Economic Modelling 124 (2023) 106277

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

Reporting heterogeneity in modeling self-assessed survey outcomes

William Greene [a,1], Mark N. Harris [b,*], Rachel Knott [c], Nigel Rice [d]

[a] Stern School of Business, New York University, USA
[b] School of Accounting, Economics and Finance, Curtin University, Australia
[c] Centre for Health Economics, Monash University, Australia
[d] Centre for Health Economics and Department of Economic and Related Studies, University of York, UK

ARTICLE INFO

Keywords:
Reporting heterogeneity
On-line data collection
Differential item functioning
Inflated ordered outcomes
Anchoring vignettes
Self-reports
Self-assessed health

ABSTRACT

The analysis of self-reports will be severely biased if they are subject to reporting heterogeneity. Moreover, there are several types of such heterogeneity, which have all shown to be widespread in the literature. We consider two predominant types of reporting heterogeneity: *differential item functioning* and *middle inflation bias*. We consider and extend approaches for adjusting for each type of reporting heterogeneity in isolation and propose models that allow for both types in combination. Monte Carlo experiments favor more complex models (that allow for reporting heterogeneity), even when the underlying data generating process is of a simpler form. The results suggest that failure to account for these nuances will lead to erroneous inference concerning the analysis of self-reported data. We apply these new methods to the important area of self-reported health outcomes.

1. Introduction

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 ordered categories. For example the ubiquitous self-assessed general health measure asks respondents to rate their health using a 5-point scale. Available response categories typically consist of *poor*, *fair*, *good*, *very good* and *excellent* health.[2] There are compelling reasons for such measures being a staple feature of household surveys, most notably the relative ease and low cost of data collection. These measures contain valuable information from which to infer differences across individuals, socio-economic groups or countries, and feature strongly in empirical social science survey research.

Although widely used, self-assessments are subject to various forms of *reporting behavior*. Responses to these types of questions are often open to subjective interpretation. Even where respondents are facing a fixed and known level of the construct under consideration, their respective assessments can vary. Accordingly, responses will reflect both the objective reality and a respondent's interpretation of the subjective scale.

Subjectivity in self-reported outcomes will not always be of concern to the researcher. For constructs such as pain or satisfaction, an individual's perception is likely to be of greater relevance than attempts at more objective measures, and accordingly, information elicited on a person's subjective view will be of primary interest. That is, reporting behavior in self-assessments will be less relevant where the measured variable reflects the underlying construct upon which we wish to base policy recommendations. Where differential reporting behavior exists and is purely random and confined to the outcome of interest it could also be of limited concern, for example, in linear models where precision but not consistency of estimates is affected. However, this does not readily extend to non-linear models typically used to model many survey self-assessments (see, for example, Hausman et al. 1998, Hausman 2001).

Often researchers are interested in comparing levels of outcomes across sub-groups of the population, for example, stratified by socio-economic status or gender (Etilé and Milcent, 2006; Bago d'Uva et al., 2008; Dowd and Zajacova, 2010; Au and Lorgelly, 2014; Davillas

[*] Corresponding author.
E-mail address: mark.harris@curtin.edu.au (M.N. Harris).

[1] The authors are grateful to Bruce Hollingsworth for comments on an earlier version of this paper. We also acknowledge extremely useful comments from the Editor, Prof. Sushanta K. Mallick, an anonymous Associate Editor and two anonymous referees. The usual caveats apply, and any remaining errors are the authors' responsibility.

[2] The wording of the questions and available response categories vary across surveys. For example, the UK Household Longitudinal Survey (UKHLS, 2022) uses the categories *very poor*, *poor*, *fair*, *good*, and *excellent* health.

https://doi.org/10.1016/j.econmod.2023.106277
Received 12 August 2022; Received in revised form 9 March 2023; Accepted 12 March 2023
Available online 20 March 2023

et al., 2019), or international comparisons where differences in cultural norms and behaviors might influence reporting styles (for example, Rice et al. 2012). There is often little compelling reason to believe, *a priori*, that self-reported outcomes will be comparable across such groups. For example, self-reported rates of work disability have been found to differ across countries, more so than might be expected on more objective measures of health (Kapteyn et al., 2007). More generally, since ill-health can be seen as a legitimate reason for being outside the labor market, individuals can also seek to justify their work status by down-reporting health. This form of reporting behavior has long been a concern in the literature on the link between health and labor market outcomes (for example, Bound 1991, Kerkhofs and Lindeboom 1995, Black et al. 2017). Evidence of non-random reporting behavior and more general forms of measurement error in self-assessments of health have been reviewed, amongst others, in Currie (2000), Crossley and Kennedy (2002) and Lindeboom (2006).

This paper is concerned with two types of reporting behavior for survey self-assessments on an ordered categorical scale where bias from misreporting is of concern for inference. The first, differential item functioning (DIF), exists where the use of different response benchmarks leads individuals with similar objective health (as in the current example) to report differently—for example, because of different expectations of health, underlying levels of optimism or pain thresholds (King et al., 2004). Note that in the context of self-assessed health this has also been referred to as *cut-point shift* or *index shift* (Sadana et al., 2000; Lindeboom and van Doorslaer, 2004), *state-dependent reporting error* (Kerkhofs and Lindeboom, 1995; Carro and Traferri, 2014), and *scale of reference bias* (Groot, 2000). The second form of reporting behavior is *middle-inflation bias* where for various reasons, individuals concentrate responses in the middle categories of the response scale. In the context of self-assessed health, both of these types of behaviors could lead individuals with identical levels of underlying latent health to respond very differently to a given survey question. Indeed, our explicit contributions to the literature are to generalize existing "structural" models of misreporting and then to also combine these with those approaches where individuals utilize different response benchmarks (DIF), into a convenient single estimation approach.

To address the issue of DIF, recent research has advocated the use of *anchoring vignettes* to detect and adjust for heterogeneity in individuals' reporting scales (Currie and Madrian, 1999; Crossley and Kennedy, 2002; King et al., 2004; Kapteyn et al., 2007; Kristensen and Johansson, 2008; Angelini et al., 2014). Essentially this entails asking individuals to rate one or more of a set of vignettes, or questions about, for example, 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 their own health. Hence, vignettes are aimed at increasing inter-respondent comparability by abstracting reporting behavior 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 mis-measurement in self-reported data. In particular, Greene et al. (2015) consider the distribution of self-assessed health (SAH), collected from a large representative sample of the Australian population (although similar issues are found in comparable surveys across the developed world). The responses to the SAH question are clearly bunched around the middle category and the one immediately to the "right" of this: *good* and *very good* in the range from *poor* to *excellent* health. They argue that this paints a rather rosy picture of the health of the population as compared to many more objective measures (such as obesity rates, exercise rates, widespread levels of elevated cholesterol levels, and so on) which tend to paint a much bleaker picture (Greene et al., 2015). However, clearly this depends on whether individuals have predominantly "inflated" these responses from higher, or lower, levels. They conclude that there is a large, 9% (prior probability) chance that a randomly selected individual will inaccurately report into the two inflated categories; which jumps to over 12% when posterior based probabilities are considered. Moreover, of the 40% (39%) estimated probability for the outcome *good* (*very good*), some 5 (4) percentage points can be attributed to inaccurate reporting. When translated into the number of implied individuals, these are substantive findings.

