed, the HILDA survey, being heavily based on U.S. and British (and other) international counterparts (such as the PSID and the British Household Panel Survey) is a widely respected and heavily used data set.

For the econometric model described above, three sets of variables are required (x_r, x_m, x_y). Clearly the choice of the variables entering these is very important, especially with regard to specifying the r^* equation that will identify our inaccurate reporters. However, akin to a more standard latent class model, there are no identification issues in having $x_m \equiv x_y$; we have no strong priors for this not to be the case. Thus three sets of covariates are considered: *common*, which feature in all equations; *inaccurate-reporting*, these uniquely identify the inaccurate/accurate reporters; and *health*, which uniquely identify the health equations for both the accurate and inaccurate reporters.

Common Variables

Here a standard set of demographics are considered: ones that have been essentially used in previous empirical studies of health production functions such as Contoyannis *et al.*, 2004. Moreover, we have no strong priors about whether or not, nor in which direction, these variables are likely to affect the mis-reporting/inaccurate decision. These consist of quadratics in (standardised) age and household income; gender; migrant status; education; employment status; marital status; number of children; place of residence; and "seifada" (an index of relative socio-economic disadvantage).

Wave 5 of the HILDA survey also contains information on the so-called "big five" personality traits. These consist of ordered scales on agreeableness, conscientiousness,

emotional stability, extroversion and openness to experience.[10] There is a significant amount of literature suggesting correlations between health, including SAH, and these personality measures. The psychology literature (for example, Jorm*et al.*, 1993, Korotkov and Hannah, 2004, and Michel, 2006), suggests evidence of personality traits affecting perceptions of health, rather than underlying health. However, even in this literature there is evidence of personality traits affecting both subjective and objective measures of health (for example, Korotkovand and Hannah, 2004) such that these are included in all parts of the model, but based predominantly on the psychology literature, our priors would be that they will be strong predictors of the inaccurate reporting equation. A drawback with the use of these variables is that they were only recorded once (Wave 5); this has the result that we lose any individual who dropped-out of the survey before Wave 5. We also have to make the assumption, as noted in the HILDA documentation, that an individual's personality traits are constant over time.[11]

Variables to Identify the Inaccurate Reporters

Arguably the most important variable selection issues we need to address relate to the instruments for inaccurate reporting. Clearly these need to be strongly related to the mis-/inaccurate reporting decision, but be independent of the individual's true health status. We note here, that although we believe that we have been quite judicious in our choice of identifying variables (as argued below) in nearly all cases it is impossible to *completely* rule out any potential reverse causation between true health and them. However, we do also believe that any such reverse causation would be relatively minimal, especially as compared to the hypothesised direct affect we believe that they will have on reporting behaviours.[12]

With respect to finding potential variables to thus identify the inaccurate reporting equation, the HILDA survey contains some useful *interviewer* based responses on several aspects of the *interviewee's* completion of the survey. These include how suspicious (the interviewer believed) the respondent was about the study after completion; whether there was another adult present; and the respondent's general understanding of the questions. Clearly, all of

[10] See http://www.fahcsia.gov.au/about/publicationsarticles/research/austsocialpolicy/Documents/austsocpolicy8/art5.htm

[11] In Wave 9 the personality variables were collected again. However, these are not identical to the Wave 5 responses, and it is not clear how these two observations per individual should be combined – as such the Wave 9 data and onwards are not used.

[12] We also conduct various robustness checks on our exclusion restrictions; see below.

these are likely to affect the accuracy of the response and/or inclinations to inaccurately report, but, in the large part, be independent of the respondent's true health levels.[13]

Based on the arguments presented above, ideally we require proxies: for "interview-trust"; to capture respondents who are more prone to issues of wanting to appear "socially desirable"; to capture individuals with "item-preference". There is a significant amount of literature suggesting that the longer a respondent spends with the interviewer, the more trusting they are of both him/her and the survey in general (see, for example, Corbin and Morse, 2003). For each respondent it is possible to calculate the total number of questions they answered. This should be a strong proxy for length of time spent completing the survey and as such is an increasing proxy for trust. Again, we would expect this to be essentially unrelated to health levels; of course, it is possible that some severe physical (or possibly more minor mental) health issues, might be reflected by this proxy for "trust" at the margin..

A further survey-based instrument of potential use is the number of questions (relative to the total number asked) that the respondent refused to answer. Again, this provides a clear indication of survey-trust, and will be independent of health levels. The final survey-based instrument considered here seeks to capture "digit-preferencing/item-preference." We used the respondent's modal response to all other *non-health related* 5-point *Likert*- scale questions. On the basis of these a dummy variable is constructed for whether this modal response was one of the middletons or not (all based on increasing scales). Once more, this should be an excellent proxy for individuals simply "ticking boxes in the middle", and by construction should be independent of health levels (we do note the possibility again, that at the margin, it may be possible that health levels, to an extent, could possibly dictate some individuals' answering of the questionnaire in this manner; however, we would expect that this would invariably not be the case though).

Variables to Identify the Health Equations

Here variables are required that directly affect health levels, but should not have an influence on reporting behaviour. Two direct health indictors are used. The first is inclusion of a simple dummy variable indicating the presence of any long-term health conditions. The second health indicator is based on the assumption of state-dependence in health levels: apart from health shocks, health levels are likely to be highly dependent over time. The approaches

[13] A potential problem with them all (apart from whether another adult was present), is that they essentially represent heterogeneity of the interviewer and not the interviewee.

summarised in Jones *et al* (2006), Disney *et al*. (2006) and others are followed here. We include individuals' "initial health stock" to capture dynamics. Following the literature, the variable is entered as the predicted latent variable from a Generalised Ordered Probit of initial SAH levels, on a range of measured health conditions (Jones *et al.*, 2006, Disney *et al*., 2006).

The remaining instruments for the health equations are essentially health inputs and risk factors: smoking; drinking; and exercise behaviours. Although these variables are clearly good predictors of health levels, and should not affect reporting behaviour, they are potentially endogenous in a model of health. To account for this potential endogeneity/reverse causality, these are instrumented using the *Generalised Residual Inclusion* approach (Terza *et al.*, 2008). This entails estimating dynamic random effects ordered probit models for each using a standard set of demographics. Due to the panel nature of the data, the over-identifying restrictions here are simply the lagged values of the dependent/endogenous variable. To account for the endogeneity in these ancillary dynamic panel probit models, the Wooldridge approach is followed, and so we include initial conditions as covariates (Wooldridge, 2005) in these equations. The generalised residuals are calculated as the derivative of the individual log density functions with respect to the constant term in the model. These are then entered into the primary equation along with the original variable(s).

