Reporting heterogeneity effects in modelling self reports of health*

WILLIAM GREENE†, MARK N HARRIS‡, BRUCE HOLLINGSWORTH§, RACHEL J KNOTT††, and NIGEL RICE§†

†*New York University, New York, USA*

‡*School of Economics and Finance, Curtin University, Perth, Australia*

§*Lancaster University, Lancaster, UK*

††*Centre for Health Economics, Monash University, Melbourne, Australia*

§†*Centre for Health Economics & Department of Economics and Related Studies, University of York, York, UK*

Abstract

Self-assessed measures of health using *Likert*-type scales are widely used to assess the health and well-being of populations, and are a feature of household surveys throughout the world. However, the self-reported and subjective nature of these measures means that different people will inherently respond in different ways - a concept known as reporting heterogeneity. In this paper we consider two types of reporting heterogeneity. The first is *differential item functioning*, which results when individuals systematically differ in their interpretation and use of response categories. The second is *middle-inflation bias*, which arises when respondents adopt a 'box-ticking' strategy - for example, because they are unsure of how to answer survey questions, or because they do not take the surveys they are completing seriously. This type of reporting heterogeneity typically materializes in the form of an artificial build-up of responses in middling response categories. We consider approaches for adjusting for each type of reporting heterogeneity, both in isolation and in combination. The results suggest that self-assessed measures of health are susceptible to both types of reporting heterogeneity, and that failure to account for these nuances may lead to erroneous inference concerning the analysis of self-reported health.

*This research was funded by an Australian Research Council Discovery Project Grant

(DP110101426), a BankWest Curtin Economics Centre (BCEC) grant, and a Monash Faculty of Business and Economics grant. The usual caveats apply.

Keywords: Self-assessed health, inflated ordered outcomes, varying individual response scales, anchoring vignettes.
JEL: I1, C1, C3

1. Introduction and Background

Social surveys typically contain multiple measurement instruments in the form of self-assessments to capture the circumstances, preferences or beliefs of respondents. These include questions relating to job and life satisfaction, satisfaction with public services, political efficacy, work disability and health status. Available responses usually consist of several ordered categories, for example the ubiquitous self-assessed general health measure typically asks respondents to rate their health using a 5-point scale ranging between *very poor* or *poor* health to *very good*, or *excellent* health. There are compelling reasons for such measures being a staple feature of most major household surveys (cross-section and panel), most notably the relative ease and low cost of data collection. In the absence of more objective measures, such measures contain valuable information from which to infer differences across individuals, socio-economic groups or countries. Moreover, and particularly in the case of self-assessed health, such measures have been shown to be good predictors of other outcomes such as mortality [13, 2, 3]. Accordingly, self-assessments feature strongly in empirical social science survey research.

Although widely used, self-assessments are subject to various forms of *measurement error*. Responses to these types of questions, where the answers are of the form: *not much*, *a bit*, *good*, *bad*, *excellent*, and so on, are by definition, subjective responses, and ones that may vary considerably across survey respondents. Since the response categories are open to subjective interpretation, even where respondents are facing a fixed and known level of the construct under consideration, their respective assessments may vary. Accordingly, responses may reflect both the objective reality and respondents' interpretation of the subjective scale. As such two individuals with identical levels of underlying true health may respond *good* and *very good* to a given question. This may occur due to a range of observed, or unobserved factors; for example, different expectations, underlying levels of optimism or pain thresholds. As a consequence, individuals when faced with a self-assessment

make use of different *reporting response* scales. In an attempt to ameliorate this effect, self-reported questions are sometimes asked relative to a scale of reference, for example: *Relative to someone of your own age, how would you rate your health?* Although, even here, response heterogeneity may remain.

To address the issue of differential reporting behaviour, recent research has advocated the use anchoring vignettes to detect and adjust for heterogeneity in individuals' reporting scales [6, 5, 12**? ? ?**]. Essentially this entails asking individuals to rate one or more of a set of vignettes, or questions about the health status of a hypothetical person. Since all respondents rate the same (set of) vignette(s), the responses can then be used to anchor, or adjust the respondent's self-assessment of the concept of interest. This is aimed at increasing inter-respondent comparability by abstracting reporting behaviour from the underlying construct under investigation.

Such response heterogeneity could be considered measurement error at the margin, although recent research suggests a potentially more systematic form of miss-measurement in such self-reports.[1] In particular [9] consider the distribution of self-assessed health (SAH), collected from a large representative sample of Australians, although similar issues are to be found in comparable surveys across the developed world. The responses to the SAH question are clearly bunched around the two "middle" responses of *good* and *very good*.[2] They argue that this paints a rather rosy picture of the health of the population in such countries, whereas many more objective measures (such as obesity rates, exercise rates, widespread levels of elevated cholesterol levels, and so on) paint a much bleaker picture[9]. They conclude that around 10% of the sample, for a range of reasons, simply tick these middle boxes and thus artificially inflate the numbers in these categories.

A potential criticism of this approach, however, is that some, or all, of the clustering of observations in the middle could actually be attributed to heterogeneity in reporting scales as opposed to a box-ticking strategy. Conversely, approaches that use anchoring vignettes, alone, could be biased if they erroneously attribute a box-ticking strategy to reporting heterogeneity

[1] It is important to note that here, and elsewhere, we use the term "measurement error" very loosely. Thus two individuals with identical health, but have differing self-reports of such, are (presumably) reporting their *self-assessments* correctly, but the measurement error is simply referring to the observation that these are different.

[2] Specifically, if one denotes the 5-outcomes, from worst to best, as $j = 0, 1, \ldots, 4$, these correspond to $j = 2$ and 3.

in response scales. This paper considers both types of reporting heterogeneity, both singularly and jointly. Our methodological approach innovates by combining the two forms of reporting heterogeneity into a single estimation approach. By using a bespoke population survey of Australians, which includes socio-demographic information together with a self-assessment of general health and vignettes, we estimate the determinants of health (and reporting behaviour, or middle inflation) allowing for reporting heterogeneity in response scales and/or middle inflation. Results support the notion that respondents when self-reporting are susceptible to both general reporting behaviour and artificial inflation of certain categories. Failure to account for these nuanced reporting effects leads to erroneous inference concerning the determinants of the construct under investigation.

2. Methods

2.1. Ordered probability models

Self-reports of health are invariably collected via survey instruments where the answers are responses on a *Likert*-type scale. Thus analysis of such data typically use ordered probability models [7]; see for example [4]. 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 (with no constant term, though this can be relaxed below) $\mathbf{x}_y$; a (standard normal) disturbance term, ε_y; and its relationship to certain boundary parameters, μ. Formally, we have

$$y^* = \mathbf{x}_y' \boldsymbol{\beta}_y + \varepsilon_y, \tag{1}$$

which translates into the observed $j = 0, \ldots, J-1$ outcomes via the mapping

$$y = \begin{cases} 0 & \text{if } \mu_{-1} < y^* \leq \mu_0 \\ 1 & \text{if } \mu_0 < y^* \leq \mu_1 \\ \vdots & \vdots \\ J-1 & \text{if } \mu_{J-2} < y^* \leq \mu_{J-1}, \end{cases} \tag{2}$$

where there are J outcomes (typically $J = 5$) and with the restriction, to guarantee well-defined probabilities, that $\mu_{-1} \leq \mu_1 \cdots \leq \mu_{J-1}$, and with $\mu_{-1} = -\infty$ and $\mu_{J-1} = \infty$.

Under the assumption of normality, the respective probabilities for each ordered outcome are

$$\Pr\left(y = j\,|\mathbf{x}_y\right) = \begin{cases} \Pr\left(j = 0\,|\mathbf{x}_y\right) = \Phi\left(\mu_0 - \mathbf{x}_y'\boldsymbol{\beta}_y\right) \\ \Pr\left(j = 1\,|\mathbf{x}_y\right) = \left[\Phi\left(\mu_1 - \mathbf{x}_y'\boldsymbol{\beta}_y\right) - \Phi\left(\mu_0 - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right] \\ \vdots \\ \Pr\left(j = J-1\,|\mathbf{x}_y\right) = \left[1 - \Phi\left(\mu_{J-2} - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right], \end{cases} \tag{3}$$

where $\Phi(.)$ denotes the standard normal distribution function evaluated at its argument. The (log) density for this model for a $n = 1, \ldots, N$ random sample of individuals is simply given by

$$\ln L_{OP}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} \left[\Pr\left(y_i = j\,|\mathbf{x}_i\right)\right], \tag{4}$$

where d_{ij} is a function returning one if individual i chose outcome j, and zero otherwise, and $\boldsymbol{\theta}$ are all of the parameters in the model.