Similarly, Brown et al. (2021) consider self-reported mental health, and explicitly the construction of the widely used *GHQ-12* instrument for such, which they found to be heavily biased away from lower levels of mental health, which they attribute to mis-reporting. In the context of self-reported drug use, Brown et al. (2018) also consider inaccurate reporting. They find that true participation rates are likely to be around double those that are self-reported for marijuana, speed and cocaine (23%, 8% and 5%, respectively, compared to reported rates of 12%, 3% and 1%). Misreporting rates were also found to vary in proportion to the "hardness" of the drug.

The general approach of Greene et al. (2015) can be extended to provide greater flexibility. Couched in terms of a *latent class model* (McLachlan and Peel, 2000), it essentially hypothesizes that there are inherently only two types of individuals in the population: those who answer the self-assessments accurately and; those who answer inaccurately (picking one of the two "inflated" responses). A more generalized approach would assume a wider range of types of individuals in the population: operationally this will correspond to the number of discrete outcomes available to the respondent (typically five). To allow for the over-representation of individuals in these "middle" responses, it must be that some individuals outside of these categories have erroneously placed themselves into them. This will be due to a number of reasons, for example psychological ones, such as "wanting to fit in" and the avoidance of extreme answers. Or respondents could be employing a "box-ticking" strategy of defaulting to reporting in the middle categories to reduce the time costs of considering conscientiously the appropriateness of each available category. This can be accommodated by considering multiple "inflation" equations. These can be specified in a number of ways. For example, to model inflation from neighboring categories only, therefore tempering from *fair* to *good* health, or *excellent* to *very good* health; or from all available outcome categories (including *poor* health) to the two middle inflated responses. This approach is a natural generalization of Greene et al. (2013) that offers flexibility in modeling inflated outcomes. Thus a significant contribution of this paper is the introduction of several new variants to these so-called "inflation models", as described above.[3]

A potential criticism of the general inflation-model approach described above though, is that some, or all, of the clustering of observations around the middle of the scale, could actually be attributed to heterogeneity in reporting scales as opposed to an "inflation" strategy. Similarly, approaches that use anchoring vignettes alone, could be biased if they erroneously attribute an "inflation" strategy to reporting heterogeneity in response scales. The paper innovates by firstly generalizing the approach of Greene et al. (2015), and then combining all of these potential forms of reporting heterogeneity (DIF and middle inflation) into single estimators. In so-doing, we propose an extremely general estimation approach, that is well-placed to handle multiple forms of survey-based misreporting/misclassification that have been discussed in the literature to-date.

We demonstrate the finite sample performance of existing and proposed models in a set of Monte Carlo experiments. Our empirical application to health outcomes support the notion that respondents, when self-reporting, are susceptible to both general reporting behavior and artificial inflation of certain categories. Failure to account for these nuanced reporting effects leads to erroneous inference of health determinants. While our particular focus is the self-reporting

[3] Similar extensions to *single inflated* outcome models, have also been considered by Brown et al. (2020) and Sirchenko (2020), for example.

of health, it should be noted that the approach is broadly applicable to categorical outcomes used across the social sciences. Indeed, it will be of use whenever the researcher is interested in self-reports on an attitudinal *Likert*-type scale, which are increasingly popular, especially in large-scale population based surveys.

2. Methods

2.1. Differential item functioning and the $HOPIT$ model

A graphical illustration of DIF is presented in the online Appendix. This shows two individuals with identical levels of underlying latent health, asked to rate their health on a 5-point *Likert*-type scale with response categories *excellent*, *very good*, *good*, *fair* and *poor*. How each respondent divides the scale into the response categories is illustrated by the placement of the boundaries or "cut-points". DIF is evident by the differing placement of cut-points across the two respondents. Despite having the same level of latent objective health, respondent 1 reports their health as *fair*, while respondent 2 reports to be in *good* health. The presence of DIF in this example invalidates interpersonal comparability of the general self-assessed health measure.

More formally, assume y^* denotes true underlying health, which is a linear function (in unknown parameters, $\boldsymbol{\beta}$) of observed characteristics $\mathbf{x}$; a standard normal disturbance term, ε_y; and its relationship to certain boundary parameters, μ_j:

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

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

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

where the total number of outcomes is J (in the example above, $J = 5$). To guarantee well-defined probabilities, $\mu_{-1} \leq \mu_1 \cdots \leq \mu_{J-1}$, with $\mu_{-1} = -\infty$ and $\mu_{J-1} = \infty$. Typical ordered response models (for example, the ordered probit, OP) assume that boundary parameters remain constant across individuals. Heterogeneity in reporting scales can be accommodated by individual varying boundary parameters, μ_j. While there are a number of ways to do this (for example, Terza (1985), Pudney and Shields (2000)), many authors adopt a hierarchical ordered probit ($HOPIT$) approach, which specifies boundary parameters as

$$\mu_0 = \mathbf{z}'\boldsymbol{\gamma}_0; \mu_j = \mu_{j-1} + \exp\left(\mathbf{z}'\boldsymbol{\gamma}_j\right); \ldots \tag{3}$$

where the $\exp(.)$ ensures the necessary ordering and identifies γ_j. For any variables that appear in both $\mathbf{x}$ and $\mathbf{z}$, however, the corresponding elements of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}$ are not separately identified without further information (as the first threshold is specified linearly).

King et al. (2004) introduced the use of anchoring vignettes to allow for identification of the $HOPIT$ model. Vignettes offer a method of anchoring individual response scales when used in conjunction with the main self-report of interest. Alongside a self-assessment, respondents are also asked to rate a set of vignettes ($k = 1, \ldots, K$) describing the situations (for example, health states) of hypothetical people. The response scale available to rate the K vignettes is the same $j = 0, \ldots, J-1$ scale used for the self-report. Example vignettes are provided in the Supplementary Materials in the online Appendix. Define the observed response to each $k = 1, \ldots, K$ possible vignette as $y^{(k)}$. The response to the main assessment in Eq. (2) is now $y^{(0)}$. The vignette responses are assumed to be dependent upon unobserved continuous latent measures $y^{(k)^*}$ and embody the mappings

$$y^{(k)} = j \text{ if } \mu_{j-1}^{(k)} < y^{(k)^*} < \mu_j^{(k)} \;\; k = 1, \ldots, K;\; j = 0, \ldots, J-1, \tag{4}$$

with $\mu_{-1}^{(k)} = -\infty$, and $\mu_{J-1}^{(k)} = \infty$. Note that with multiple vignettes, an unobserved individual-specific effect can be included in the specification of the thresholds such that in Eq. (3) $\mu_{i0} = \mathbf{z}_i'\boldsymbol{\gamma}_0 + u_i$ (Kapteyn et al., 2007). The $y^{(k)^*}$'s are assumed to be a function of a constant and random errors

$$y^{(k)^*} = \alpha^{(k)} + \varepsilon^{(k)}, \tag{5}$$

with $\varepsilon^{(k)} \sim N\left(0, (\sigma^{(k)})^2\right)$ and independent of all observed covariates in the model. Eq. (5) follows from an identifying assumption of *vignette equivalence* (VC) - that the underlying level of the construct of interest described by a vignette is perceived by all respondents in the same way and on the same unidimensional scale, except for random error; the alternative being to extend Eq. (5) to include a function of observed characteristics as in Eq. (1). Often the simplifying assumption that the variance is the same across all vignettes, $(\sigma^{(k)})^2 = \sigma^2$, is also imposed.[4] Heterogeneity across the response scales is once more allowed for by specifying the boundaries as a function of threshold variables, $\mathbf{z}$, and having the same form as Eq. (3).