Variable definitions of the variables used, as well as some summary descriptive statistics, are given in Table 1 below. In brief, we see that average SAH is high, at somewhere between *good* and *very good*. About half the sample is male, and the majority are employed. About 20% are current smokers; nearly 30% are classified as being in the "risky-drinking" category; and some 40% undertake only low amounts of physical activity. With regard to the instruments for the inaccurate reporting equation, it can be seen that another adult was present in 36% of interviews; and respondents were generally cooperative, non-suspicious and showed a good understanding of the questions. Respondents refused to answer a relatively small number of questions. Interestingly, some 42% of observations corresponded to the modal choice of picking the "middle/right-of-middle" responses in the *Likert*-scale response items for all other non-health questions.

Insert Table 1 about here

Results

Firstly, in Table 2 we present some summary results of predicted probabilities, along with sample proportions. From this, the high observed proportions of *good* and *very good* outcomes are clearly evident: nearly 73% of responses fall into these categories (top panel, column 1). One of the key results of our findings relates to the probabilities of inaccurate reporting: clearly if there is little, or no, inaccurate reporting into these hypothesised inflated categories one would reject the basic inflation hypothesis and favour a more standard econometric approach that does not embody such outcome-inflation. In such a situation, we would simply conclude that the observed outcomes are, indeed, a true reflection of the nation's health. However, the inflation hypothesis is clearly supported by the results presented in Table 2. The model predicts some 9% prior probability – from (10), this is

$$\Pr\left(\text{inaccurate}, y=2\right)+\Pr\left(\text{inaccurate}, y=3\right)=\Phi(-x_r'\beta_r, -x_m'\beta_m; \rho_{rm})+\Phi(-x_r'\beta_r, x_m'\beta_m, -\rho_{rm}) \tag{11}$$

- that a randomly selected observation will inaccurately report into one of these two categories (bottom panel, column 1).Moreover, the estimated posterior probability of an individual inaccurately reporting (that is, conditional on them being in one of these two categories),

$$\Pr\left(\text{inaccurate} \mid y=2 \text{ or } y=3\right) \;=\; \frac{\Pr\left(\text{inaccurate}, y=2\right) \;+\; \Pr\left(\text{inaccurate}, \; y=3\right)}{\Pr\left(y=2\right) \;+\; \Pr\left(y=3\right)} \tag{12}$$

is over 12% (bottom panel).[14] Moreover, with these estimated probabilities, and others reported elsewhere, the small standard errors on them suggests that we can be confident in their magnitudes.[15]

Next, against the overall sample probabilities, a comparison is made of (averaged) overall model-predicted probabilities in the *Total* column (column 2, top panel). Thus it can be seen that sample and predicted marginal probabilities for all outcomes are reasonably close to the sample frequencies. Moreover, again the very small standard errors on these estimated

[14] Note that these, and other estimated probabilities are estimated for each observation in the sample, and then averaged.

[15] The delta method was used to estimate standard errors here, taking into account the dependence across observations due to the presence of common parameters.

probabilities, indicates that we can place quite a strong degree of confidence on these point estimates.

We next take a closer look at the probabilities for the hypothesised inflated outcomes. In particular, the *Joint Prior* column (column 3) contains the joint (prior) probability arising from inaccurate reporting and these two outcomes. In essence it provides a metric by which we can judge how much the outcome has been inflated by the hypothesized inaccurate reporting. Thus for a randomly selected observation from the population, it can be seen that of the total 39.7% probability estimated for the *good* outcome, some 4.9 percentage points (pp) of this can be attributed to inaccurate reporting into this category. This number is marginally lower (at 3.6pp) for the 37.8% in the *very good* category, but still suggests significant over-inflation here as well.

In the *Conditional Probability (Inaccurate-Reporting)* column (column 4), we present conditional probabilities of both *Good* and *Very good*, respectively, where the conditioning is on being an inaccurate reporter (all of the components for these calculations can be found in result (10)). Essentially, these probabilities give us the split of the choice between *Good* and *Very good*, for the inaccurate reporters. Thus we see that conditional on an individual being an inaccurate reporter, they are much more likely to pick *Very good* than *Good* (69% compared to 31%).

Insert Table 2 about here.

The full set of parameter estimates are given in Tables 3 and 4. The parameter results are only briefly discussed here. As these are not partial effects, it is only possible to consider significance levels and directions of effects (we return to partial effects shortly). With regard to the index function for ordered outcomes (y^* and x_y), it can be seen that the model does a very good job in explaining the health equation purged of any bias arising from inaccurate reporting. That is, with the exception of only a couple of variables, all covariates are significant predictors of this equation, and typically with the direction of effects as expected and found previously in the literature. Thus quite significant and non-linear, age and income effects; and strong negative and positive, respectively, effects of smoking and exercise can be seen. It is also noted in passing that generalised residuals for both smoking and exercise are strongly significant (but not for alcohol), indicating probable endogeneity of these variables

in the health equations.[16] Initial heath stock is strongly significant, as is the presence of any long-term health condition. Interestingly, all of the personality scale variables are all strong predictors of self-assessed health.

The equation that determines the inaccurate reporters is the one linking r^* and x_r. Of high importance here are the identifying variables. Firstly, it can be seen that neither the perceived cooperativeness nor suspicion of the respondent help to identify this equation. The suspicion of the survey in general might be better captured by the total number of refused questions in the survey–which is, indeed, marginally significant, with more refusals being associated with a greater probability of inaccurate reporting. On the other hand, if another adult was present at the interview, the respondent was estimated to be significantly more likely to report accurately, as were observations where there was a perceived worse understanding of the survey in general. Respondents apparently do gain significantly more trust with the length of time spent with the interviewer, as the total number of questions answered is a very strong positive predictor of being an accurate reporter. Finally, there is clear evidence of "digit-preferencing" and/or "middle-box-ticking", as if the respondent's modal choice of non-health related questions was in either the central response box, or the one immediately to the "right" of this, there was strong evidence that they would also tick one of these outcomes with regard to SAH.

Next the health equation is considered for those individuals identified as inaccurate reporters. Recall, that for these individuals, the choice is only one of *Good versus Very good*. This health equation appears to be well-explained by the assumed health production function, with high levels of significance across-the-board. However, there are some interesting differences with the effects estimated for the accurate reporters. For example, accurately reporting males are more likely to report higher SAH levels, whereas gender is insignificant for the latent class identified as inaccurate reporters. And whilst migrant status has no apparent association with SAH levels for the latent class of observations identified as accurate reporters, a strong negative effect is found for those identified as inaccurate reporters.