2.2. Generalised ordered probit models

For a given level of true health, y^*, differences in reporting scales across respondents can be accommodated by individual varying boundary parameters, μ_{ij}. Allowing the boundary parameters of the OP model to vary has a long history; see, for example, [15], [14], [1], [7], [10]. One way to allow for this, is to specify the boundaries as a function of observed characteristics $\mathbf{z}_i$ (where $\mathbf{z}_i$ must include a constant) such that

$$\mu_{ij} = \mathbf{z}_i'\gamma_j. \tag{5}$$

Such models that allow for varying boundary parameters are usually referred to as generalised ordered probit (GOP) models. Two unappealing facets of such a specification are: firstly, the effects of any variables that overlap across $\mathbf{z}$ and $\mathbf{x}$ cannot be uniquely identified in the boundary and structural equations; and secondly, the required ordering of the thresholds is in no way ensured. For these reasons, many authors adopt a hierarchical ordered probit (HOPIT) approach. This differs from equation (5) in that the boundaries are forced to be increasing (that is, *hierarchical*) by specifying them as

$$\begin{aligned} \mu_{i0} &= \mathbf{z}_i'\boldsymbol{\gamma}_0 \\ \mu_{ij} &= \mu_{ij-1} + \exp\left(\mathbf{z}_i'\boldsymbol{\gamma}_j\right) \\ &\vdots \end{aligned} \tag{6}$$

where the $\exp(.)$ ensures the necessary ordering and also aids in identification. Once again however, for any variables that appear in both $\mathbf{x}_y$ and $\mathbf{z}$, the corresponding elements of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_y$ are not separately identified as the first threshold is specified linearly.

2.3. Use of anchoring vignettes

Anchoring vignettes offer a method of anchoring individual response scales when used in conjunction with the main self-report of interest. In general, interpersonal incompatibility in response scales is known as *differential item functioning* (DIF). Alongside a self-assessment of a construct of interest, for example health, respondents are also asked to rate a set of vignettes ($k = 1, \ldots, K$) describing a hypothetical person. The response scale available to rate the K vignettes is the same $j = 0, \ldots, J-1$ scale used for the self-report of interest.

An example of a vignette, and one used in the empirical example below, is:

> KEVIN walks for one to two kilometres and climbs three flights of stairs every day without tiring. He keeps himself neat and tidy and showers and dresses himself each morning in under 15 minutes. He works in an office and misses work one or two days per year due to illness. Kevin has a headache once every two months that is relieved by taking over-the-counter pain medication. He remains happy and cheerful most of the time, but once a week feels worried about things at work. He feels very sad once a year but is able to come out of this mood within a few hours.
>
> In GENERAL, would you say KEVIN'S HEALTH is:
> · Excellent
> · Very Good
> · Good
> · Fair
> · Poor

Define the observed response to each $k = 1, \ldots, K$ possible vignette as y_{ik}; then, as above, this is assumed to be dependent upon an unobserved continuous latent measure, y^*_{ik} and embodies the mapping

$$y_{ik} = j \text{ if } \mu^{j-1}_{ik} \leq y^*_{ik} < \mu^j_{ik},\ k = 1, \ldots, K;\ j = 0, \ldots, J-1, \tag{7}$$

with $\mu_{-1} = -\infty$, and $\mu_{J-1} = \infty$.[3] The y^*_{ik}'s are assumed to be a function of a constant and random error,

$$y^*_{ik} = \alpha_k + \varepsilon_{ik}, \tag{8}$$

with $\varepsilon \sim N(0, \sigma^2_k)$ and orthogonal to all observed covariates in the model. Often the simplifying assumption that the variance (σ^2) is the same across the vignettes is imposed, such that $\sigma^2_k = \sigma^2\ \forall k$. Heterogeneity across the response scales is once more allowed for by specifying the cut-points as a function of threshold variables, $\mathbf{z}_i$, and having the same form as equation (6).

To identify the model the literature advocates assuming response consistency (RC) and vignette equivalence (VE) [12]. The RC assumption amounts to stating that individuals apply the same reporting scale for their vignette responses as they do for the key self report of interest (for us, self-assessed health). That is, there may be DIF across individuals, but not within. Formally, this amounts to restricting all coefficients in the reporting parts of the model (the boundary parameters: $\boldsymbol{\gamma}_j$) to be equivalent across the vignettes and self-assessment. That is, $\boldsymbol{\gamma}$ in the $HOPIT$ part of the self-report of interest is identical to that in all of the $k = 1, \ldots, K$ $HOPIT$ parts in the vignette equations. In this way, the model is over-identified such that it is possible, indeed very common, to have $\mathbf{X} \equiv \mathbf{Z}$ (even with linear boundaries as in equation (5)).

The assumption of VE requires that the level of health described in a particular vignette is perceived by all respondents in the same way, and on the same scale. Differences across respondents is attributed to random error. This leads to the specification described in equation (8). Indeed,

[3] With multiple vignettes an unobserved individual-specific effect can be included in the specification of the thresholds such that in equation (6) $\mu_{i0} = \mathbf{z}'_i\boldsymbol{\gamma}_0 + u_i$ (see Kapteyn, Smith and van Soest[**?**]). In the interests of parsimony we do not include such an extension.

it is this assumption that ensures that the vignettes are indeed anchoring vignettes. The remaining restrictions usually employed in the literature are those of location and scale - here we restrict the constant in the first boundary equation to be zero and the variance in the self-assessed component to be one.

The (log-)likelihood function for the $HOPIT$ model consists of two distinct parts: one relating to the self-report of interest, and a second to the vignette component of the model. For the vignette component, we have for a given k ordered probabilities of the form

$$\Pr(y_{ijk} = j \,|\mathbf{z}_i) = p^v_{ijk}(\mathbf{z}_i) = \begin{cases} \Pr\left(j = 0 \,|\mathbf{z}_i\right) = \Phi\left(\left[\mu_{i0} - \alpha_k\right]/\sigma\right) \\ \Pr\left(j = 1 \,|\mathbf{z}_i\right) = \left[\Phi\left(\left[\mu_{i1} - \alpha_k\right]/\sigma\right) - \Phi\left(\left[\mu_{i0} - \alpha_k\right]/\sigma\right)\right] \\ \vdots \\ \Pr\left(= J - 1 \,|\mathbf{z}_i\right) = \left[1 - \Phi\left(\left[\mu_{i,J-2} - \alpha_k\right]/\sigma\right)\right], \end{cases} \tag{9}$$

where the μ_{ij} are of the form of equation (5); such that the (log-)likelihood contribution arising from the vignettes component over $k = 1, \ldots, K$ vignettes is

$$\ln L_V = \sum_{i=1}^{N} \sum_{k=1}^{K} \ln \sum_{j=0}^{J-1} d_{ijk} \times p^v_{ijk}\left(\mathbf{z}_i\right), \tag{10}$$

where d_{ijk} is now the vignette-specific indicator variable. The contribution from the self-report ($HOPIT$) part of the model will be

$$\ln L_{HOPIT} = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} \left[\Pr\left(y_i = j \,|\mathbf{x}_i, \mathbf{z}_i\right)\right], \tag{11}$$

where y_i corresponds to the observed value of y for the self-report of interest. Importantly the inherent boundary parameters of equation (11) are also driven by equations of the form (6) with identical coefficients $\boldsymbol{\gamma}$, as those in the vignettes component of equation (10). The overall (log-)likelihood is simply the sum of these two components such that

$$\ln L = \ln L_V + \ln L_{HOPIT}, \tag{12}$$

where the first term is a function of α_k, σ and $\mu_{ij}\left(\boldsymbol{\gamma}_j\right)$; and the second term

a function of $\boldsymbol{\beta}$ and $\mu_{ij}\left(\boldsymbol{\gamma}_j\right)$. Thus these two components are linked through the common cut-point (or boundary) parameters μ_{ij}, and so do not factorise into two independent models.

2.4. Middle inflation models of reporting heterogeneity

We assume here that both the self-report of interest and the corresponding vignettes are labelled as running from worst to best (*poor* is coded as $y = 0$, *fair* is $y = 1$ and so on). With this in mind, heterogeneity in reporting behaviour that can be addressed with the use of vignettes, can be considered to primarily operate at the margin. For example, two individuals with identical true health but who embody DIF, are (presumably) more likely to exhibit this with self-reports into neighbouring categories, but much less so anything more extreme than that. However, recent literature [9] considers a much more structural form of reporting heterogeneity.