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. Under the assumption of normality, the respective probabilities for each ordered outcome of the self-assessment are

$$\Pr(y = j|\mathbf{x}) = \begin{cases} \Pr(y = 0|\mathbf{x}) = \boldsymbol{\Phi}\left(\mu_0 - \mathbf{x}'\boldsymbol{\beta}\right) \\ \vdots \\ \Pr(y = J-1|\mathbf{x}) = \left[1 - \boldsymbol{\Phi}\left(\mu_{J-2} - \mathbf{x}'\boldsymbol{\beta}\right)\right] \end{cases}, \tag{6}$$

where $\boldsymbol{\Phi}(.)$ denotes the standard normal distribution function evaluated at its argument (note that if the boundary parameters were constant across individuals, these would simply be of the standard OP form). The (log) density for the self-report ($HOPIT$) component, for a random sample of individuals, $i = 1, \ldots, N$, is given by

$$\ln L[HOPIT, \theta] = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} \left[\Pr\left(y_i = j \left| \mathbf{x}_i, \mathbf{z}_i, HOPIT\right.\right)\right], \tag{7}$$

where d_{ij} is a binary indicator equal to one if individual i chooses outcome j, and zero otherwise, $\sum_j d_{ij} = 1$, and $\boldsymbol{\theta}$ is a vector of parameters in the model. In what follows, to simplify the notation, observation subscript i will be omitted and $\boldsymbol{\theta}$ will be implicit in the formulations of log likelihood functions.

For the vignette component and a particular k, we have ordered probabilities

$$\Pr(y_{jk} = j\,|\mathbf{z}) = p_{jk}^{v}(\mathbf{z})$$
$$= \begin{cases} \Pr\left(y^{(k)} = 0\,|\mathbf{z}\right) = \boldsymbol{\Phi}\left(\left[\mu_0^{(k)} - \alpha^{(k)}\right]/\sigma\right), j = 0; \\ \Pr\left(y^{(k)} = 1\,|\mathbf{z}\right) = \left[\boldsymbol{\Phi}\left(\left[\mu_1^{(k)} - \alpha^{(k)}\right]/\sigma\right) - \boldsymbol{\Phi}\left(\left[\mu_0^{(k)} - \alpha^{(k)}\right]/\sigma\right)\right], j = 1; \\ \vdots \\ \Pr\left(y^{(k)} = J-1\,|\mathbf{z}\right) = \left[1 - \boldsymbol{\Phi}\left(\left[\mu_{J-2}^{(k)} - \alpha^{(k)}\right]/\sigma\right)\right], j = J-1 \end{cases} \tag{8}$$

where the $\mu_j^{(k)}$ are of the form of Eq. (3). The (log-)likelihood contribution arising from the vignettes component over $k = 1, \ldots, K$ vignettes is

$$\ln L[V] = \sum_{i=1}^{N} \sum_{k=1}^{K} \ln \sum_{j=0}^{J-1} d_{ij}^{(k)} \Pr\left(y_{ij}^{(k)} = j|\mathbf{z}, V(k)\right), \tag{9}$$

where $d_{ij}^{(k)}$ is now the vignette-specific indicator variable.

The second identifying assumption of the $HOPIT$ approach is *response consistency* (RC) which implies that the boundary parameters are equivalent across the self-report and the K vignettes, such that $\gamma_j^{(k)} \equiv \gamma_j^{(0)}$, $j = 0, \ldots, J-1; k = 1, \ldots, K$. With independence of ε_y

[4] Imposing the assumptions of VE and *response consistency* (see below) allows the parameter σ^2 to be freely estimated given the normalization of scale and location in Eq. (1); see Kapteyn et al. (2007).

and ϵ^k, the overall (log-)likelihood is the sum of these two components

$$\ln L = \ln L[V] + \ln L[HOPIT], \tag{10}$$

where the first term is a function of $[\mu_{ij}, \alpha^{(1)}, \ldots, \alpha^{(K)}, \sigma]$ and the second is a function of $[\mu_{ij}, \gamma_j]$. The two terms are linked through the common boundary parameters μ_{ij}, and so do not factorize into two independent models. Model identification rests on the two underlying assumption of VC and RC (Greene and Hensher, 2010); a simple parametric test for which has recently been proposed by Greene et al. (2020). However, it is not clear how such a test would be applied to the models suggested below which combine this approach for reporting behavior with other middle inflation.

2.2. Middle inflation models of reporting heterogeneity

Recent innovations in the literature (Greene et al., 2015) consider a much more structural form of reporting heterogeneity than that described by the $HOPIT$ model with vignettes. 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. 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* $(j = 4)$, these would correspond to outcomes $j = 2$ and $j = 3$. Reasons as to why some individuals may misreport into these middle categories include, amongst others: a general distrust of surveys; the opportunity cost-of-time; a desire to want to appear more socially acceptable, or to "play it safe". Traditional ordered response models cannot accommodate this phenomenon or test it as a hypothesis.

To account for middle inflation, Greene et al. (2015) propose a middle-inflated ordered probit ($MIOP$) model (illustrated conceptually in the on-line Appendix). Formally, consider a latent variable, r^*, which represents an individual's propensity to report accurately/inaccurately (*i.e.*, *middle inflate*). Let this latent variable be a function of a set of observed covariates, $\mathbf{w}$, with unknown weights λ, and a (standard normal) disturbance term, υ such that

$$r^* = \mathbf{w}'\lambda + \upsilon, \tag{11}$$

with $\upsilon \sim N[0, \sigma_\upsilon^2]$ and σ_υ normalized to one. When this index reaches a critical level (normalized 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

$$\Pr(r = 1|\mathbf{w}) = \Pr\left(r^* > 0|\mathbf{w}\right) = \Phi\left(\mathbf{w}'\lambda\right). \tag{12}$$

Conditional on being in the *non-box-ticking* regime, a standard formation given by Eq. (6) applies; driven by covariates $\mathbf{x}$ with associated weights $\boldsymbol{\beta}$, and for now assuming constant boundary parameters (μ_j) across respondents (as in the traditional OP model). However, for individuals with a box-ticking propensity, they will essentially make a binary choice between the categories *good* and *very good*. This can be determined by a further latent variable of the form

$$m^* = \mathbf{f}'\delta + h, \tag{13}$$

where it is expected that observed covariates $\mathbf{f}$ will be the same as those driving the health equation for accurate reporters; δ are unknown coefficients and; h is a standard normal error term . Again, once the index reaches a threshold value normalized to zero, this triggers the choice of *very good* relative to *good*. Thus under independence of the stochastic elements of the system, joint probabilities of inaccurate and of *good* reporting, $\Pr(r = 0|\mathbf{w})$ and $\Pr(m = 0|\mathbf{f})$ respectively, and of inaccurate and *very good* reporting, $\Pr(r = 0|\mathbf{w})$ and $\Pr(m = 1|\mathbf{f})$ respectively, will be

$$\begin{aligned}\Pr(r = 0, m = 0|\mathbf{w}, \mathbf{f}) &= \Phi\left(-\mathbf{w}'\lambda\right)\Phi\left(\mathbf{f}'\delta\right) \\ \Pr(r = 0, m = 1|\mathbf{w}, \mathbf{f}) &= \Phi\left(-\mathbf{w}'\lambda\right)\Phi\left(-\mathbf{f}'\delta\right).\end{aligned} \tag{14}$$