Insert Table 3 about here.

The ancillary parameters of the model are presented in Table 4. The boundary parameters are all strongly significant, which is often taken as an indication of the assumed ordering in the

[16]Denoted GR(SMOKE, EXERCISE, ALCOHOL).

response variable. Extremely strong evidence is found of unobserved effects in all three equations, with, unsurprisingly, unobserved effects in the health equation for inaccurate reporters (σ_m) being the largest. Strong evidence is also apparent for the *a priori* expected correlations between the unobservables in both the r^* and y^*equations and those in the r^* and m^* ones ($\rho_{r,y}$, $\rho_{r,m}$). These are of a similar magnitude in both equations, and imply that, all other things equal, the more likely an observation is to be in the identified latent class of accurate reporting, the higher reported SAH levels will be. Conversely, all other things being equal, the more likely an observation is to be in the identified latent class of inaccurate reporting, the lower reported SAH levels will be, within this class.

Insert Table 4 about here.

There are several probabilities that may be of potential interest to policymakers, some of which have been previously discussed. It is possible to estimate partial effects of covariates for each of these probabilities. Given space considerations and the focus of the paper, we consider only those partials concerning both the posterior and prior probability of an inaccurate reporter.[17] To be specific, in Table 5 we report a selection (of significant) partial effects on the posterior probability of being an inaccurate reporter given the observed response; $\text{Prob}(\text{inaccurate} \mid y = 2 \text{ or } y = 3)$. Table 6 reports the partial effects for the prior probability of an inaccurate reporter; $\text{Prob}(\text{inaccurate})$. In both cases these are split into positive and negative effects, and ranked in order of magnitude. Although, for reasons of space only a selection of such partials is presented, we note that high levels of statistical significance across most of the variables and the various partials were found. In particular, although not presented here, age exerted a relatively large effect.

Insert table 5 about here.

Turning first to Table 5, these partials would be interpreted along the lines of *given that an individual was observed to choose one of the inflated categories, what are the effects of explanatory variables on the probability that they are an inaccurate reporter?*. Drilling down to this level can help provide more information as to who is actually more or less likely to inaccurately report in this specific context. Due to the conditioning in these posterior probabilities, they are affected by variables in all parts of the model, not just those entering the inaccurate reporting equation. Thus we see an effect coming from the number of

[17] The full set of partials are available on request.

questions answered: one of the variables specifically hypothesized to identify these individuals. So, for a 10% rise in the number of questions answered, a clear proxy for overall trust in the survey, the individual is some ½pp less likely to report inaccurately.

The more highly educated are less likely to report inaccurately; for example, those individuals who choose an inflated outcome and who have a university education are 5pp less likely to report inaccurately. Similarly, males are some 3pp less likely to report inaccurately. Interestingly, some of our further instruments for inaccurate reporting also afford a relatively large effect: for example, a unit increase in the perceived understanding of the survey questions reduces inaccurate reporting probabilities by some 1.7pp; there being another adult present at the time of interview, decreases the probability of inaccurately reporting by some 1pp.

Turning to those variables that exert a positive effect on inaccurate reporting, we can see that for those individuals who choose an inflated outcome, being employed and married increase the probability of inaccurately reporting by some 10 and 7pp, respectively. Interestingly, whether the modal choice of all other non-health related questions was a middleton, exerts a significant and relatively large effect on the inaccurate reporting probability: these individuals are nearly 1.3pp more likely to be a "serial digit preferencer" and so report inaccurately here.

The results in Table 6 correspond to the prior probability of inaccurate reporting, and can be interpreted as *given a randomly selected person from the population, with no knowledge of their SAH outcome, what are the effects of explanatory variables on the probability that they are an inaccurate reporter?*. As might be expected, these bear a strong resemblance to the posterior effects, but, only variables that appear in the inaccurate reporting equation can have an affect here. Thus we see that, once again, a 10% rise in the number of questions answered results in a random selected individual being over 5pp less likely to report inaccurately. Having a university education and completing high school, respectively, reduce this probability by 4.7pp and 2.4pp. Being male is likely to reduce prior inaccurate probabilities by just over 2pp. In addition, two of our identifying variables exert a relatively large effect: a unit increase in the perceived understanding reduces this probability by 1.6pp, whilst there being an adult present reduces the same by nearly 1pp.Being employed and married both exert a strong positive influence (of 9 and 6pp, respectively)on the prior probability of inaccurate reporting. And once more, there is strong evidence of the validity of one of our

identifying variables, in that "box-tickers" have a 1.3pp higher probability of inaccurately reporting into their SAH levels.

The exercise of using an in-depth analysis of the model predictions can aid policymakers in identifying the potential inaccurate reporters and offer some insights into how such potentially inaccurate reporting can also be minimised.

Robustness checks

Thus far we have considered the results from what we would consider our preferred model. We have been judicious in our choice of identifying variables for both the health and inaccurate reporting equations. These variables were, in the main, significant drivers of their respective equations and the results that they yielded in terms of summary probability statistics (and the small standard errors of such) were in-line with our broad priors and eminently plausible. However, as we have no metrics against which to compare our results, the question of how confident in these results can we be, naturally arises. To address such concerns, we take a three-pronged approach: a comparison between more standard models; scenario analyses regarding the exclusion restrictions utilised in the model; and finally a small Monte Carlo analysis.

Firstly, if the researcher's aim is to essentially consistently estimate the parameters of the ordered health equation, that is β_y, then an obvious comparison is between those estimated from our approach (explicitly taking into account any potential inflation into the middleton outcomes; a variant of the Middle Inflated Ordered Probit, MIOP, model) with those from a more standard simple OP approach. Note that as these are strictly comparable across models, we need only compare the estimated coefficients (and not, for example, the implied partial effects). We report the results of such an exercise in Table 7. Thus it is apparent there are quite substantial differences if inflation is not allowed for. In particular, with regard to statistical significance levels, whilst being an ex-smoker is a strongly significant driver in the MIOP approach, it would be deemed uninfluential in the OP one; and whilst the attainment of a trade Certificate/Diploma is significant for the latter, it is only weakly so for the former. If one regards the MIOP parameters as the "true" ones, the final column of Table 7 reports percentage biases of the OP ones. None of these appear negligible: the average is just under 25%; the lowest still nearly 9%; and the highest some 230% (for the employed indicator), and even higher if one considers the insignificant variables.