The approach involves explicitly allowing for the outcomes corresponding to the middle two outcomes to be *inflated*: in some sense they are an over-representation of a population's true health status in these outcomes [9]. For example, it is typical to find about 70% of observations in the *very good* and *good* categories in reports of self-assessed health - for a 5-point likert scale, running from "poor"($j = 0$) to "excellent"($j = 4$), this would correspond to outcomes $j = 2, 3$. 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 (i.e. *middle inflate*). Importantly, this index will be uncorrelated to their reporting scales. Let this latent variable be a function of a set of observed covariates, $\mathbf{x}_r$, with unknown weights $\boldsymbol{\beta}_r$, and a (standard normal) disturbance term, ε_r such that

$$r^* = \mathbf{x}_r'\boldsymbol{\beta}_r + \varepsilon_r. \tag{13}$$

When this index reaches a critical level (normalised to zero), the individual will accordingly report accurately ($r = 1$); otherwise, they will employ a *box-ticking* strategy. Under normality, the probability that an individual will report "accurately" is therefore a probit probability of the form

$$\Pr\left(r = 1\right) = \Pr\left(r^* > 0\right) = \Phi\left(\mathbf{x}_r'\boldsymbol{\beta}_r\right). \tag{14}$$

Conditional on being in the *non-box-ticking* regime, a standard OP formation given by equation (3) applies; driven by covariates $\mathbf{x}_y$ with associated

weights $\boldsymbol{\beta}_y$. However, for individuals with a box-ticking propensity, they will essentially make a binary choice between *good* and *very good*. Without loss of generality this can be determined by a further latent variable of the form

$$m^* = \mathbf{x}'_m\boldsymbol{\beta}_m + \varepsilon_m, \tag{15}$$

where it is expected that $\mathbf{x}_m \equiv \mathbf{x}_r$. Again, once this index reaches a threshold value, normalised to zero, this triggers the choice of *good* relative to *very good*. Thus under independence of the stochastic elements of the system, joint probabilities of inaccurate reporting and *very good*, and of inaccurate and *good*, will be[4]

$$\begin{aligned} \Pr\left(inaccurate, good\right) &= \Phi\left(-\mathbf{x}'_r\boldsymbol{\beta}_r\right)\Phi\left(\mathbf{x}'_m\boldsymbol{\beta}_m\right) \\ \Pr\left(inaccurate, very\ good\right) &= \Phi\left(-\mathbf{x}'_r\boldsymbol{\beta}_r\right)\Phi\left(-\mathbf{x}'_m\boldsymbol{\beta}_m\right). \end{aligned} \tag{16}$$

Conversely, for all accurately reporting respondents, they can choose freely across the 5-point choice set, but for them to do this, one has to take into account the conditioning. Thus joint probabilities here are now, by independence, OP probabilities as given by equation (3), but now weighted by the probabilities of accurate reporting, $\Phi\left(\mathbf{x}'_r\boldsymbol{\beta}_r\right)$,

$$\Pr(accurate, y) = \begin{cases} \Pr\left(j=0\right) = \Phi\left(\mathbf{x}'_r\boldsymbol{\beta}_r\right) \times \left[\Phi\left(\mu_0 - \mathbf{x}'_y\boldsymbol{\beta}_y\right)\right] \\ \Pr\left(j=1\right) = \Phi\left(\mathbf{x}'_r\boldsymbol{\beta}_r\right) \times \left[\Phi\left(\mu_1 - \mathbf{x}'_y\boldsymbol{\beta}_y\right) - \Phi\left(\mu_0 - \mathbf{x}'_y\boldsymbol{\beta}_y\right)\right] \\ \Pr\left(j=2\right) = \Phi\left(\mathbf{x}'_r\boldsymbol{\beta}_r\right) \times \left[\Phi\left(\mu_2 - \mathbf{x}'_y\boldsymbol{\beta}_y\right) - \Phi\left(\mu_1 - \mathbf{x}'_y\boldsymbol{\beta}_y\right)\right] \\ \Pr\left(j=3\right) = \Phi\left(\mathbf{x}'_r\boldsymbol{\beta}_r\right) \times \left[\Phi\left(\mu_3 - \mathbf{x}'_y\boldsymbol{\beta}_y\right) - \Phi\left(\mu_2 - \mathbf{x}'_y\boldsymbol{\beta}_y\right)\right] \\ \Pr\left(j=4\right) = \Phi\left(\mathbf{x}'_r\boldsymbol{\beta}_r\right) \times \left[1 - \Phi\left(\mu_3 - \mathbf{x}'_y\boldsymbol{\beta}_y\right)\right]. \end{cases} \tag{17}$$

Finally, marginal probabilities for the middle-inflated $\left(P^{MIOP}\right)$ of the full

[4]Note that once more we are only using the terms (in)accurately, as convenient labels. Indeed, they may be appropriate, or they simply may represent what the respondent truely believes their health is. In the latter case, they would still be an accurate representation of a *self*-report.

choice set are simply the sum of these two (16,17) such that $P^{MIOP} =$

$$\begin{cases} \Pr(j=0) = \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r) \times \left[\Phi\left(\mu_0 - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right] \\ \Pr(j=1) = \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r) \times \left[\Phi\left(\mu_1 - \mathbf{x}_y'\boldsymbol{\beta}_y\right) - \Phi\left(\mu_0 - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right] \\ \Pr(j=2) = \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r) \times \left[\Phi\left(\mu_2 - \mathbf{x}_y'\boldsymbol{\beta}_y\right) - \Phi\left(\mu_1 - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right] + \left[\Phi\left(-\mathbf{x}_r'\boldsymbol{\beta}_r\right)\Phi\left(-\mathbf{x}_m'\boldsymbol{\beta}_m\right)\right] \\ \Pr(j=3) = \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r) \times \left[\Phi\left(\mu_3 - \mathbf{x}_y'\boldsymbol{\beta}_y\right) - \Phi\left(\mu_2 - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right] + \left[\Phi\left(-\mathbf{x}_r'\boldsymbol{\beta}_r\right)\Phi\left(\mathbf{x}_m'\boldsymbol{\beta}_m\right)\right] \\ \Pr(j=4) = \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r) \times \left[1 - \Phi\left(\mu_3 - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right]. \end{cases} \tag{18}$$

Such a model now embodies the inflation hypothesis of the *good* and *very good* outcomes [9], whereby these probabilities are inflated from the inaccurate reporters. Once the form of the probabilities, dependent on unknown parameters and observed data, is known, the model can be estimated by ML techniques where the (log-)likelihood function for the middle-inflated ($MIOP$) model is now

$$\ln L_{MIOP} = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} P_{ij}^{MIOP}, \tag{19}$$

where P_{ij}^{MIOP} are those as given by equation (18). Such models, by inflating middle categories, are generally referred to as middle-inflated OP, or $MIOP$, models.

Several potential reasons arise as to *why* some individuals may misreport into these middle categories [9]. These include amongst others: a general distrust of surveys; an opportunity cost-of-time argument; a desire to want to appear more socially acceptable; and so on.

2.5. Tempered inflation models of reporting heterogeneity

Implicitly in the middle inflation models, as described above, there is an inherent ordering: first an individual decides if they have a propensity to report accurately, or not, and conditional on this decision then reports accordingly (such that, in the set-up above, a box-ticker, would only ever tick the *good* and *very good* outcomes). However, [8] consider reversing this implicit ordering, albeit in a much simpler set-up with only three (ordered) outcome choices and hypothesised inflation in only the middle one of these. Assume that an individual first has a propensity to translate their notions of her true underlying health, y^*, into one of the five observed categories (where the generalisations to more or less than five categories is implied). However, for similar reasons to those noted above with regard to inaccurate

reporting, extreme values of a preferred outcome/choice are tempered by equations that similarly allow for a tendency for individuals to be pulled towards the middle categories.[5] As an example, consider a respondent who has an underlying propensity for either *good* or *very good*. These outcomes are already in the middle such that there would be no further forces acting to pull them towards these middle outcomes. However, consider a different respondent who has a true propensity for the neighbouring choice of *excellent*. Clearly there will remain a non-zero probability that they will still choose this outcome, but these are likely to be tempered, for some, with a pull towards the middle outcomes. Although our set-up and hypothesis here is somewhat more complicated than that of [8], we show how their general approach can be adopted.

Firstly, we assume a standard OP set-up as described in equations (1) to (3). We will label these "first stage" probabilities, $P_{y0}, P_{y1}, \ldots, P_{y,J-1}$. For individuals with a $j = 4$ (*excellent*) propensity, we wish our approach to simultaneously allow for tempering towards the inflated middle outcomes, and also for the respondent to simply choose *excellent*. Thus conditional on this first stage propensity, a *probit* model applies with potential outcomes: *excellent*, and *very good*. Let this tempering equation be determined by a latent equation of the form

$$t_4^* = \mathbf{x}_r'\boldsymbol{\beta}_4 + \varepsilon_4, \tag{20}$$

with resultant probabilities of $P_{4|4}$ and $P_{3|4}$, where the conditioning indicates that these are tempering probabilities *from* (or conditional on), the *excellent* ($j = 4$) outcome in the first stage.

Moving along the choice scale (assuming a $J = 5$, 5-point *Likert*-type scale), the first stage probabilities of the middle outcomes (P_{y2}, P_{y3}) will be left untempered. Next, consider the $j = 1$ choice, of *fair*. Once more, to allow for tempering from this choice, we can envisage a further secondary latent equation of the form

$$t_1^* = \mathbf{x}_r'\boldsymbol{\beta}_1 + \varepsilon_1 \tag{21}$$

which will now drive this conditional choice to either *fair* (that is, there is

[5] Where extreme would essentially be any outcome outside of the inflated middle categories.

no tempering); or to the inflated neighbouring outcome of *good*. Recognising the binary nature of these choices, equation (21) will translate itself into a further probit equation with resultant probabilities of $P_{1|1}$, and $P_{2|1}$.