Accurately reporting respondents will choose freely across the 5-point choice set. The main outcome probabilities conditional on $r = 1$ are given in Eq. (6). Marginal probabilities based on probit models for respondent type, r, and middle inflation outcome, m, are given by Eqs. (12) and (15), respectively. Marginal probabilities for the observed outcomes are determined as follows

$$\begin{aligned}\Pr(y = 0|\mathbf{x}, \mathbf{w}, \mathbf{f}) &= \Pr(r = 1|\mathbf{x}, \mathbf{w}, \mathbf{f})\Pr(y = 0|r = 1, \mathbf{x}, \mathbf{w}, \mathbf{f}) \\ &+ \Pr(r = 0|\mathbf{x}, \mathbf{w}, \mathbf{f})\Pr(y = 0|r = 0, \mathbf{x}, \mathbf{w}, \mathbf{f})\end{aligned}$$

However, $\Pr(y = 0|r = 0, \mathbf{x}, \mathbf{w}, \mathbf{f}) = 0$. Collecting the remaining terms,

$$\Pr(y = 0|\mathbf{x}, \mathbf{w}, \mathbf{f}) = \Phi\left(\mathbf{w}'\lambda\right)\Phi\left(\mu_0 - \mathbf{x}'\boldsymbol{\beta}\right). \tag{15}$$

By the same logic, outcomes $y = 1$ and $y = 4$ arise only when $r = 1$. Thus,

$$\begin{aligned}\Pr(y = 1|\mathbf{x}, \mathbf{w}, \mathbf{f}) &= \Phi\left(\mathbf{w}'\lambda\right)\left[\Phi\left(\mu_1 - \mathbf{x}'\boldsymbol{\beta}\right) - \Phi\left(\mu_0 - \mathbf{x}'\boldsymbol{\beta}\right)\right], \\ \Pr(y = 4|\mathbf{x}, \mathbf{w}, \mathbf{f}) &= \Phi\left(\mathbf{w}'\lambda\right)\left[1 - \Phi\left(\mu_3 - \mathbf{x}'\boldsymbol{\beta}\right)\right].\end{aligned} \tag{16}$$

Outcomes $y = 2$ and $y = 3$ arise from an accurate report or a misreport. Thus,

$$\begin{aligned}\Pr(y = 2|\mathbf{x}, \mathbf{w}, \mathbf{f}) &= \Pr(r = 1|\mathbf{x}, \mathbf{w}, \mathbf{f})\Pr(y = 2|r = 1, \mathbf{x}, \mathbf{w}, \mathbf{f}) \\ &+ \Pr(r = 0|\mathbf{x}, \mathbf{w}, \mathbf{f})\Pr(m = 0|r = 0, \mathbf{x}, \mathbf{w}, \mathbf{f}) \\ &= \Phi\left(\mathbf{w}'\lambda\right)\left[\Phi\left(\mu_2 - \mathbf{x}'\boldsymbol{\beta}\right) - \Phi\left(\mu_1 - \mathbf{x}'\boldsymbol{\beta}\right)\right] \\ &+ \left[1 - \Phi\left(\mathbf{w}'\lambda\right)\right]\left[1 - \Phi\left(\mathbf{f}'\delta\right)\right]. \\ \Pr(y = 3|\mathbf{x}, \mathbf{w}, \mathbf{f}) &= \Pr(r = 1|\mathbf{x}, \mathbf{w}, \mathbf{f})\Pr(y = 3|r = 1, \mathbf{x}, \mathbf{w}, \mathbf{f}) \\ &+ \Pr(r = 0|\mathbf{x}, \mathbf{w}, \mathbf{f})\Pr(m = 1|r = 0, \mathbf{x}, \mathbf{w}, \mathbf{f}) \\ &= \Phi\left(\mathbf{w}'\lambda\right)\left[\Phi\left(\mu_3 - \mathbf{x}'\boldsymbol{\beta}\right) - \Phi\left(\mu_2 - \mathbf{x}'\boldsymbol{\beta}\right)\right] \\ &+ \left[1 - \Phi\left(\mathbf{w}'\lambda\right)\right]\left[\Phi\left(\mathbf{f}'\delta\right)\right].\end{aligned} \tag{17}$$

Once the form of the probabilities, dependent on unknown parameters and observed data, is known, the model can be estimated by maximum likelihood, ML, techniques where the (log-)likelihood function for the middle-inflated ($MIOP$) model is now

$$\ln L[MIOP] = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} \Pr\left(y_i = j|\mathbf{x}_i, \mathbf{f}_i, \mathbf{w}_i, MIOP\right). \tag{18}$$

2.3. Tempered inflation models of reporting heterogeneity

There is an inherent sequencing implicit in the $MIOP$ model: first an individual decides if they have a propensity to report accurately, or not, and conditional on this decision then reports accordingly. However, Greene et al. (2013) consider reversing this implicit ordering using a tempered ordered probit (TOP) model. Importantly, their approach considered only three outcome choices, with hypothesized inflation in the singleton middle outcome only. Here, assume that an individual first has a propensity to translate their notions of their true underlying health, y^*, into one of the five observed categories (where the generalizations to more or less than five categories is implicit). 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. In essence, this is the so-called TOP approach described among others by Brown et al. (2020), Greene et al. (2013) and Sirchenko (2020).

We extend this general approach in several important ways. A relatively simple extension, is to consider *multiple*, that is > 3 outcomes. However, as explained in the case of self-elicited responses on *Likert* scales, "outcome inflation" is unlikely to be evident in a single middle category, but predominantly in two (or more): in the current example, *good* and *very good*. The combination of these factors means that such extensions are at the same time more complicated, but also potentially more flexible. With such multiple (> 3) outcomes, an “extreme outcome” could be any outside of the inflated middle categories. For example, consider a respondent who has an underlying propensity for either *good* or *very good*. These outcomes are already in the “middle” so 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 neighboring 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.

To expand on these new developments, firstly, assume a standard set-up for an OP model as described in Eqs. (1), (2) and (6), with constant μ_j across individuals. We will label these “first stage’ probabilities, $\Pr\left(y^{(0)} = j|\mathbf{x}, OP\right) = \Pr\left(j|OP\right)$, $j = 0, \ldots, J-1$. For individuals with a $j = 4$ (*excellent*) propensity, our approach simultaneously allows for tempering towards the inflated middle outcomes, and also for the respondent to simply “stay where they are” and choose *excellent*. Thus conditional on this first stage propensity, a binary $Probit$ model applies with potential outcomes: *excellent*, and *very good*. Let this tempering equation be determined by a latent equation of the form (with observed covariates $\mathbf{w}$ with unknown weights λ and a standard normal error term u)

$$t_4^* = \mathbf{w}'\lambda_4 + u_4, \tag{19}$$

with resultant probabilities of $\Pr\left(y = 4|t_4 = 4\right)$ and $\Pr\left(y = 3|t_4 = 4\right)$, where the conditioning indicates that these are tempering probabilities *from* (or conditional on), the *excellent* ($y = 4$) outcome in the first stage.

For the next cells on the choice scale, the first stage probabilities of the middle outcomes $y = 2$ and $y = 3$, will be left untempered: $\Pr\left(y = 2|t_4 = 1\right) = \Pr\left(y = 2\right)$ and likewise for $y = 3$. Next, consider the $j = 1$ choice, of *fair*. Once more, to allow for tempering from this choice, we can envisage a latent equation of the form

$$t_1^* = \mathbf{w}'\lambda_1 + u_1, \tag{20}$$

which will now drive this conditional choice to either *fair* (that is, there is no tempering); or to the inflated neighboring outcome of *good*. Again, recognizing the binary nature of these choices, Eq. (20) will translate itself into a further $Probit$ equation with resultant probabilities of $\Pr\left(y = 1|t_1 = 1\right)$, and $\Pr\left(y = 2|t_1 = 1\right)$.