Insert Table 7 about here

The next set of robustness checks involves re-estimating the model under a range of differing scenarios with regard to variable selection. That is, we explicitly employ differing identification strategies with regard to the exclusion restrictions in the model; three additional scenarios to the baseline model are considered. Firstly, although we have been careful in our choice of variables to identify the mis-reporting equation with variables that are (ostensibly) independent of true health levels, some minor associations may remain for some individuals. For example, consider the variable indicating whether the interviewer deemed the respondent to be suspicious of the survey or not. On face value, this would appear to be orthogonal to true health levels. However, particularly vulnerable individuals, in poor health, may simply be more suspicious of the interviewer, and so on. Thus, we rerun the model excluding all of the identifying variables for the mis-reporting equation. By the same logic, we also consider another specification where we exclude the identifying variables for the health equations, whilst now retaining those for the mis-reporting equation. We finally consider a specification where we retain only a subset of the inaccurate reporting identifying variables.

For reasons of brevity, we do not provide the full set of estimation results, but instead some key summary statistics in the form of estimated probabilities. We consider the key ones here to be: the probability of inaccurate reporting (prior and posterior); the marginal (or total) probabilities of each outcome; and the joint probabilities of *Good* and inaccurate reporting and *Very good* and inaccurate (as these, in comparison to the marginal probabilities of these outcomes, gives a sense for the amount of artificial "inflation" in these categories).

Insert Table 8 about here

In Table 8, we firstly report our "benchmark/baseline" results in column 1. "No health" corresponds to dropping the additional identifying variables used in the former in the health equations; [18] "No Mis-Reporting" excludes all identifying variables for the mis-reporting equation; and finally, "Subset Mis-Reporting" excludes only the interviewer-based identifying variables for the mis-reporting equation, but includes those constructed from the survey responses. Average predicted outcomes are essentially invariant to the choice of identifying variables (especially once one takes into account the sampling variability of these

[18] We retain the more objective measures of health, as these are standard covariates in empirical models of health, but drop the health input(s) and risk factors.

estimates); for example, take the *Good* outcome, compared to the baseline prediction of 39.7%, average predictions vary only from a low of 83.7 to a high of 40.1%.

Probably of more importance here though, is the extent of "inflation" the various models predict relative to the baseline one. As can be seen all models essentially predict a slightly smaller prior probability of inaccurate reporting, dropping to a low of 6% (compared to the baseline of 8.5%) for the instance where *all* identifying mis-reporting variables were removed. However, as soon as we retain at least of couple of the (arguably) more strongly legitimate variables, we see this figure rise again to the 8-9% mark (see Subset Mis-Reporting column). If one considers the posterior probabilities, again these are remarkably constant at around just over 10% (again, with the notable exception of the No Mis-Reporting results). Finally, there is also a consistent finding of around 5pp inflation of the *Good* category and around 3-4% in the *Very good* one. In summary of these results then, it appears firstly that it is important to include at least one identifying variable in the mis-reporting equation. However, notwithstanding this finding, there is a strong consensus, irrespective of the identifying set of variables used, that mis-reporting probabilities are around 8-9%, and over 10% if one considers posterior ones.

As a final set of robustness checks, we consider some Monte Carlo experiments, and explicitly with two distinct sets of data generating processes (*dgp*'s). In both, we utilise the actual data used in the empirical example, along with the estimated coefficients, and then simply draw all remaining stochastic elements of the model to complete the *dgp*. The first set of experiments consists of generating the model explicitly as per the system of equations described above. Based on these we generate draws of "observed" SAH for each individual, and then use these, as well as the observed covariates, to estimate our model. We repeat this process a large number of times (500), and compare the known probability of mis-reporting with what, on average, our model estimates. By varying the constant term in the mis-reporting equation, we consider three different values for the probability of inaccurate reporting: the baseline model estimates (of just over 8%); as well as a small number (5%); and a much larger one (20%). The key quantity of interest here is the estimated (prior) probability of mis-reporting, compared to the known true one.

In a second set of *dgp*'s, we consider what we hypothesise to be individuals' true behaviour in the population, but that which is unobserved to the researcher. Thus we explicitly split the sample into two groups: the accurate and the inaccurate reporters. This split is based on a

first-stage prediction model determined by the mis-reporting equation. In the "accurate" and "inaccurate" samples, the individuals therefore behave quite differently, according to the *m** and *y** equations as described above. However, to mimic what the researcher actually observes we then randomly combine these sub-samples back into one larger sample, and then use the model to try and disentangle them. We believe that such an approach should clearly provide a strong test of whether we can be confident on the baseline model's findings, by explicitly mimicking what we believe happens in practice. As before, we focus on the model's estimated probabilities of inaccurate reporting, and control the actual level to the baseline estimated, "high" (20%) and "low" (5%), as before. The results of both exercises can be found in Table 9.

Insert Table 9 about here

As we can see in Table 9, the model does an excellent job of correctly predicting the true prior probability of inaccurate reporting. Compared to the baseline results of some 8.5% prior probability we see that generating as per the set-up described in the Empirical Approach section, results in an averaged estimated value of marginally over this, at 8.9% (column "Average estimated"). However, if the data were generated as we hypothesise in the population, so that we explicitly split the sample into "accurate" and "inaccurate" reporters, *but then assume that this split is unobserved to the researcher*, we see that the model does even better, essentially estimating the prior probability exactly (and moreover has a smaller standard deviation as well; see column headed "Split-sample average estimated"). The model performs similarly well when the prior probability is large (at 20%), with again the split-sample *dgp* yielding marginally more accurate results. Even when the true prior probability is small (at 5%), where we would expect the model to perform worse due to smaller effective sample sizes, it performs well as estimates this 5% at 5.3 and 4.5%, respectively for the full-sample, and split sample *dgp*'s respectively. In summary of this small set of Monte Carlo experiments, it is clear that the model essentially works extremely well in correctly predicting the true, but unknown, prior probability of inaccurate reporting, even when the sample has been split in a manner unobserved to the researcher, and even when true prior probabilities of such are relatively low, and therefore much harder to identify.

Overall, these robustness checks show that the model, and are broad conclusions, appear to be very robust to specification of the exclusion restriction strategy employed, and moreover,

from the experimental evidence, we are similarly highly confident in the model's ability to accurately estimate and correctly identify inaccurate reporting levels.

Discussion

We hypothesise that in survey questions related to self-assessed health there may be mis-reporting/inaccurate reporting. Given the apparent over representation of responses in the categories *Good* and *Very good*, responses may be in some sense inflated when compared to more objective measures of health.