In the case of a 5-point scale, there is the further choice of *poor* $(j = 0)$. It might be considered that this outcome is sufficiently far from the inflated outcomes such that no, or very little tempering is likely. Or, it might be that tempering from this extreme is still present in the data. However, given the low frequency of reports observed in our data falling within the *poor* category (5.2%), we do not consider tempering from this outcome.

Thus overall probabilities for the tempered approach $\left(P^{TOP}\right)$ are given by

$$\Pr(y = j) = P^{TOP} = \begin{cases} j = 0 & P_{y0} \\ j = 1 & P_{y1,1} \\ j = 2 & P_{y2} + P_{y1,2} \\ j = 3 & P_{y3} + P_{y4,3} \\ j = 4 & P_{y4,4} \end{cases} \tag{22}$$

The resulting (log-)likelihood function for the tempered (TOP) model is

$$\ln L^{TOP} = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} P_{ij}^{TOP}. \tag{23}$$

In the simple case with $J = 3$ and inflation of only the middle category, [8] show that the standard middle-inflated OP model is actually a restricted version of their TOP model, such that the latter can be used as a specification case test for the former. However, in our case, such a comparison is not particularly obvious due, for example, to the presence of the additional boundary parameters, in our tempering equations. Note that this specification of the TOP model is implicitly tempering from near neighbours only. It would be possible to generalise this approach to allow for tempering across a range of neighbouring alternatives.[6]

2.6. Inflation and tempered models with vignette reporting adjustments

The methods, described above, have been developed and employed in isolation. When viewed in isolation, one could erroneously attribute reporting

[6]However, in estimation this model seemed overtly prone to convergence issues and was therefore not pursued.

heterogeneity to say DIF when in fact there was only inflation heterogeneity present in the data, and *vice versa*. However, clearly there exists the possibility that both forms of reporting heterogeneity operate jointly. Accordingly, we combine the above approaches to simultaneously account for both forms of reporting behaviour. To combine DIF with the $MIOP$ model for middle-inflation, it is necessary to first allow the reporting-scale parameters of the $MIOP$ model, $\boldsymbol{\mu}^{MIOP}$ to be person-specific, such that the μ's of equation (18) are now

$$\begin{aligned} \mu_{i0}^{MIOP} &= \mathbf{z}_i'\boldsymbol{\gamma}_0 \\ \mu_{ij}^{MIOP} &= \mu_{ij-1} + \exp\left(\mathbf{z}_i'\boldsymbol{\gamma}_j\right) \\ &\vdots \end{aligned} \tag{24}$$

and as before, with the more standard $HOPIT$ approach, of equation (12), a separate $HOPIT$ model for the vignette responses applies. Again, enforcing the anchoring of individual specific reporting scales via the vignettes requires equality of the boundary parameters across the two models, such that we maintain identical coefficients $\boldsymbol{\gamma}$ in both the vignettes part of the $HOPIT$ model as those in the self-assessment part of the $HOPIT$ model.

To estimate this augmented model, one would simply replace the likelihood contribution arising from the $HOPIT$ part of the model, $\ln L_{HOPIT}$, in equation (12) with that from the $MIOP$, $\ln L_{MIOP}$, from equation (19), once the definition of the boundaries has been changed along the lines of equation (24) in the latter; such that

$$\ln L_{V/MIOP} = \ln L_V + \ln L_{MIOP}. \tag{25}$$

Again, as these two components are linked through the common cut-point (or boundary) parameters and the overall model does not factorise into two independent models. Conceptually, the identifying assumption that the reporting behaviour captured by DIF is independent of that which drives the accurate/inaccurate reporting behaviour as described by equation (13). That is, the key identifying assumption is that the vignette responses are not similarly affected by box-ticking behaviours.

Similarly we can also nest the $HOPIT$ approach within the general TOP setting. Once more, we simply alter the constant cut-point parameters in-

herent in equation (22) to again be of the form

$$\begin{aligned} \mu_{i,0}^{TOP} &= \mathbf{z}_i'\boldsymbol{\gamma}_0 \\ \mu_{i,j}^{TOP} &= \mu_{i,j-1} + \exp\left(\mathbf{z}_i'\boldsymbol{\gamma}_j\right) \\ &\vdots \end{aligned} \tag{26}$$

In this set-up is the identifying assumption that vignette responses are not similarly affected by *tempering*. The likelihood function will now be

$$\ln L_{V/TOP} = \ln L_V + \ln L_{TOP}. \tag{27}$$

3. Data and variable selection

3.1. Survey and socio-economic characteristics

We use data from an online survey involving a representative sample of Australians aged 18 to 65 years. The sample was collected in two waves - the first in April 2014 (n=2,007) and the second in August 2015 (n=3,027), resulting in a pooled sample size of 5,034. Summary statistics of the sample are provided in Table 1. Importantly with respect to self-assessed health, there appears to be inflation within the middle categories, with the sample proportions very similar to those reported in [9], where the latter was based on a large nationally representative sample. Thus we see that nearly 70% of respondents reported either *very good* or *good* health; 13% reported *excellent* health; 15% reported fair; and only 5% reported poor health.

Insert Table 1 about here

Following the literature (see for example, [4]) we include the following set of covariates to model health and allow these to similarly affect the $HOPIT$ parts of all of the models considered (that is, to affect the boundary equations capturing DIF). This includes variables for age/100 and $(\text{age}/100)^2$, gender, highest educational qualification received, employment status, marital status, and migrant status.

Mean age for the sample is 41.4 years and 52% are female. Highest qualification received is characterised by three categories, namely tertiary education (which includes both university and post-school diplomas and certificates, these respondents constitute 69% of the sample), high school (15%

of respondents), and has not completed high school (16% of respondents, the reference). In terms of employment status, 75% of respondents are employed (the reference category), 10% are unemployed, and 14% are not in the labour force. 59% of the sample are married or in de facto relationships, 11% are divorced/separated or widowed, and 30% have never married (the reference). Finally, 25% of respondents were born overseas.

3.2. Vignettes to identify DIF

The survey contained three anchoring vignettes from which to identify DIF. The first describes a relatively favourable state of health, the second describes a moderate health state, and the third describes a comparatively poor state of health. The order in which the vignettes were presented was randomized across respondents. Vignettes were gender specific in terms of the names used to describe the hypothetical individuals (as suggested by [12]), and are presented in the appendix.

3.3. Variables to identify inflation and tempering

Clearly an important set of covariates are those to be used to identify the *inflation* parts of the models considered. As the inflation and tempering equations, described above, can be considered to be ostensibly driven by similar covariates, the same set is utilised across these two models. Thus here we require variables that are orthogonal to true health levels, as well as to any potential DIF present in the data. That is, we require measures to identify those individuals who report inaccurately, and are hence drawn to modal responses. Such measures could include proxies for individuals who do not understand the questions and/or the survey in general and individuals who may not be taking the survey seriously. We have several potential candidates in the dataset to proxy these factors [9]. In terms of individual's accuracy of responses to survey questions, respondents were asked at the end of the survey if they understood the questions of the survey - 84% indicated that they did. The data also contains information on whether others were present while the survey was being completed, which we utilize, as [9] found that having others present increased the likelihood of accurate reporting. This variable is represented by a binary indicator where 1 indicates that the respondent completed the survey alone (92% of respondents). Finally, respondents were asked whether the money they receive from participating in online surveys contributes significantly to their household income (also represented by a dummy variable, where *1* is *yes* - 30% of the sample), which may provide

an indication for whether the survey was taken seriously, and hence whether reporting was accurate, or not.

These variables should be ostensibly unrelated to true health and differing response item scales, such that jointly, they should adequately identify these inflation and tempering equations. We also include the standard range of demographics, as discussed above, in all misreporting equations, although we have very few priors about the direction of these effects. Accordingly, $\mathbf{x}_r$ and $\mathbf{x}_m$ contain the same set of covariates as $\mathbf{x}_y$ with the addition of the set of variables described above that act to identify inflation and tempering.

4. Results

We consider a range of models. Firstly, the reference OP model followed by the $MIOP$ of [9] and the TOP model based on that of [8], but with tempering to the two (hypothesised) inflated outcomes.

Full results of all variants with constant thresholds are presented in Tables 2 and 3. Table 2 contains the parameter estimates and Table 3 contains corresponding marginal effects for reporting *very good* and *excellent* health.[7] We first present, and briefly discuss direct parameter estimates and partial effects for the mean functions $\boldsymbol{\beta}_y$.

Insert Tables 2 and 3 about here

Note that the self-reported health measure is increasing in health, such that positive (negative) coefficients are associated with increasing (decreasing) health. There is reasonable consistency across the estimated coefficients and partial effects for these models with respect to sign and significance, but magnitudes vary across models. Health levels are decreasing with age (at a decreasing rate). None of the models find a statistically significant difference between men and women in reported health at conventional levels of significance (5%), although in absolute terms the OP and $MIOP$ models find the largest effects which are negative (i.e. females are in worse health than males)

[7]Partial effects for other categories are available on request, but omitted here to conserve space. Partial effects were evaluated by differentiating the full probabilistic expression for the final choice outcomes for the respective model with respect to the relevant covariate. These were evaluated at sample means of the covariates and the Delta method used to estimate standard errors. Depending on the model of interest, this overall effect could be composed of various components from middle inflation and HOPIT models.

and significant at the 10% level. Respondents who have a tertiary or high school qualification are in better health than those who have not completed high school. Unemployed respondents experience worse health compared to employed respondents (i.e. the reference group), in general, and respondents not in the labour force, in turn, experience worse health than the unemployed (which may be due, in part, to selection out of the labour market due to ill states of health). Marriage also appears beneficial for health; and migrants report better health compared to respondents born in Australia. In general, these effects appear to be in line with evidence found elsewhere (see, for example, [4]).