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. In our example we do not consider tempering from this outcome. Thus overall probabilities for the tempered approach are given by

$$\Pr(y = j|TOP) = \begin{cases} \Pr\left(y = 0|\mathbf{x}, \mathbf{w}, \mathbf{f}, TOP\right) = \Pr\left(y = 0, TOP\right) \\ \Pr\left(y = 1|\mathbf{x}, \mathbf{w}, \mathbf{f}, TOP\right) = \Pr\left(y = 1|t_1 = 1, TOP\right)\Pr\left(t_1 = 1\right) \\ \Pr\left(y = 2|\mathbf{x}, \mathbf{w}, \mathbf{f}, TOP\right) = \Pr\left(y = 2|OP\right) + \Pr\left(y = 2|t_1 = 0, TOP\right)\Pr\left(t_1 = 0\right) \\ \Pr\left(y = 3|\mathbf{x}, \mathbf{w}, \mathbf{f}, TOP\right) = \Pr\left(y = 3|OP\right) + \Pr\left(y = 3|t_4 = 0, TOP\right)\Pr\left(t_4 = 0\right) \\ \Pr\left(y = 4|\mathbf{x}, \mathbf{w}, \mathbf{f}, TOP\right) = \Pr\left(y = 4|t_4 = 1, TOP\right)\Pr\left(t_4 = 1\right) \end{cases} \tag{21}$$

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

$$\ln L[TOP] = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} \Pr\left(y_i = j|\mathbf{x}_i, \mathbf{f}_i, \mathbf{w}_i, TOP\right). \tag{22}$$

This particular form of the TOP model is also summarized conceptually in the Appendix. Note that this specification is implicitly tempering from near neighbors only. A further generalization would be to allow for tempering across a range of neighboring alternatives. It would appear though, that computationally such a model would be data demanding whilst being conceptually sound.

Identification in tempered and inflated models is achieved through the underlying assumption of normality of the various stochastic elements, along with the inherent non-linearities in the resulting joint probabilities— see, for example, Eqs. (15)–(21). However, it is also buttressed by imposing variable exclusion restrictions in the various components of the models; this is discussed below in regard to variable selection. Such identification strategies are ubiquitous amongst the full suite of related “hurdle/inflation” models, that these model extensions are based upon. Moreover, even the assumption of normality appears to be rather benign one in such models, as evidenced in Harris and Zhao (2007). Note that whilst the assumed independence of the stochastic elements of the models does, as noted, help in terms of identification issues, it is not strictly required. However, for such relatively small sample sizes as we have, this was not deemed an appropriate additional complexity to already quite “data-hungry” approaches. Moreover, in general, we would recommend identification based on exclusion restrictions whenever possible.

2.4. 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 mistakenly attribute reporting heterogeneity to say DIF when in fact there was only inflation heterogeneity present in the data, or *vice versa*. However, clearly there exists the possibility that both forms of reporting heterogeneity operate jointly. We combine the above approaches to account simultaneously for both forms of reporting behavior. That is, we extend both the inflation models, $MIOP$ and TOP, described above to incorporate information from anchoring vignettes to adjust for reporting heterogeneity.

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

$$\begin{aligned} \mu_0^{MIOP} &= \mathbf{z}'\boldsymbol{\gamma}_0 \\ \mu_j^{MIOP} &= \mu_{j-1} + \exp\left(\mathbf{z}'\boldsymbol{\gamma}_j\right) \\ &\ \ \vdots \end{aligned} \tag{23}$$

and as before, with the more standard $HOPIT$ approach, 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 simply replaces the likelihood contribution arising from the $HOPIT$ part of the model, $\ln L[HOPIT]$, with that from the $MIOP$, $\ln L[MIOP]$, once the definition of the boundaries has been changed along the lines of Eq. (23) in the latter; such that

$$\ln L[V/MIOP] = \ln L[V] + \ln L[MIOP]. \tag{24}$$

Again, as these two components are linked through the common cut-point (or boundary) parameters and the overall model does not factorize into two independent models. Conceptually, the identifying assumption that an individual's subjective reporting scale that represents DIF is uncorrelated with their propensity to report accurately or not. That is, an individual's response scale represents a subjective assessment of the location of boundaries between different levels of health which is likely to be related to an individual's health resilience. This is distinct from a propensity to box-tick which is more strategic relating to preferences over time use, a need to “fit in” or not to “stand out” from the perceived norm, or simply a lack of understanding of the question asked. Accordingly, the key identifying assumption is that the vignette responses are not similarly affected by box-ticking behaviors.

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

$$\begin{aligned}\mu_0^{TOP} &= \mathbf{z}'\boldsymbol{\gamma}_0 \\ \mu_j^{TOP} &= \mu_{j-1} + \exp\left(\mathbf{z}'\boldsymbol{\gamma}_j\right) \\ &\vdots\end{aligned} \tag{25}$$

and the likelihood function is now

$$\ln L[V/TOP] = \ln L[V] + \ln L[TOP]. \tag{26}$$

We note here that it would also be possible to allow for unobserved heterogeneity in both structural and boundary components of the model (Greene et al. 2014, Greene and Hensher 2010 and Greene et al. 2015, for example); however, identification of such would be much aided if panel data were to hand. It would also be possible to allow all of the unobserved elements to be correlated, although this would lead to an extremely complicated system of equations, with likely very limited benefits.

Before moving on to the empirical results and experimental evidence, it is worth noting that simply observing high frequencies in particular categories is not *prima facie* evidence of reporting behavior and/or category inflation (although it is often used as such empirically). The key here, is that the same observationally equivalent outcome can arise as the result of two, or more, processes. It is then down to the model, and in particular the identifying variables, to try to disentangle these. The identification strategy can also be informed by any findings that unusually high frequencies in certain categories occur for some individuals in the presence of covariates (attributes) while smaller frequencies of that category occur for other people with similar attributes.

We note that it is *not* sufficient to simply allow for bunching of responses in particular categories via boundary shifts in $(C)HOPIT$ models, if the bunching is a result of two, or more, processes. It is similarly not appropriate to model any bunching by "inflation" methods, if in-fact the bunching was a result of simply boundary shifts. If these two quite separate processes are erroneously conflated by the researcher, whilst measures of model fit/appropriateness might appear favorable, completely wrong inference and policy-advice will result. For example, a solely DIF-correction approach could well reallocate a cohort of individuals into an inflated category as a "true" reflection of their underlying subjective assessment, whereas in reality this cohort is simply categorized by individuals who are inherent "box-tickers", whereby these outcomes are quite distinct from their true self-assessments.

Importantly, as we show below in the experimental evidence, if the assumed inflation is not present in the data, then the model works to collapse to a simpler version which does not embody this. Furthermore, any concerns that individuals responding, say, *good* or *very good* are responding accurately but are simply interpreting the categories differently, will be allayed by the use of the vignettes (as described above). A possible limitation of the suggested models here, is that vignette questions are not always available. However, Harris et al. (2020) have shown how it is possible to use (essentially merge in) vignettes collected in other datasets along with the primary data of interest.

3. Empirical application

3.1. Self-reported (assessed) health outcomes

The issue of the determinants of SAH has been widely studied in the (predominantly) health economics literature. Examples and differing foci abound, but importantly include one of our key driving sources Greene et al. (2015), as well as Bound (1991), Etilé and Milcent (2006), Jones et al. (2010), Lalji et al. (2018), Kesavayuth et al. (2020) and, as recent as Chen et al. (2023), as just a very small subset of such literature. Given the importance of SAH, as well as the strong potential for such hypothesized reporting behaviors here (as evidenced by the voluminous existing literature on such), this will form the basis of our empirical example, as detailed below.