Using a large nationally representative panel sample, we propose and test an appropriate econometric specification to identify potential outcome-inflation and provide more information as to who is more likely to inaccurately report their responses. We find that a significant number of respondents inaccurately report into these categories, even after controlling for other effects such as long term illness. We estimate that a randomly selected member of the population has some 9% chance of inaccurately reporting into one of these inflated states; and conditional on a person choosing a middleton outcome, a conditional probability of inaccurately reporting of over 12%. Moreover, a range of robustness checks suggested that we can be rather confident in the magnitudes of these findings.

This inaccurate reporting appears to be driven by many of the standard demographic variables, including age, gender, education, employment status and personality traits. We also significant effects of "digit preferencing", trust, understanding of the survey in general and the presence of another adult. In terms of magnitudes of effects, age, education, survey trust, gender, general understanding of the questions/survey and employment and marital status all, amongst others, appear to be highly influential.

We saw how failing to account for such inflated responses in measures of self-assessed health could lead to significant changes in policy that are based on potentially unreliable responses and similarly affected econometric results. Further investigation and reflection on the implications of these issues is warranted, especially given the economic significance of the health sector, increasing costs in developed countries and the heavy policy reliance on such self-assessed measures of health.

There may valid reasons for this over inflation – reasons we cannot account for here, such as adaptation to conditions, or moving in-and-out of conditions, such as mental illness. However, we demonstrate that over reliance upon a simple metric such as the 5-point health

scale may be misleading if taken at face value. Other more sophisticated measures of health and illness alongside such scales, and the exposition of the potential over inflation in survey responses, may help make the measurement of SAH a much more useful tool in targeting effective use of scare health resources. Indeed here, we have demonstrated exactly that: how there are ways in which it may be possible to make such measures of health more useful to policymakers.

Finally, we note that such "box ticking" behaviour is also likely in the analysis of many other survey related data, not just health. Indeed, the methodology, along with the predominantly survey-based instruments we derive and suggest, respectively, to identify these inaccurate reporters are likely be widely applicable in many such instances.

References

Adams, A.S., Matthews, C.E., Ebbeling, C.B., Moore, C.G., Cunningham, J.E., Fulton, J. and Herbet, J.R. (2005). The Effect of Social Desirability and Social Approval on Self-Reports of Physical Activity', *American Journal of Epidemiology*, 161(4), 389-398.

Arce-Ferrer, A.J. (2006). 'An Investigation into the Factors Influencing Extreme-Response Style: Improving Meaning of Translated and Culturally Adapted Rating Scales', *Educational and Psychological Measurement*, 66(3): 374-392.

Bago d'Uva, T., van Doorslaer, E., Lindeboom, M., O'Donnell, O.A. and Chatterji, S. (2006) 'Does Reporting Heterogeneity Bias the Measurement of Health Disparities', *Tinbergen Institute Discussion Paper* No. 2006-033/3 Available at SSRN: http://ssrn.com/abstract=895085

Baker, M., Stabile, M., and Deri, C. (2004). 'What do self-reported, objective, measures of health measure? *Journal of Human Resources*, 39, 1067–1093.

Böckenholt, U., Barlas, S. and van der Heijden, P.G.M. (2009). 'Do randomized-response designs eliminate response biases? An empirical study of non-compliance behavior', *Journal of Applied Econometrics*, 24(3): 377-392.

Bound J. (1991).'Self-reported versus objective measures of health in retirement models', *Journal of Human Resources*, 26(1): 106-138.

Burström, B. and Fredlund, P. (2001). 'Self-rated health: Is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes?',*Journal of Epidemiology Community Health* 2001, 55:836-840.

Butler, J.S., Burkhauser, R.V., Mitchell, J.M. and Pincus, T.P. (1987). 'Measurement Error in Self-Reported Health Variables', *The Review of Economics and Statistics*, 69(4): 644-650.

Carro, J.M. and Traferri, A. (2012). 'State dependence and heterogeneity in health using a bias-corrected fixed effects estimator', *Journal of Applied Econometrics*, published online 2nd September, 2012.

Contoyannis, P., Jones, A. M. and Rice, N. (2004).'The dynamics of health in the British Household Panel Survey,' *Journal of Applied Econometrics*, 19: 473–503.

Corbin, J. and Morse, J.M. (2003). 'The Unstructured Interactive Interview: Issues of Reciprocity and Risks when Dealing with Sensitive Topics', *Qualitative Enquiry* 9(3); 335-354.

Cronbach, L.J. (1950). 'Further Evidence on Response Sets and Test Design', *Educational and Psychological Measurement*, 10(1): 3-31.

Crossley, T. F. and Kennedy, S. (2002). 'The reliability of self-assessed health status,' *Journal of Health Economics*, 21(4): 643-658.

Currie, J. and Madrian, B.C. (1999), 'Health, Health Insurance and the Labour Market' in O.C. Ashenfelter and D. Card (eds), *Handbook of Labour Economics, Volume 3C*, Amsterdam: Elsevier Science Publishers BV, pp. 3309-416.

Disney, R., Emmerson, C. and Wakefield, M. (2006). 'Ill-health and retirement in Britain: a panel data-based analysis', *Journal of Health Economics*, 25: 621-649.

Dowd, J.B., Zajacova, A. (2010). 'Does Self-Rated Health Mean the Same Things Across Socioeconomic Groups? Evidence from Biomarker Data', *Annals of Epidemiology*, 20(10): 743-749.

Dunstan, D.W., Zimmet, P.Z., Welborn, T.A., Cameron, A.J., Shaw, J., de Courten, M., Jolley, D., McCarty, D.J. (2002). 'The Australian Diabetes, Obesity and Lifestyle Study (AusDiab)—methods and response rates.' *Diabetes Research and Clinical Practice*, 57 (2): 119-129.

Ezzati M., Martin H., Skjold S., Vander Hoorn, S. and Murray C.J. (2006). 'Trends in national and state-level obesity in the USA after correction for self-reported bias: analysis of health surveys', *Journal of the Royal Society of Medicine*, 99(5):250–257.

Fry, T.R.L. and Harris, M.N. (2005). 'The DOGIT Ordered Generalised Extreme Value Model', *Australian and New Zealand Journal of Statistics*, 47(4): 531-542.

Garland, R. (1991). 'The Mid-Point on a Rating Scale: Is it Desirable?',*Marketing Bulletin*, 2, 66-70, Research Note 3.