Turning to the results for the binary health equation for the middle inflation ($MIOP$) model (equation (15)), it appears that age, tertiary education and unemployment are the main drivers behind the binary choice of reporting *very good* vs *good* health for the inaccurate reporters. For these respondents, age and unemployment have a negative influence on the likelihood of reporting *very good* health, while tertiary education has a positive influence. In the *accurate-reporting* equation which identifies accurate from inaccurate reporters, a range of both the (standard) demographics variables and two of the identifying variables appear to be significant drivers of membership for this latent group. Note that in this equation, a positive coefficient indicates that a respondent is likely to report accurately, while a negative coefficient signifies middle inflation. Females are more likely to inflate compared to males, as are respondents who are married. Respondents who are not in the labour force are more likely to report accurately compared to employed and unemployed respondents. For the identifying variables, the likelihood of reporting accurately increases for individuals who rely on the money they receive from responding to online surveys; this may indicate that they are taking the survey seriously. On the other hand, the likelihood of accurate reporting decreases for individuals who did not have others present when completing the survey, which is similar to the effect found by [9].

Next we focus on tempering equations (equations (20) and (21)) for the TOP model. The first equation represents the propensity to report accurately at *fair* health, as opposed to tempering towards *good* health; while the second represents the propensity to report accurately at *excellent* health, as opposed to tempering towards *very good* health. For both equations, a positive coefficient represents accurate reporting, while a negative coefficient indicates tempering towards the middle categories. It appears that the model is not particularly reliable at distinguishing between the binary choice

of tempering between *fair* and *good* health, as evidenced by the low significance levels in the majority of the parameter estimates for the first tempering equation, with only migrant status being significant (migrants are more likely than Australian-born respondents to temper away from *fair* health). For the second tempering equation, female respondents are more likely to temper away from reporting *excellent* health towards the middle. For the inflation variables, respondents stating that others were not present during the survey were more likely to temper away from *excellent* health, while those that understood all questions where less likely to do so. It is also noted that the model fails to differentiate between μ_3 and μ_4. This could indicate a problem with the specification of this variant of the model.

The coefficients presented in Table 2 and discussed above are not directly comparable across different versions of the models. For example, the parameter estimates in the structural part of the $MIOP$ model are conditional on the inflation equation for misreporting and the binary health equation for the choice between *good* and *very good* for respondents who do misreport into the middle categories. A more direct comparison of the estimates across the models can be obtained through partial effects. These are presented in Table 3 for the partial effect of reporting *very good* and *excellent* health.[8] Interpretation of the partial effects confirms the general relationships outlined above - the probability of reporting *very good* or *excellent* health decreases with age, is lower for the unemployed and respondents not in the labour force, and increases with education and marital status. Females generally are less likely to report *excellent* health, while migrant respondents are more likely to report *very good* health. Estimates of partial effects vary across the models. In general, there is a rather even balance across probabilities of reporting *excellent* and *very good* health observed under the partial effects generated by the OP. However, this balance is not observed under the $MIOP$ or TOP partial effects, where the probabilities in absolute terms are much larger for *very good* health compared to *excellent* health, reflecting inflation towards the middle.

The bottom panel of Table 2 presents Vuong non-nested test statistics for the various models, where a large positive value favours the null, and a large negative value favours the alternative. Both the $MIOP$ and the TOP are

[8] All partial effects are obtained by differentiating the respective probabilistic outcomes with respect to the variable of interest.

favoured over the OP model, suggesting that middle-inflation bias is present in our measure of self-assessed health. The test cannot distinguish between the $MIOP$ and the TOP model (i.e. the final row); however information criteria (not shown) favour the $MIOP$ over TOP across four measures (BIC, AIC, HQIC and CAIC).[9]

Results from the $MIOP$ and TOP models display evidence of reporting heterogeneity in the form of middle inflation.[10] We next consider general reporting heterogeneity existing across the thresholds of the $HOPIT$ model. Structural parameter estimates, $\boldsymbol{\beta}_y$ are provided in Table 4, boundary coefficients, $\boldsymbol{\gamma}_\mathbf{j}$ are presented in Table 5, and partial effects are presented in Table 6. Age, educational status, marital status and migrant status are predictors of reporting behaviour in the first boundary equation ($j = 0$); age, unemployment and migrant status in the second boundary ($j = 1$), while age, education, employment status and marital status are predictors in the fourth boundary ($j = 3$). Broadly, the results indicate that older respondents tend make use of the extreme categories with some individuals down reporting their level of health by making greater use of the categories *poor* and *fair* for a given underlying latent value of health and other over reporting by making greater use of the *excellent* health category compared to younger counterparts. This is observed through the positive and significant (at the 1% level) coefficients in the first ($j = 0$) and second ($j = 1$) threshold and the significant negative coefficient in the forth ($j = 3$) threshold. Educated respondents also tend to down report a given level of underlying health in comparison to less educated counterparts, this can be seen from the positive and significant coefficients in the first and fourth thresholds. Married respondents were less likely to make use of the *poor* category compared to those who were not, while divorced or widowed respondents were least likely to make use of the *excellent* category compared to married and single respondents.

Insert Tables 4 and 5 about here

In general the sign and significance of the coefficient estimates in the structural part of the $HOPIT$ model follow those of the OP with the exception of age effects (which becomes insignificant; but note that the marginal

[9]While information criteria allow a ranking of models, we use the Voung statistic as this provides a direct test of the null hypothesis.

[10]Both the $MIOP$ and TOP models and their vignette variants are robust to including one or two of the three identifying variables. These results are available on request.

effects are significant and consistent with the OP model) and divorced or widowed respondents (where the negative effect becomes significant at the 5% level under the $HOPIT$). The role of marriage and unemployment on health is lesser in the $HOPIT$ compared to the OP model. The effects of education and particularly migrant status, however, are greater. The differences in estimates illustrate the role of that DIF plays in the structural parameters.

The results from the middle inflation models and the $HOPIT$ model controlling for DIF are strongly suggestive of reporting behaviour in self-reports of health. We next combine these models to control for both middle inflation and general forms of reporting behaviour. These models, outlined in section 2.6 are termed $V/MIOP$ and V/TOP respectively when the $HOPIT$ (or vignette) approach is combined with the idea of middle inflation or tempered inflation respectively. Results of these models are presented in columns alongside the results of the $HOPIT$ model in Table 4.

The effects of education (both tertiary and year12) and migrant status are larger in absolute terms in both the $V/MIOP$ and V/TOP models compared with the $HOPIT$ model, while the effect of employment status (both unemployment and not in the labour force) and marriage is lower. The inflation model for the $V/MIOP$ and tempering equations for V/TOP contain significant effects for one or more of the identifying variables. For example, respondents who reported that no-one else was present at the time of completing the survey were again more likely to inflate their responses towards the middle in the case of the $V/MIOP$, or temper away from the *excellent* category towards *very good* health under the V/TOP. On the whole, the direction and significance of these effects follow a similar pattern to those for the corresponding $MIOP$ and TOP models of Table 2, with the exception of covariate for having no-one present in first tempering equation, which was positive under the $V/MIOP$ variant, albiet, insignificant at conventional levels. Finally, the Vuong non-nested tests are presented at the bottom of Table 4. Once again, these tests favour the $V/MIOP$ and V/TOP models over the $HOPIT$ model, signifying that middle-inflation bias is present; though the tests are unable to distinguish between the $V/MIOP$ and V/TOP models (information criteria favour the $V/MIOP$ over V/TOP).

Insert Table 6 about here

Partial effects for reporting *very good* and *good* are presented in Table 6.[11] Focussing first on the $HOPIT$ estimates, note that we no longer observe the balance that was present for the OP model across the probabilities of reporting *very good* and *excellent* health, reflecting the effects of the heterogeneous boundary parameters. This is particularly so across variables for which the estimated parameters of the j = 3 boundary (i.e. the boundary between *very good* and *excellent*) are relatively large in magnitude. For the $V/MIOP$ and V/TOP models we once again see that the probabilities of reporting *very good* health are much larger in absolute terms than the probabilities of reporting *excellent* health, though they are somewhat mediated by the effects of the heterogeneous boundary equations.