3.2. Data description and variable selection

We conducted an online survey on a sample of Australian respondents aged 18 to 65 years recruited using a survey panel company. The sampling strategy targeted a representative sample according to gender-age-state of residence splits of the Australian population. Following an initial pilot, we collected data 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. The two waves are cross-sectional and do not form a panel. Ethics approval was obtained by the Monash University Human Research Ethics Committee, Monash University Australia.

Survey respondents were asked to rate their own health, together with the health of hypothetical individuals described in three vignettes. The available response categories are *poor*, *fair*, *good*, *very good*, and *excellent* health. The vignettes were developed to describe overall states of health at differing levels of severity, and are provided in the online Appendix. The hypothetical people described in the vignettes were assigned names that were gender-matched to the respondents as suggested by King et al. (2004). For further information on vignette construction, see Knott et al. (2017).

The inflation equation of the $MIOP$ and the tempering ones of the TOP could be considered to be largely driven by similar covariates; thus we use the same set in the inflation components of both models. These included binary variables constructed from: 1) a question at the end of the survey asking respondents how well they understood the questions of the survey (*Understood all questions* = 1 if respondents understood all question; = 0 if not); 2) a question asking whether others were present while the survey was being completed, since this should influence how truthfully people respond to survey questions (*No one else present* = 1 if others were not present; = 0 if others were present); and 3) whether the financial incentives received from participating in online surveys contributed significantly to respondents' household income, which could provide an indication for how seriously people took the survey (*Money received* = 1 if contributed significantly to household income; = 0 if not).

The above variables are crucial for identification and were chosen following similar research. For example, Brown et al. (2018) considered such variables to identify misreporting equations in a drug consumption framework. Since the variables mostly relate to the how the particular survey was administered, they should ostensibly affect reporting behavior, and not underlying health levels. Moreover, such factors are also in accordance with a long history of literature suggesting that they correlate strongly with misreporting/misclassification dating back to Mensch and Kandel (1988) and more recently Kraus and Augustin (2001) and Berg and Lien (2006). Finally, a very similar set of identifying variables were shown to perform well in a similar context in Greene et al. (2015).

The survey also contained a standard set of demographic and socioeconomic questions including age, gender, highest level of education, employment, marital and migrant status. Following the literature (for example, Contoyannis et al. 2004), these are included as covariates in the structural (health) component of all models estimated (*i.e.*, $\mathbf{x}$). They are also included in the boundary equations of models allowing for DIF ($\mathbf{z}$); and equations determining middle-inflation behaviors in middle-inflated and tempered models ($\mathbf{w}$, and $\mathbf{f}$). Summary statistics of the sample are provided in the online Appendix.

With respect to self-assessed health, superficiality there appears to be inflation within the middle categories, with nearly 70% of respondents reporting *very good* or *good* health. These sample proportions

Table 1
Regression results; OP, $MIOP$ and TOP.

	OP		$MIOP$		TOP	
Age/10	−0.443***	(0.081)	−0.44***	(0.112)	−0.482***	(0.088)
$(\text{Age}/10)^2$	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)
University educated	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)
Fixed boundary parameters:						
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/10			−0.443**	(0.206)	0.538	(1.198)
$(\text{Age}/10)^2$			0.041*	(0.024)	−0.021	(0.141)
Female			−0.064	(0.082)	0.816	(0.719)
University educated			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 equation		Tempering from *Excellent* health	
Constant			0.19	(0.386)	−0.144	(0.296)
Age/10			0.241	(0.186)	−0.006	(0.154)
$(\text{Age}/10)^2$			−0.023	(0.021)	−0.011	(0.018)
Female			−0.172**	(0.069)	−0.144**	(0.058)
University educated			−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:						
Model 1: $MIOP$; Model 2: TOP			−0.112			

***$p \le 0.01$.
**$p \le 0.05$.
*$p \le 0.1$.

are very similar to those reported in Greene et al. (2015), who used data from the Household, Income and Labour Dynamics in Australia ($HILDA$) Survey, a large representative household panel.

3.3. Results

We consider a range of models. Firstly, the reference OP model followed by the $MIOP$ of Greene et al. (2015) and the TOP model based on that of Greene et al. (2013), but with tempering to the two (hypothesized) inflated outcomes. To conserve space, we only report a selection of parameter results for these models. These are provided in Table 1. In the online Appendix we further report partial effects for reporting *very good* and *excellent* health.

We first discuss briefly direct parameter estimates for the mean functions $\boldsymbol{\beta}$. Note that the self-reported health measure is increasing in health, such that a positive (negative) coefficient means that the probability of reporting *excellent* (*poor*) health increases with increases in the variable. Categories of health between these two extremes can go either way. Increasing probability of *excellent* health is what we mean by "increasing health". There is reasonable consistency in the estimated coefficients across models with respect to sign and significance, but magnitudes vary. Health levels are decreasing with age (at a decreasing rate). Respondents who have a tertiary or high school qualification report better health than those who have not completed high school. Unemployed respondents report worse health compared to employed respondents (the reference group), in general, and respondents not in the labour force, in turn, report worse health than the unemployed (which could 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, Contoyannis et al. 2004).

Results for the binary health equation for the middle inflation ($MIOP$) model (Eq. (13)), indicate 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 (second panel of Table 1). Age and unemployment have a negative influence on the likelihood of reporting *very good* health, while tertiary education has a positive influence. In the *reporting* equation which distinguishes between accurate and inaccurate reporters, a range of both the (standard) demographic variables and two of the identifying variables appear significant (third panel of Table 1). Females are more likely to report inaccurately (inflate) compared to males, as are married respondents. Respondents who are not in the labour force are more likely to report accurately compared to employed and unemployed respondents. Of the identifying variables, the likelihood of reporting accurately increases for individuals who rely on the money they receive from responding to online surveys and decreases for individuals who did not have others present when completing the survey. This is similar to the effect found by Greene et al. (2015).

For the tempering equations (Eqs. (19) and (20)) of the TOP model, the first equation represents the propensity to report accurately at *fair* health, as opposed to tempering towards *good* health (second panel of Table 1); while the second represents the propensity to report accurately at *excellent* health, as opposed to tempering towards *very good* health (third panel of Table 1). Positive coefficients indicate a greater likelihood of accurate reporting, and negative coefficients indicate tempering toward the middle categories. The data are 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 *very good* health. The model, however, does appear to be identified with respondents stating that others were not present during the survey more likely to temper away from *excellent* health, while those that understood all questions were less likely to do so. The model fails to notably differentiate between the boundaries μ_3 and μ_4, effectively removing the distinction between *very good* and *excellent* health once account for tempering has been taken. Interpretation of the partial effects, presented and briefly discussed in the Appendix confirms the general relationships outlined above.

In the bottom panel of Table 1 we present *Vuong* non-nested test statistics where a large positive value favors the model denoted as Model 1, and a large negative value favors the alternative, denoted Model 2. There is evidence that this test is not appropriate for testing inflation in such model (Wilson, 2015), such that is not used as such here. However, comparing across the $MIOP$ and the TOP models, the test is inconclusive; whilst on the other hand, information criteria (not shown) favor the $MIOP$ over TOP across four measures considered: AIC (Akaike, 1987), corrected AIC, $CAIC$ (Bozdogan, 1987); BIC (Schwarz, 1978); and Hannan–Quinn $HQIC$ (Hannan and Quinn, 1979).