Greene, W.H. (2012). *Econometric Analysis 7e*, Prentice Hall.

Greene, W. and Hensher, D. (2010), *Modeling Ordered Choices*, Cambridge University Press.

Hauck, K. and Hollingsworth, B. (2011). 'Health dynamics, adaptation to illness and resource allocation', *Applied Economics Letters*, 18(16): 1545-1548

Hausman, J., Scott-Morton, F. and Abrevaya, J. (1995). 'Misclassification of a Dependent Variable in Qualitative Response Models,' *Journal of Econometrics*, 63(6): 1445-1476.

Hebert, J.R., Ebbeling, C.B., Matthews, C.E., Hurley, T.G., Yunsheng, M.A., Druker, S. and Clemow, L. (2002). 'Systematic Errors in Middle-Aged Women's Estimates of Energy Intake: Comparing Three Self-Report Measures to Total Energy Expenditure from Doubly Labelled Water', *Annals of Epidemiology*, 12(8): 577-586.

Jones A.M., Rice N, Contoyannis P. (2006). 'The dynamics of health'. In: *The Elgar Companion to Health Economics,* edited by A.M. Jones. Elgar.

Jones, A.M., Rice, N. and Robone, S. (2012). 'A comparison of parametric and non-parametric adjustments using vignettes for self-reported data', HEDG Working paper, WP 12/10, University of York.

Jones, A. M., Rice, N. and Roberts, J. (2010). 'Sick of work or too sick to work? Evidence on self-reported health shocks and early retirement from the BHPS', *Economic Modelling*, 27 (4): 866-880.

Jones, A.M. and Schurer, S. (2011). 'How does heterogeneity shape the socioeconomic gradient in health satisfaction?',*Journal of Applied Econometrics*, 26: 549-579.

Jorm, A., Christensen, H., Hendersen, S., Korten, A., Mackinnon, A. and Scott, R. (1993). 'Neuroticism and self-reported health in an elderly community sample.' *Personality and Individual Differences*, 15(5): 515-521.

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

Klesges, R.C., Debon, M., Ray, J.W. (1995) 'Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey' *Journal of Clinical Epidemiology*, 48(10): 1225-1233.

Korotkov, D. and Hannah, E. (2004), 'The five-factor model of personality: strengths and limitations in predicting health status, sick-role and illness behaviour', *Personality and Individual Differences*, 36: 187-199.

Krosnick, J.A. and Alwin, D.F. (1987). 'An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement', *Public Opinion Quarterly*, 51(2): 201-219.

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

Matell, M.S. and Jacoby, J. (1972). 'Is there an Optimal Number of Alternatives for Likert Scale Items? Effects of Testing Time and Scale Properties', *Journal of Applied Psychology*, 56(6), 506-509.

Mátyás, L. and Sevestre, P. (2008). *The econometrics of panel data: fundamentals and recent developments in theory and practice 3e*, Springer.

Michel, G. (2006). 'The influence of neuroticism on concurrent symptom reporting; a multilevel modeling approach', *Personality and Individual Differences*, 41: 549-560.

Mormorstein, H., Grewal, D. and Fishe, R.P.H. (1992). 'The Value of Time Spent in Price-Comparison Shopping: Survey and Experimental Evidence,' *Journal of Consumer Research*, 19(1); 52-61.

Mossey, J.M. and E Shapiro. (1982). 'Self-rated health: a predictor of mortality among the elderly', *American Journal of Public Health*,72(8): 800-808.

Mundlak, Y. (1978). 'On the Pooling of Time Series and Cross Section Data', *Econometrica*, 46 (1), 69-85.

Prochaska, F.J. and Schrimper, R.A. (1973). 'Opportunity Cost of Time and Other Socioeconomic Effects of Away-From-Home Food Consumption', *American Journal of Agricultural Economics*, 55 (4); 595-603.

Pudney, S. and Shields, M. (2000). 'Gender, race, pay and promotion in the British nursing profession: estimation of a generalized ordered probit model', *Journal of Applied Econometrics*, 15 (4): 367-399.

Terza, J., Basu, A., and Rathouz, P. (2008). 'Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling', *Journal of Health Economics*, 27(3): 531-543.

Train, K. (2009).*Discrete Choice Methods with Simulation*, 2e, Cambridge University Press.

Wildt A.R. andMazis M.B. (1978). 'Determinants of scale response: label versus position', *Journal of Marketing Research*, 15: 261-267.

Winkelmann, R. (1996). 'A Markov Chain Monte Carlo Analysis of Underreported Count Data with an Application to Worker Absenteeism', *Empirical Economics*, 21: 575-587.

Wooden, M. and Watson, N. (2007). 'The HILDA Survey and its Contribution to Economic and Social Research (So Far)',*The Economic Record*, 83(261): 208-231.

Worcester, R.M. and T.R. Burns (1975). 'A Statistical Examination of the Relative Precision of Verbal Scales', *Journal of Market Research Society*, 17(3), 181-197.

Wooldridge, J. (2005). 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', *Journal of Applied Econometrics*, 20(1): 39-54.

Table 1: Summary Descriptive Statistics; Based on Estimation Sample of 42,120 Observations

	Mean	Standard Deviation	Brief Description
SAH	2.340	0.9479	Self-assessed health
MALE	0.460	0.4984	
MIGRANT	0.217	0.4120	
UNIVERSITY	0.237	0.4250	Highest qualification
CERT/DIP	0.315	0.4647	Highest qualification
HIGHSCHOOL	0.132	0.3385	Highest qualification
EMPLOYED	0.649	0.4774	
MARRIED	0.666	0.4718	
# CHILD	0.262	0.6999	
CITY	0.603	0.4893	
SEIFADA	5.760	2.8285	
AGREEABLENESS	5.407	0.9069	Personality Scale (1-7; increasing in the trait)
CONSCIENTIOUS	5.191	1.0114	As above
EMOTIONAL	5.254	1.0748	As above
EXTROVERTNESS	4.385	1.0743	As above
OPENNESS	4.207	1.0448	As above
SMOKER	0.197	0.3975	
EX-SMOKER	0.312	0.4633	
RISKY-DRINKER	0.270	0.4440	Drinks 1 or more days per week and 3 or more standard drinks on any drinking occasion
LOW-DRINKER	0.580	0.4936	Drinks no more than 3 days per month OR no more than 2 standard drinks on any drinking occasion
LOW PHYSICAL ACTIVITY (PA)	0.399	0.4898	Moderate or vigorous exercise 2 times or less per week
MODERATE PA	0.160	0.3667	As above but 3 times per week
HIGH PA	0.334	0.4718	As above but more than 3 times per week
LR HEALTH CONDITION	0.284	0.4508	Presence of any long-term health condition
ADULT	0.361	0.4802	Another adult present at interview
COOPERATIVE	1.157	0.3936	Cooperative scale (1-5); decreasing in trait
SUSPICIOUS	1.012	0.1164	Suspicious scale (1-3); increasing in trait
UNDERSTANDING	1.255	0.4834	Understanding scale (1-5); decreasing in trait
LN(#QUESTIONS)	5.397	0.2720	Log of total number of questions asked
% REFUSED	0.073	0.6061	% of refused questions
MODE34	0.420	0.4935	Dummy for whether modal *Likert*-scale response was “middle/right of middle”
GR(ALCOHOL, EXERCISE, SMOKING)			Generalised residuals