4.1. Summary measures

We now turn to summary measures, as represented by partial average predicted probabilities. When correcting for reporting heterogeneity it is intuitive to consider the outcome probabilities "purged" of reporting heterogeneity and how these compare to the estimated probabilities from the standard ordered probit model. Table 7 contains the orderd probit probabilites, estimated average probabilities for the $HOPIT$ model accounting for DIF, together with estimated average probabilities purged of variations in reporting behaviour across models that consider both types of reporting heterogeneity ($V/MIOP$ and V/TOP). The bottom panel expresses estimates from the ordered probit model relative to these (purged) probabilities (averaged over the sample). To be explicit, we use the term "purged" to represent probabilities where any reporting bias has been effectively removed. In the $HOPIT$ this simply involves evaluating the overall probabilistic expressions for the outcomes, as the use of the vignettes should have removed any DIF bias. For the $TOP/MIOP$ variants, this involves evaluating *just* the ordered probit components of the overall probabilities: that is, the parts of the model unaffected by the assumed reporting behaviour. For the V/TOP and variants $V/MIOP$ variants, both of these apply.

Substantive variation is evident across the estimated probabilities. Compared to the ordered probit results, accounting for DIF only ($HOPIT$) leads to an increae in estimated probabilities in the excellect (by 8%) and fair categories (by 24%). The remaining categories see a decrease in average esti-

[11] Implicilty in the vignettes models, these will be 'purged' of DIF

mated probabilities. Overall, it appears that between 17% (in the case of the *V/TOP*) and 22% (for the *V/MIOP*) of responses in the *good* and *very good* categories were artificially inflated. After purging of reporting behaviour, the proportion of respondents in the *good* category decreased by 22% for the *V/MIOP*, and 16% for the *V/TOP*; while the proportion in the *very good* category declined by 23% for the *V/MIOP*, and 18% for the *V/TOP* model. The proportion of respondents in the categories for *excellent* and *fair* health increased by 52% and 41% for the *V/MIOP* model, and by 48% and 34% for the *V/TOP* model, respectively. The proportion of respondents in the *poor* category of the *V/MIOP* model also increased by 39%.

Insert Table 7 about here

5. Conclusions

This paper considers the analysis of self-reports of health, with a particular focus on measurement error brought about through differential item functioning, DIF, (the use of different thresholds that separate the reported levels of self-assessed health) and the adoption of a "box-ticking strategy " leading to inflation of the middle/middle-right categories of the health variable. Employing the methods of anchoring vignettes within the $HOPIT$ model to identify DIF, and two new approaches to identify so-called middle inflation bias, we find compelling evidence for the existence of both types of reporting behaviour. We then develop a model that combines the two approaches and again

find evidence that middle inflation exists once general reporting bias has been controlled for by anchoring against vignettes. After adjusting for these apparent nuances in reporting behaviour, in our particular example we find that health levels on average are rather different than the raw self- reported data suggest. For example, in our preferred specification, there is evidence that the observed categories of very good and good health are overestimated by around 22% each. When adjusting for reporting behaviour, we find a much higher proportion of respondents in the neighbouring categories of excellent and fair health.

The use of anchoring vignettes has gained popularity in the social sciences as a means to anchor self-reported data to some common scale to increase cross-respondent comparability. For example, Kapteyn et al., (2007) [11] apply the HOPIT approach to anchor self-reports of work disability in The

Netherlands to the scale adopted by Americans when drawing inference on comparable levels of underlying disability across the two countries. While the approach appears useful in identifying and correcting for general forms of reporting behaviour observed across the levels of the underlying construct of study (for example, self-assessed health), it does so by linking DIF to observed levels of respondent characteristics. In applications these tend to be socio-demographic characteristics such as age, gender, education, and income, together with the wider social and cultural context of area or nationality of respondent. These measures may exclude more nuanced reporting behaviour brought about by a simple box-ticking approach that leads to the artificial inflation of specific categories. In the case of subjective health, this type of reporting appears to be supported by the observation that the distribution of self-reported health is most often bunched around the middle categories of good and very good, suggesting a more favourable distribution of health than might be inferred from more objective measures. Identifying such behaviour is likely to be best achieved through exclusion restrictions using variables related to the implementation of the survey instrument and specific to the circumstances of the individual when responding to subjective questions (both widely available in survey data). Our results confirm that both types of reporting behaviour exist in self-reports of health. While this presents challenges for researchers wishing to analyse self-reported health questions, the general finding is likely to have similar implications for other self-reported variables commonly observed in survey data.

References

[1] S. Boes and R. Winkelmann. Ordered response models. *AStA Advances in Statistical Analysis*, 90(1):167–181, 2006.

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

[3] B. Burstrm and P. Fredlund. 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*, 55:836–840, 2001.

[4] P. Contoyannis, A. Jones, and N. Rice. The dynamics of health in the british household panel survey. *Journal of Applied Econometrics*, 19:473–503, 2004.

[5] T. F. Crossley and S. Kennedy. The reliability of self-assessed health status. *Journal of Health Economics*, 21(4):643–658, 2002.

[6] J. Currie and B.C. Madrian. Health, health insurance and the labour market. In O.C. Ashenfelter and D. Card, editors, *Handbook of Labour Economics*, pages 3309–3416. Amsterdam: Elsevier Science Publishers BV, 1999.

[7] W. Greene and D. Hensher. *Modeling Ordered Choices*. Cambridge University Press, 2010.

[8] W. Greene, M Gillman, M.N. Harris, and C. Spencer. The tempered ordered probit (top) model with an application to monetary policy. Discussion Paper Series 2013-10, Department of Economics, Loughborough University, 2013.

[9] W. Greene, M.N. Harris, and B. Hollingsworth. Inflated responses in measures of self-assessed health. Working Paper EC-14-12, Stern Business School, New York University, 2014.

[10] W. Greene, M.N. Harris, B Hollingsworth, and P. Maitra. A latent class model for obesity. *Economics Letters*, 123:1–5, 2014.

[11] Arie Kapteyn, James P Smith, and Arthur Van Soest. Vignettes and self-reports of work disability in the united states and the netherlands. *The American Economic Review*, pages 461–473, 2007.

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

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

[14] S. Pudney and M. Shields. Gender, race, pay and promotion in the british nursing profession: estimation of a generalized ordered probit model. *Journal of Applied Econometrics*, 15:367399, 2000.

[15] J. Terza. Ordered probit: A generlization. *Communications in Statistics - A. Theory and Methods*, 14:1–11, 1985.

6. Appendix A

Anchoring vignettes for self-assessed health

(Note that vignettes were gender specific)

Vignette 1:

Rob (Rebecca) is able to walk distances of up to 500 metres without any problems but feels puffed and tired after walking one kilometre or walking up more than one flight of stairs. He (she) is able to wash, dress and groom himself/herself, but it requires some effort due to an injury from an accident one year ago. His (her) injury causes him (her) to stay home from work or social activities about once a month. Rob (Rebecca) feels some stiffness and pain in his (her) right shoulder most days however his (her) symptoms are usually relieved with low doses of medication, stretching and massage. He (she) feels happy and enjoys things like hobbies or social activities around half of the time. The rest of the time he (she) worries about the future and feels depressed a couple of days a month.

Vignette 2:

Chris (Christine) is suffering from an injury which causes him (her) a considerable amount of pain. He (she) can walk up to a distance of 50 metres without any assistance, but struggles to walk up and down stairs. He (she) can wash his (her) face and comb his (her) hair, but has difficulty washing his (her) whole body without help. He (she) needs assistance with putting clothes on the lower half of his (her) body. Since having the injury Chris (Christine) can no longer cook or clean the house himself (herself), and needs someone to do the grocery shopping for him (her). The injury has caused him (her) to experience back pain every day and he (she) is unable to stand or sit for more than half an hour at a time. He (she) is depressed nearly every day and feels hopeless. He (she) also has a low self-esteem and feels that he (she) has become a burden.

Vignette 3:

Kevin (Heather) walks for one to two kilometres and climbs three flights of stairs every day without tiring. He (she) keeps himself neat and tidy and showers and dresses himself each morning in under 15 minutes. He (she) works in an office and misses work one or two days per year due to illness. Kevin (Heather) has a headache once every two months that is relieved by taking over-the-counter pain medication. He (she) remains happy and cheerful most of the time, but once a week feels worried about things at work. He (she) feels very sad once a year but is able to come out of this mood within a few hours.