Results from the $MIOP$ and TOP models display evidence of reporting heterogeneity in the form of middle inflation. We next consider general reporting heterogeneity existing across the thresholds of the $HOPIT$ model. Structural parameter estimates, $\boldsymbol{\beta}$ are provided in Table 2. The full set of estimated boundary coefficients, $\boldsymbol{\gamma}_j$, are presented in the online Appendix. Age, educational attainment, marital status, employment and migrant status appear as predictors of reporting behavior in the boundary equations (for $j = 0, 1$ and 3). Broadly, the results indicate that older respondents tend to make use of the extreme categories with some individuals under reporting their level of health by making greater use of the categories *poor* and *fair* for a given underlying latent value of health and others over reporting by making greater use of the *excellent* health category compared to younger counterparts. Educated respondents also tend to under report a given level of underlying health in comparison to less educated counterparts; married respondents are 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.

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 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 are not as large in the $HOPIT$ compared to the OP model, however, the effects of education and migrant status are larger. The differences in estimates illustrate the role that DIF plays in the structural parameters.

The above results are strongly suggestive of reporting behavior in self-reports of health. We combine the models to control for both middle inflation and general forms of reporting behavior as outlined in Section 2.4, and refer to these as $V/MIOP$ and V/TOP where V indicates the HOPIT model (via vignettes). Results are presented in Table 2. Both the inflation model for the $V/MIOP$ and tempering equations for V/TOP contain significant effects for one or more of the identifying variables which operate in the same direction as for the single versions of these models described above. On the whole, the direction and significance of the model covariates follow a similar pattern to those for the corresponding $MIOP$ and TOP models of Table 1. The *Vuong* tests are unable to distinguish between the $V/MIOP$ and V/TOP models (information criteria again favor the $V/MIOP$ over V/TOP).

Partial effects for reporting *very good* and *good* are presented in the Appendix. For the $HOPIT$ model 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. For the $V/MIOP$ and V/TOP models again we 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.

We further compute and compare summary measures, 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 3 contains the ordered probit probabilities, estimated average probabilities for the $HOPIT$ model accounting for DIF, together with estimated average probabilities purged of variations in reporting behavior across models that consider both types of reporting heterogeneity ($V/MIOP$ and V/TOP). The bottom panel reports estimates from the ordered probit model relative to these (purged) probabilities (averaged over the sample—see Section 4 for a description).

Substantive variation is evident across the estimated probabilities. Compared to the ordered probit results, accounting for DIF only ($HOPIT$) leads to an increase in estimated probabilities in the excellent (by 8%) and fair categories (by 24%). The remaining categories see a decrease in average estimated probabilities, with some decreasing by more than 20%. 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. That is, without a tendency towards inflation these responses would have reported a different outcome. After purging of reporting behavior, 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 $V/MIOP$

Table 2
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/10	−0.108	(0.096)	−0.189*	(0.113)	−0.202*	(0.104)
$(\text{Age}/10)^2$	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)
University educated	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/10			−0.374	(0.273)	−0.034	(0.476)
$(\text{Age}/10)^2$			0.036	(0.032)	0.005	(0.053)
Female			−0.039	(0.102)	0.194	(0.181)
University educated			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 equation		Tempering from *Excellent* health	
Constant			0.341	(0.491)	0.398	(0.785)
Age/10			0.302	(0.241)	0.344	(0.413)
$(\text{Age}/10)^2$			−0.027	(0.028)	−0.035	(0.05)
Female			−0.223**	(0.09)	−0.388**	(0.159)
University educated			−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:						
Model 1: $V/MIOP$; Model 2: V/TOP			0.523			

***$p \le 0.01$.
**$p \le 0.05$.
*$p \le 0.1$.

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%. These differences between standard and purged predicted probabilities are substantial and emphasize the need to adjust for reporting behavior when aiming for meaningful inference on population health status.

4. Experimental evidence

To examine the finite sample performance of the set of models we carry out three Monte Carlo experiments. The first considers the ability of standard test statistics to select the correct specification. The second extends the first to include the anchoring vignettes in the model specification. The third considers an alternative to the normal distribution as the base platform for the data generating process.

We simulate data under the data-generating processes ($DGPs$) for the various ordered probit specifications; standard ordered probit (OP), middle inflated ($MIOP$), and tempered (TOP). These three are also augmented with the anchoring vignettes ($HOPIT$, $V/MIOP$, V/TOP) as described earlier. To generate data under each scenario, we begin with parsimonious specifications of the models considered in the empirical examples described in Section 3. We use the observed outcome data and covariates to estimate each model separately. The estimated

Table 3
SAH probabilities purged of inaccurate reporting.

	Predicted probabilities		Probabilities purged of inaccurate reporting	
SAH	OP	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 OP		
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

parameters are then used as the true values in the respective simulated $DGPs$. Each model was then simulated 1,000 times by re-drawing only the stochastic components (and consequent outcomes) for each DGP. On occasion, constant terms were perturbed to yield a more natural distribution of observed outcomes for the model under consideration. For example, in a replication in which the $MIOP$ is "correct", if the DGP for the model produced very little inflation, then additional inflation was achieved by perturbing the relevant constant term (in $\mathbf{w}$ in Eq. (11)).

A natural question relates to the role that the normality assumption might play in the statistical outcomes. For example, excess kurtosis might be manifested through greater prevalence of the extreme outcomes. To examine the likely consequences of non-normality on the performance of the test statistics, we modified the $V/MIOP$ model. The latent stochastic elements of the model are generated as non-normal (*Type 1 Extreme Value*) variates, but the estimator remains based on the normal log-likelihood. This provides a robustness check of the estimator against skewness and kurtosis in the conditional distribution of the latent outcomes. The $V/MIOP$ model was chosen for this exercise for parsimony and because it was the preferred specification in our empirical example.

Models that rely on vignettes, $HOPIT$, $V/MIOP$, V/TOP and Non-normal $V/MIOP$, can be regarded as having multiple outcomes (self-reported health plus the vignette outcomes). This will be reflected in larger likelihood contributions compared to the simpler non-nested counterparts; OP, $MIOP$, and TOP. Accordingly, for each replicate drawn under a given simulated DGP, we estimate all the available models within its class (OP, $MIOP$, TOP for ordered probit $DGPs$; and separately $HOPIT$, $V/MIOP$ and V/TOP for vignette DGP variants). Our performance indicator is the proportion of 1,000 replicates that information criterion correctly selects as the true model under each DGP. We consider a range of criteria based on information metrics: AIC; $CAIC$; BIC and; $HQIC$. We also consider *Vuong* tests for models that are not nested in the usual parameter-restriction sense, and report the power performance of the *Vuong* test in selecting the null model under the DGP against alternatives. As with the information criteria this is also undertaken for models within the same class as the DGP. The proportion of Monte Carlo replicates for which the information criterion selects the correct model under the assumed DGP are presented in the top panel of Table A6 in the Appendix; the bottom panel reports the proportion of replicates where the model under the DGP is preferred against the within-class alternative using the *Vuong* test.