Table 2: Summary Estimated Average Probabilities (standard errors in parentheses)

	(1)	(2)	(3)	(4)
SAH Category	**Sample**	**Marginal (Total) Probability**	**Joint Prior Probability (Inaccurate Reporting)**	**Conditional Probability (Inaccurate Reporting)**
Poor	0.031	0.023*** (3E-4)		
Fair	0.146	0.133*** (0.001)		
Good	0.371	0.397*** (0.007)	0.0490*** (0.005)	0.310*** (0.051)
Very Good	0.355	0.378*** (0.008)	0.036*** (0.006)	0.690*** (0.051)
Excellent	0.097	0.070*** (0.010)		
Marginal Prior Inaccurate Reporting			0.085*** (0.004)	
Posterior Inaccurate-Reporting			0.121*** (0.006)	

(Estimated standard errors in parentheses, ***, **, * denote significance at 1%, 5%, 10% level)

Table 3: Estimated Model Coefficients

Variable	Index Function for Ordered Choices		Index Function for Inaccurate-Reported Outcomes		Index Function for Inaccurate-Reporting Equation	
AGE	-0.6120***	(0.045)	-0.7701	(0.523)	-0.6982***	(0.147)
AGE^2	0.395***	(0.044)	0.1882	(0.475)	0.3271***	(0.133)
Household INC	0.0873***	(0.013)	0.1969	(0.186)	0.0775	(0.050)
Household INC^2	-0.0473***	(0.010)	-0.0469	(0.412)	0.0609	(0.089)
SMOKER	-0.1413***	(0.028)	-0.7559***	(0.292)	-	-
EX-SMOKER	-0.0314**	(0.015)	0.0613	(0.132)	-	-
LOW PHYICAL EXERCISE (PA)	0.5153***	(0.024)	-0.0982	(0.199)	-	-
Moderate PA	0.8905***	(0.035)	1.541***	(0.348)	-	-
High PA	1.008***	(0.026)	.9757***	(0.245)	-	-
RISKY-DRINKER	0.1457***	(0.022)	-0.2770	(0.205)	-	-
LOW-DRINKER	0.1500***	(0.021)	0.4067**	(0.198)	-	-
GR (SMOKE)	0.0625***	(0.007)	0.1469*	(0.076)	-	-
GR (ALCOHOL)	-0.0106	(0.015)	-0.2046	(0.154)	-	-
GR (EXERCISE)	-0.1569***	(0.013)	-0.9628***	(0.165)	-	-
HEALTH	-0.9321***	(0.015)	0.5658***	(0.144)	-	-
AGREEABLENESS	0.0582***	(0.008)	0.1082*	(0.064)	0.0498**	(0.022)
CONSCIENTIOUS	0.0389***	(0.007)	0.3061***	(0.074)	-0.0809***	(0.020)
EMOTIONAL	0.1077***	(0.007)	0.0266	(0.056)	-0.0135	(0.019)
EXTROVERT	0.0943***	(0.006)	0.0469	(0.049)	0.0951***	(0.018)
OPENESS	0.0380***	(0.007)	0.1915***	(0.058)	0.0943***	(0.019)
COOPERATIVE	-	-	-	-	0.0779	(0.047)
SUSPICIOUS	-	-	-	-	-0.0274	(0.133)
UNDERSTANDING	-	-	-	-	0.1348***	(0.038)
LN(#QUESTIONS)	-	-	-	-	0.4213***	(0.059)
REFUSED	-	-	-	-	-0.034*	(0.020)
MODE3 or 4	-	-	-	-	-0.1041***	(0.029)
MALE	0.1114***	(0.015)	-0.370***	(0.131)	0.2018***	(0.040)
MIGRANT	-0.0076	(0.016)	-0.5075***	(0.146)	0.0671	(0.046)
UNIVERSITY	0.2056***	(0.021)	1.079***	(0.220)	0.3875***	(0.061)
CERT/DIP	0.04017**	(0.018)	-0.1835	(0.117)	-0.0107	(0.045)
HIGHSCHOOL	0.1959***	(0.023)	0.5748***	(0.196)	0.2029***	(0.068)
EMPLOYED	0.0493***	(0.018)	0.1953	(0.194)	-0.7449***	(0.057)
MARRIED	-0.1400***	(0.015)	-0.566***	(0.162)	-0.5003***	(0.050)
# CHILDREN	-0.0045	(0.010)	.2861***	(0.085)	-0.0544**	(0.027)
CITY	-0.0142	(0.015)	0.0999	(0.109)	-0.0765*	(0.039)
SEIFADA	0.0143***	(0.003)	-0.0550***	(0.210)	-0.0096	(0.007)
Y_0	0.495***	(0.008)	1.343***	(0.184)	-	-
ADULT	-	-	-	-	.0900***	(0.031)

GR = Generalized residuals

Table 4: Estimated Model Coefficients: Ancillary Parameters

Coefficients	Parameter Value	Standard Error
Boundary Parameters		
μ_1	1.6316***	(0.020)
μ_2	3.216***	(0.027)
μ_3	4.831***	(0.033)
Correlations		
$\rho_{r,y}$	.8237***	(0.014)
$\rho_{r,m}$	.8600***	(0.076)
Constant Terms		
Ordered Respones	-.7173***	(0.068)
Inaccurate-Reporting	-.7203*	(0.378)
Inaccurate-Reporters	-7.276***	(1.29)
Standard Deviations of Unobserved Effects		
Ordered Equation (σ_y)	0.7231***	(0.007)
Inaccurate-Reporting Equation (σ_r)	0.3700***	(0.018)
Inaccurate-Reporters Equation (σ_m)	1.013***	(0.134)
Maximised Log-Likelihood	-45,637.82	