Table 1: Regression sample descriptive statistics

Variable	*Mean*	*Std Dev*	*Variance*	*Minimum*	*Maximum*
N = 5034					
SAH	2.340	1.047	1.096	0	4
SAH excellent	0.128	0.334	0.112	0	1
SAH very good	0.339	0.474	0.224	0	1
SAH good	0.330	0.470	0.221	0	1
SAH fair	0.151	0.358	0.128	0	1
SAH poor	0.052	0.223	0.050	0	1
Vignette 1	3.132	0.873	0.761	0	4
Vignette 2	1.442	0.815	0.664	0	4
Vignette 3	0.361	0.784	0.615	0	4
Age/100	4.139	1.328	1.764	1.8	6.5
$(\text{Age}/100)^2$	18.90	11.19	125.29	3.24	42.25
Female	0.517	0.500	0.250	0	1
Tertiary	0.694	0.461	0.212	0	1
Year 12 or certificate/diploma	0.149	0.356	0.127	0	1
Unemployed	0.104	0.305	0.093	0	1
Not in labour force	0.144	0.351	0.123	0	1
Married	0.589	0.492	0.242	0	1
Divorced/Separated/Widowed	0.112	0.315	0.099	0	1
Migrant	0.246	0.430	0.185	0	1
No one else present	0.915	0.279	0.078	0	1
Understood all questions	0.835	0.371	0.138	0	1
Money received	0.296	0.457	0.208	0	1

Table 2: Regression results; OP, MIOP and TOP

	OP		*MIOP*		*TOP*	
Age/100	-0.443***	(0.081)	-0.44***	(0.112)	-0.482***	(0.088)
(Age/100)^2	0.03***	(0.009)	0.024*	(0.013)	0.035***	(0.01)
Female	-0.06*	(0.031)	-0.081*	(0.042)	-0.023	(0.033)
Tertiary	0.248***	(0.043)	0.231***	(0.06)	0.272***	(0.046)
Year 12 or certificate/diploma	0.27***	(0.055)	0.277***	(0.075)	0.27***	(0.058)
Unemployed	-0.366***	(0.05)	-0.362***	(0.07)	-0.407***	(0.054)
Not in labour force	-0.488***	(0.045)	-0.521***	(0.062)	-0.518***	(0.047)
Married	0.235***	(0.037)	0.277***	(0.051)	0.261***	(0.041)
Divorced/Separated/Widowed	-0.016	(0.057)	-0.001	(0.079)	0.001	(0.06)
Migrant	0.113***	(0.035)	0.155***	(0.051)	0.108***	(0.038)
μ_1	-2.76***	(0.164)	-2.626***	(0.224)	-2.789***	(0.181)
μ_2	-1.906***	(0.162)	-1.592***	(0.234)	-1.903***	(0.179)
μ_3	-0.929***	(0.161)	-0.95***	(0.249)	-0.96***	(0.178)
μ_4	0.18	(0.161)	-0.311	(0.266)	-0.96***	(0.178)
			Binary Health Equation		Tempering From *fair* health	
Constant			0.869**	(0.412)	1.064	(2.891)
Age/100			-0.443**	(0.206)	0.538	(1.198)
(Age/100)^2			0.041*	(0.024)	-0.021	(0.141)
Female			-0.064	(0.082)	0.816	(0.719)
Tertiary			0.376***	(0.123)	-0.219	(0.736)
Year 12 or certificate/diploma			0.265*	(0.146)	-0.509	(0.409)
Unemployed			-0.492***	(0.158)	-0.361	(0.802)
Not in labour force			-0.142	(0.168)	0.359	(0.994)
Married			0.068	(0.121)	-0.797	(0.609)
Divorced/Separated/Widowed			-0.046	(0.155)	0.609	(1.523)
Migrant			0.025	(0.086)	-1.228***	(0.001)
No one else present					-0.378	(1.167)
Understood all questions					-0.34	(0.889)
Money received					1.28	(1.08)
			"Accurate" reporting eqn		Tempering From *excellent* health	
Constant			0.19	(0.386)	-0.144	(0.296)
Age/100			0.241	(0.186)	-0.006	(0.154)
(Age/100)^2			-0.023	(0.021)	-0.011	(0.018)
Female			-0.172**	(0.069)	-0.144**	(0.058)
Tertiary			-0.12	(0.097)	-0.014	(0.096)
Year 12 or certificate/diploma			-0.096	(0.122)	0.072	(0.113)
Unemployed			0.22*	(0.121)	0.022	(0.104)
Not in labour force			0.436***	(0.107)	-0.035	(0.099)
Married			-0.283***	(0.088)	-0.049	(0.07)
Divorced/Separated/Widowed			-0.101	(0.133)	-0.084	(0.125)
Migrant			-0.132*	(0.077)	0.002	(0.064)
No one else present			-0.401**	(0.165)	-0.486***	(0.091)
Understood all questions			0.1	(0.091)	0.303***	(0.084)
Money received			0.265**	(0.107)	0.052	(0.066)
Vuong non-nested tests:						
H_0 : OP; H_A : MIOP				-4.500		
H_0 : OP; H_A : TOP				-4.602		
H_0 : MIOP; H_A : TOP				-0.112		

Table 3: Partial effects

	OP		MIOP		TOP	
	V Good	Excellent	V Good	Excellent	V Good	Excellent
Age/100	-0.09***	-0.086***	-0.127***	-0.055**	-0.142***	-0.049**
	(0.017)	(0.016)	(0.037)	(0.025)	(0.035)	(0.025)
$(Age/100)^2$	0.006***	0.006***	0.011**	0.002	0.012***	0.002
	(0.002)	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)
Female	-0.012*	-0.012*	0.004	-0.028***	0.015	-0.024**
	(0.006)	(0.006)	(0.014)	(0.009)	(0.013)	(0.009)
Tertiary	0.051***	0.048***	0.089***	0.029**	0.083***	0.025*
	(0.009)	(0.008)	(0.022)	(0.015)	(0.02)	(0.015)
Year 12 or certificate/diploma	0.055***	0.052***	0.071***	0.039	0.069***	0.038**
	(0.011)	(0.011)	(0.027)	(0.018)	(0.024)	(0.018)
Unemployed	-0.075***	-0.071***	-0.128***	-0.044***	-0.124***	-0.038**
	(0.011)	(0.01)	(0.027)	(0.016)	(0.022)	(0.016)
Not in labour force	-0.01***	-0.094***	-0.105***	-0.053***	-0.148***	-0.057***
	(0.01)	(0.009)	(0.023)	(0.016)	(0.02)	(0.015)
Married	0.048***	0.045***	0.061***	0.024**	0.085***	0.019*
	(0.008)	(0.007)	(0.017)	(0.011)	(0.016)	(0.011)
Divorced/Separated/Widowed	-0.003	-0.003	0.004	-0.008	0.013	-0.012
	(0.012)	(0.011)	(0.027)	(0.019)	(0.025)	(0.019)
Migrant	0.023***	0.022***	0.029*	0.016	0.032**	0.011
	(0.007)	(0.007)	(0.015)	(0.01)	(0.015)	(0.01)
No one else present			0.046***	-0.033**	0.072***	-0.072***
			(0.017)	(0.013)	(0.014)	(0.014)
Understood all questions			-0.011	0.008	-0.045***	0.045***
			(0.01)	(0.008)	(0.012)	(0.012)
Money received			-0.03*	0.021***	-0.008	0.008
			(0.018)	(0.007)	(0.01)	(0.01)

Table 4: Regression results; HOPIT, V/MIOP and V/TOP

	HOPIT		*V/MIOP*		*V/TOP*	
Constant	2.451***	(0.212)	2.267***	(0.235)	2.395***	(0.218)
Age/100	-0.108	(0.096)	-0.189*	(0.113)	-0.202*	(0.104)
(Age/100)^2	0.004	(0.011)	0.007	(0.013)	0.012	(0.012)
Female	-0.055	(0.036)	-0.079*	(0.042)	-0.012	(0.04)
Tertiary	0.304***	(0.051)	0.305***	(0.061)	0.3***	(0.055)
Year 12 or certificate/diploma	0.353***	(0.065)	0.356***	(0.076)	0.339***	(0.070)
Unemployed	-0.278***	(0.06)	-0.279***	(0.068)	-0.327***	(0.066)
Not in labour force	-0.472***	(0.053)	-0.511***	(0.063)	-0.438***	(0.058)
Married	0.116***	(0.044)	0.167***	(0.051)	0.139***	(0.048)
Divorced/Separated/Widowed	-0.137**	(0.068)	-0.066	(0.08)	-0.109	(0.072)
Migrant	0.263***	(0.042)	0.257***	(0.05)	0.236***	(0.047)
			Binary Health Equation		Tempering From *fair* health	
Constant			0.744	(0.538)	0.498	(1.008)
Age/100			-0.374	(0.273)	-0.034	(0.476)
(Age/100)^2			0.036	(0.032)	0.005	(0.053)
Female			-0.039	(0.102)	0.194	(0.181)
Tertiary			0.31**	(0.154)	-0.094	(0.243)
Year 12 or certificate/diploma			0.259	(0.205)	0.09	(0.368)
Unemployed			-0.58**	(0.258)	0.131	(0.284)
Not in labour force			0.012	(0.181)	0.939	(0.657)
Married			-0.015	(0.136)	-0.178	(0.234)
Divorced/Separated/Widowed			-0.166	(0.235)	0.021	(0.304)
Migrant			-0.006	(0.105)	-0.304	(0.19)
No one else present					0.223	(0.309)
Understood all questions					-0.068	(0.209)
Money received					0.151	(0.18)
			"Accurate" reporting eqn		Tempering equation From *Excellent* health	
Constant			0.341	(0.491)	0.398	(0.785)
Age/100			0.302	(0.241)	0.344	(0.413)
(Age/100)^2			-0.027	(0.028)	-0.035	(0.05)
Female			-0.223**	(0.09)	-0.388**	(0.159)
Tertiary			-0.048	(0.135)	-0.029	(0.269)
Year 12 or certificate/diploma			0.072	(0.178)	0.146	(0.325)
Unemployed			0.362*	(0.196)	0.615	(0.461)
Not in labour force			0.381***	(0.144)	0.237	(0.267)
Married			-0.28**	(0.114)	-0.26	(0.19)
Divorced/Separated/Widowed			0.009	(0.197)	-0.092	(0.4)
Migrant			-0.229**	(0.098)	-0.162	(0.172)
No one else present			-0.554***	(0.212)	-0.907***	(0.295)
Understood all questions			0.127	(0.109)	0.518***	(0.17)
Money received			0.346***	(0.106)	0.159	(0.173)
Vuong non-nested tests:						
H_0 : HOPIT; H_A : V/MIOP				-5.578		
H_0 : HOPIT; H_A : V/TOP				-5.628		
H_0 : V/MIOP; H_A : V/TOP				0.523		