(, **, * denote significance at 1%, 5%, 10% level)***

Table 5: Selected Estimated Marginal Effects on Posterior Probability of Inaccurate Reporting (standard errors in parentheses; all p-values ≤ 0.05)

Variable	Negative Effect		Variable	Positive Effect	
LN(#QUESTIONS)	-0.054	(0.008)	**EMPLOYED**	0.095	(0.008)
UNIVERSITY	-0.051	(0.008)	**MARRIED**	0.065	(0.007)
HIGHSCHOOL	-0.028	(0.009)	**MODE34**	0.013	(0.004)
MALE	-0.027	(0.005)	**CONSCIENTIOUS**	0.010	(0.003)
UNDERSTANDING	-0.017	(0.005)	**HEALTH**	0.008	(6E-04)
EXTROVERT	-0.013	(0.002)	**# CHILDREN**	0.007	(0.003)
ADULT	-0.012	(0.004)	**SMOKER**	0.001	(3E-04)
OPENESS	-0.012	(0.003)	**CERT/DIP**	0.001	(0.006)
HIGH PA	-0.009	(7E-04)			
MOD PA	-0.008	(6E-04)			
AGREEABLENESS	-0.007	(0.003)			
LOW PA	-0.004	(4E-04)			
LOWDRK	-0.001	(2E-04)			
RISKY-DRINKER	-0.001	(2E-04)			

Table 6: Selected Estimated Marginal Effects on Prior Probability of Inaccurate Reporting (standard errors in parentheses; all p-values ≤ 0.05)

Variable	Negative Effect		Variable	Positive Effect	
LN(#QUESTIONS)	-0.051	(0.007)	**EMPLOYED**	0.090	0.007)
UNIVERSITY	-0.047	(0.008)	**MARRIED**	0.060	0.006)
HIGHSCHOOL	-0.024	(0.008)	**# CHILDREN**	0.007	0.003)
MALE	-0.024	(0.005)	**CONSCIENTIOUS**	0.010	0.002)
UNDERSTANDING	-0.016	(0.005)	**MODE34**	0.013	0.004)
ADULT	-0.011	(0.004)			
OPENESS	-0.011	(0.002)			
EXTROVERT	-0.011	(0.002)			
AGREEABLENESS	-0.006	(0.003)			

Table 7: Comparison of Ordered Index Parameters from (Middle) Inflated Ordered Probit Model (MIOP), with those from Standard Ordered Probit Model (OP);

	MIOP	OP	Bias (%)
CONSTANT	-0.7173***	-0.4736***	-34.0
AGE	-0.6120***	-0.4909***	-19.8
AGE^2	0.3950***	0.3445***	-12.8
HOUSEHOLD INC	0.0873***	0.0571***	-34.6
HOUSEHOLD INC^2	-0.0473***	-0.0388***	-18.0
SMOKER	-0.1413***	-0.1218***	-13.8
EX-SMOKER	-0.0314***	-0.0217	-30.9
LOW PHYICAL EXERCISE (PA)	0.5153***	0.4175***	-19.0
MODERATE PA	0.8905***	0.7594***	-14.7
HIGH PA	1.0080***	0.8412***	-16.6
RISKY-DRINKER	0.1457***	0.1055***	-27.6
LOW-DRINKER	0.1500***	0.1363***	-9.2
GR (SMOKE)	0.0625***	0.0560***	-10.4
GR (ALCOHOL)	-0.0106	0.0040	-137.4
GR (EXERCISE)	-0.1569***	-0.1435***	-8.5
HEALTH	-0.9321***	-0.7558***	-18.9
AGREEABLENESS	0.0582***	0.0429***	-26.2
CONSCIENTIOUS	0.0389***	0.0482***	23.8
EMOTIONAL	0.1077***	0.0903***	-16.2
EXTROVERT	0.0943***	0.0574***	-39.2
OPENESS	0.0380***	0.0158***	-58.3
MALE	0.1114***	0.0392***	-64.8
MIGRANT	-0.0076	-0.0193	154.5
UNIVERSITY	0.2056***	0.1307***	-36.4
CERT/DIP	0.0402**	0.0239*	-40.6
HIGH SCHOOL	0.1959***	0.1442***	-26.4
EMPLOYED	0.0493***	0.1613***	227.2
MARRIED	-0.1400***	-0.0366***	-73.9
# CHILDREN	-0.0045	0.0093	-306.2
CITY	-0.0142	-0.0007	-94.9
SEIFADA	0.0143***	0.0167***	16.6
Y_0	0.4950***	0.4254***	-14.1
μ_1	1.6316***	1.3802***	-15.4
μ_2	3.2160***	2.8053***	-12.8
μ_3	4.8310***	4.2477***	-12.1

(, **, * denote significance at 1%, 5%, 10% level)***

Table 8: Robustness Checks: Exclusion Restrictions (standard errors in parentheses; all p-values ≤ 0.05)

	Baseline	**No Health**	**No Mis-reporting**	**Subset Mis-reporting**
Mis-reporting (prior)	0.085	0.079	0.057	0.078
	(0.004)	(0.004)	(0.004)	(0.005)
Mis-reporting (posterior)	0.121	0.112	0.073	0.111
	(0.006)	(0.005)	(0.005)	(0.007)
Poor	0.022	0.018	0.024	0.021
	(3E-4)	(0.001)	(0.001)	(5E-4)
Fair	0.133	0.127	0.135	0.131
	(0.001)	(0.001)	(0.001)	(0.008)
Good	0.397	0.409	0.387	0.396
	(0.007)	(0.008)	(0.005)	(0.008)
Very good	0.378	0.382	0.373	0.380
	(0.008)	(0.009)	(0.006)	(0.001)
Excellent	0.070	0.064	0.081	0.071
	(0.001)	(0.001)	(0.001)	(0.006)
Joint (*good*, mis-reporting)	0.049	0.055	0.028	0.050
	(0.005)	(0.007)	(0.005)	(0.006)
Joint (*very good*, mis-reporting)	0.036	0.024	0.029	0.028
	(0.006)	(0.007)	(0.004)	(0.007)

Table 9: Monte Carlo Experiments: True and Averaged Estimated Probabilities (standard deviations in parentheses)

	Mis-reporting (prior): true	Average estimated	Split-sample average estimated
Baseline	0.085	0.089	0.084
		(0.015)	(0.009)
High	0.200	0.207	0.201
		(0.027)	(0.007)
Low	0.050	0.053	0.044
		(0.011)	(0.009)