Table 5: Regression results; HOPIT boundary parameters

	HOPIT		*V/MIOP*		*V/TOP*	
j=0 boundary						
Age/100	0.178**	(0.071)	0.128**	(0.059)	0.139**	(0.063)
$(\text{Age}/100)^2$	-0.014*	(0.008)	-0.011	(0.007)	-0.011	(0.007)
Female	0.019	(0.027)	0.026	(0.022)	0.026	(0.024)
Tertiary	0.082**	(0.037)	0.08**	(0.032)	0.084**	(0.033)
Year 12 or certificate/diploma	0.066	(0.048)	0.066	(0.04)	0.068	(0.042)
Unemployed	-0.011	(0.043)	-0.03	(0.037)	-0.031	(0.039)
Not in labour force	0.031	(0.038)	-0.011	(0.032)	0.011	(0.034)
Married	-0.159***	(0.032)	-0.112***	(0.027)	-0.127***	(0.029)
Divorced/Separated/Widowed	-0.086*	(0.049)	-0.067	(0.041)	-0.075*	(0.044)
Migrant	0.093***	(0.031)	0.088***	(0.026)	0.089***	(0.027)
j=1 boundary						
Constant	-0.315**	(0.142)	-0.472***	(0.143)	-0.424***	(0.144)
Age/100	0.15**	(0.07)	0.156**	(0.07)	0.168**	(0.071)
$(\text{Age}/100)^2$	-0.011	(0.008)	-0.011	(0.008)	-0.012	(0.008)
Female	-0.044*	(0.027)	-0.035	(0.026)	-0.05*	(0.027)
Tertiary	-0.022	(0.035)	-0.038	(0.035)	-0.033	(0.037)
Year 12 or certificate/diploma	-0.021	(0.046)	-0.049	(0.046)	-0.044	(0.048)
Unemployed	0.085**	(0.041)	0.092**	(0.042)	0.094**	(0.043)
Not in labour force	0.033	(0.037)	0.053	(0.037)	0.018	(0.041)
Married	-0.011	(0.032)	-0.017	(0.032)	-0.02	(0.033)
Divorced/Separated/Widowed	-0.075	(0.047)	-0.074	(0.047)	-0.075	(0.049)
Migrant	0.059**	(0.03)	0.06**	(0.03)	0.064**	(0.03)
j=2 boundary						
Constant	-0.192	(0.137)	-0.43***	(0.166)	-0.294*	(0.16)
Age/100	0.048	(0.07)	0.057	(0.085)	0.021	(0.083)
$(\text{Age}/100)^2$	-0.005	(0.008)	-0.005	(0.01)	-0.003	(0.01)
Female	0.03	(0.027)	0.009	(0.033)	0.055*	(0.033)
Tertiary	-0.06*	(0.036)	-0.054	(0.045)	-0.062	(0.044)
Year 12 or certificate/diploma	0.023	(0.046)	0.047	(0.056)	0.038	(0.055)
Unemployed	0.059	(0.042)	0.061	(0.053)	0.057	(0.053)
Not in labour force	-0.037	(0.039)	0.031	(0.047)	0.018	(0.049)
Married	0.05	(0.032)	0.009	(0.04)	0.049	(0.04)
Divorced/Separated/Widowed	0.029	(0.049)	0.011	(0.061)	0.036	(0.059)
Migrant	-0.005	(0.031)	-0.043	(0.039)	-0.031	(0.038)
j=3 boundary						
Constant	0.153	(0.135)	-0.132	(0.173)	-0.119	(0.168)
Age/100	-0.144**	(0.069)	-0.114	(0.088)	-0.096	(0.086)
$(\text{Age}/100)^2$	0.018**	(0.008)	0.014	(0.01)	0.014	(0.01)
Female	0.012	(0.027)	-0.019	(0.034)	-0.037	(0.033)
Tertiary	0.125***	(0.04)	0.113**	(0.048)	0.117**	(0.048)
Year 12 or certificate/diploma	0.092*	(0.05)	0.099	(0.061)	0.1*	(0.061)
Unemployed	-0.091*	(0.047)	-0.022	(0.057)	-0.026	(0.057)
Not in labour force	-0.107**	(0.042)	-0.111**	(0.052)	-0.082	(0.051)
Married	0.071**	(0.033)	0.066	(0.041)	0.047	(0.04)
Divorced/Separated/Widowed	0.109**	(0.051)	0.143**	(0.061)	0.116*	(0.061)
Migrant	0.004	(0.03)	-0.018	(0.039)	-0.023	(0.039)
Vignettes						
α_1	3.232*	(0.147)	2.63***	(0.136)	2.81***	(0.136)
α_2	1.428***	(0.142)	1.172***	(0.122)	1.255***	(0.127)
α_3	-0.117	(0.141)	-0.121	(0.118)	-0.124	(0.125)
$1/\sigma_v$	1.017***	(0.016)	1.241***	(0.033)	1.164***	(0.024)

Table 6: Partial effects

	HOPIT		V/MIOP		V/TOP	
	V Good	Excellent	V Good	Excellent	V Good	Excellent
Age/100	-0.123***	-0.071***	-0.134***	-0.056**	-0.15***	-0.058**
	(0.025)	(0.02)	(0.035)	(0.025)	(0.035)	(0.025)
$(\text{Age}/100)^2$	0.01***	0.004	0.012***	0.002	0.013***	0.003
	(0.003)	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)
Female	-0.008	-0.014*	0.003	-0.025***	0.011	-0.024***
	(0.01)	(0.008)	(0.013)	(0.009)	(0.013)	(0.009)
Tertiary	0.083***	0.036***	0.086***	0.035**	0.084***	0.034**
	(0.014)	(0.011)	(0.02)	(0.015)	(0.02)	(0.016)
Year 12 or certificate/diploma	0.074***	0.041***	0.068***	0.045**	0.068***	0.045**
	(0.017)	(0.014)	(0.026)	(0.018)	(0.025)	(0.019)
Unemployed	-0.097***	-0.067***	-0.139***	-0.043**	-0.136***	-0.036
	(0.016)	(0.013)	(0.034)	(0.017)	(0.03)	(0.023)
Not in labour force	-0.118***	-0.082***	-0.1***	-0.064***	-0.132***	-0.059***
	(0.015)	(0.012)	(0.021)	(0.016)	(0.021)	(0.016)
Married	0.06***	0.035***	0.058***	0.025**	0.077***	0.021*
	(0.012)	(0.009)	(0.016)	(0.011)	(0.016)	(0.011)
Divorced/Separated/Widowed	0.024	-0.023	0.006	-0.009	0.026	-0.021
	(0.018)	(0.015)	(0.029)	(0.02)	(0.029)	(0.023)
Migrant	0.022***	0.022***	0.029**	0.014	0.03**	0.013
	(0.011)	(0.009)	(0.014)	(0.01)	(0.014)	(0.01)
No one else present			0.046***	-0.035***	0.056***	-0.056***
			(0.017)	(0.013)	(0.018)	(0.018)
Understood all questions			-0.01	0.008	-0.032***	0.032***
			(0.009)	(0.007)	(0.011)	(0.011)
Money received			-0.029***	0.022***	-0.01	0.01
			(0.009)	(0.006)	(0.011)	(0.011)

Table 7: SAH probabilities purged of inaccurate reporting

	Predicted probabilities		Probabilities purged of inaccurate reporting	
SAH	*Ordered Probit*	*HOPIT*	*V/MIOP*	*V/TOP*
SAH excellent	0.128	0.139	0.195	0.189
SAH very good	0.340	0.318	0.261	0.280
SAH good	0.330	0.319	0.259	0.276
SAH fair	0.151	0.187	0.213	0.203
SAH poor	0.052	0.038	0.072	0.053
		Probabilities relative to Ordered Probit		
SAH		*HOPIT*	*V/MIOP*	*V/TOP*
SAH excellent		1.086	1.524	1.479
SAH very good		0.935	0.768	0.824
SAH good		0.967	0.785	0.836
SAH fair		1.238	1.412	1.344
SAH poor		0.731	1.388	1.010