Information criteria select the true model under the assumed DGP for the vast majority of replicates. The only exception is $CAIC$ for the $V/MIOP$ with non-normal error. This model is generated with *Type 1 Extreme Value* errors, but estimated assuming normally distributed errors. Under this criterion the corresponding $V/MIOP$ model is selected in the majority of replicates indicating robustness to skewness and kurtosis in the error. Unsurprisingly, the AIC performs marginally worse than other criteria when considering the most parsimonious within-class model under the DGP (OP, and $HOPIT$) due to the relatively small penalty placed on the additional number of parameters in the $MIOP$ and TOP model *versus* OP, and $V/MIOP$ and V/TOP *versus* the $HOPIT$ model. Overall, the *Vuong* test favors the $MIOP$ and TOP models when these form the DGP. Within the vignette class of models, when the DGP is $V/MIOP$, the *Vuong* test favors this model over the V/TOP model. When V/TOP forms the DGP it is favored over $V/MIOP$ in 55% of replicates and inconclusive across 45%. When *Type 1 Extreme Value* errors are included in the $V/MIOP$ under the DGP, but estimated assuming normality, the $V/MIOP$ is favored for the majority of replicates (64% *versus* 36%) when the alternative is V/TOP.

As for the empirical application in Section 3 we also compare the predicted probabilities from the various estimated models against probabilities generated under each DGP. For all estimated models, predicted probabilities are purged of reporting heterogeneity effects. For the inflation models, these are the probabilities arising solely from the OP component of the model. For models with DIF that rely on vignettes for identification, we evaluate the underlying OP model at fixed boundary parameters to provide predicted probabilities purged of reporting heterogeneity. This is achieved by evaluating the boundary equations at the sample means of the boundary covariates. For the models with both inflation and vignette components, these approaches were jointly applied. Probabilities are evaluated at an individual level and averaged over individuals and Monte Carlo replicates.

Table 4
Monte Carlo: Difference between predicted purged probabilities and true purged probabilities.

	Estimation Model				
	Models without vignettes		Models with vignettes		
DGP	$MIOP$	TOP	$HOPIT$	$V/MIOP$	V/TOP
OP	0.0450	*0.0108*	*0.0002*	*0.0048*	*0.0035*
$MIOP$	**0.0098**	*0.0178*	0.0940	*0.0148*	0.0462
TOP	*0.0342*	**0.0120**	*0.0108*	0.0588	*0.0001*
$HOPIT$	0.0770	0.0638	**0.0001**	*0.0020*	*0.0013*
$V/MIOP$	*0.0226*	0.0402	0.0410	**0.0001**	*0.0384*
V/TOP	*0.0300*	*0.0194*	0.0448	0.0464	**0.0002**
Non-Normal	*0.0054*	0.0432	*0.0360*	*0.0079*	*0.0152*

The purpose of comparing estimated purged probabilities with those produced under the DGP is twofold. Firstly, to check that the true models perform appropriately when the assumed DGP is correct, and secondly, to investigate the performance of models where the assumed DGP is incorrect. When incorrect, it is of interest to ascertain the likely finite sample performance of the more complex (middle-inflation and DIF) models, when in fact, no such behavior is present in the DGP. For example, if a middle inflation model ($MIOP$) essentially collapses to a simple OP when the latter is true, this implies that the former model is a 'safe option' in the sense that model over-specification does not bias predicted probabilities. On the other hand, it is also of interest to gauge the performance of models assuming no reporting heterogeneity when in fact it is present in some form or another.

Table 4 compares the purged probabilities predicted by the estimated models indicated by the column headers when the true DGP is the model labeled in the row (probabilities are purged by not including the heterogeneous inflation component.) We computed the five outcome probabilities for the true DGP and the estimated model, then obtained the mean absolute differences in the five cell probabilities for the categorical outcomes. Values as large as 0.040 seem indicative of misspecification. Moving across the columns in each row, the figures display the ability of each estimated model to match (or not) the probabilities generated by the same background DGP. Thus, as might be expected, the column labeled models tend to predict best when they are, in fact, the correct model for the DGP (results shown in bold).

As expected, estimation models that follow the true DGP model tend to perform best among competing models. When the DGP is an OP (topmost row), that is when middle-inflation and DIF are not present, all models except $MIOP$ perform well (results in italics). Interestingly, the more complex $HOPIT$ model appears to outperform others when considering differences in probabilities. When middle-inflation is present but not DIF (*i.e.*, the $MIOP$ model), the $HOPIT$ and V/TOP models perform poorly. When the DGP is TOP all models except $V/MIOP$ perform well. When the $HOPIT$ model represents the DGP, estimation of the OP and $MIOP$ models that do not account for DIF do poorly. For $DGPs$ that exhibit both DIF and middle-inflation bias ($V/MIOP$ and V/TOP) all models perform poorly in predicting true purged probabilities except for the associated estimation models themselves ($MIOP$ and $V/MIOP$ when the DGP is $V/MIOP$, and TOP and V/TOP when the DGP is V/TOP). Finally, under the non-Normal DGP, the $V/MIOP$, which is associated with the true DGP and the $MIOP$ and V/TOP all perform reasonably well.

Table A7 in the accompanying Appendix presents $RMSEs$ for the difference between predicted and 'true' purged probabilities. The general pattern observed in Table 4 is replicated. Overall, more complex versions of the models under the DGP tend to perform better than their less complex counterparts, indicating that over-specification of these models can be beneficial.

Finally, reflecting the fact that the analyst's primary concern will be estimation of relevant partial effects (PEs), we consider the estimation of the single PE of age and age^2 in our parsimonious specification. These are compared to the true value (PE_j) by taking the bias to be: $bias_j = E(\widehat{PE}_j) - PE_j$; and then percentage absolute bias: $\left(bias_j/|PE_j|\right) \times 100$, where j indicate the response category. In this way we can ascertain not only the percentage bias, but whether $\widehat{PE}_j$ over (under) estimates PE_j (with positive and negative values, respectively, due to the fact that $\sum_j PE_j = 0$). Note that these biases are not strictly comparable *across* DGPs though, as the true PE will differ. Table 5 reports the results. Unsurprisingly, biases are smallest for those models corresponding to the true DGP; for example, the simple OP model clearly performs best in the top panel (and so on). However, the more complex models have relatively little bias, such that they can be considered 'safe options' even if over-specified here. The possible exception to this is the $HOPIT$ model, with relatively high (percentage biases of 30, 43 and 54%). Under the $MIOP$ DGP the OP model appears severely biased, and to a lesser extent the $HOPIT$ as well. The TOP model fares well apart from those for $j = 0, 3$; whilst the vignette versions of both the $MIOP$ and TOP models, perform very well (with the singular exception of V/TOP_3). For the TOP DGP, most approaches, including the null model, appear to struggle to accurately estimate PE_1. Outside of this the under-specified OP appears to perform rather well, and the $MIOP$ is on a par with the null model. Although the $HOPIT$ model here appears to suffer somewhat from the misspecification, the $HOPIT$ versions of both the inflated models perform very well.

For experiments where DIF is introduced, we see that for all of the models where this is not accounted for (OP, $MIOP$ and TOP), they all appear to be quite severely biased, with the $MIOP$ appearing to fare best of the three. However, remarkably both the vignettes augmented versions of the $MIOP$ and TOP models perform well, and not much worse than the null model. A similar story arises when the DGP is $V/MIOP$ with the $HOPIT$ model performing reasonably well compared to those not allowing for DIF, whilst the V/TOP model performs exceptionally well, almost on a par with the null model. Finally, similarly for the V/TOP DGP, the non-DIF models tend to perform quite poorly, whilst the $HOPIT$ works well, and the $V/MIOP$ has performance just shy of the null model.

In summary the findings, as before, tend to suggest that without DIF both of the inflated model versions tend to quite accurately estimate the PEs, even when being over-specified. The same cannot be said of the under-specified OP model, when there is some form of middle-inflation. The DIF-augmented inflation models similarly perform extremely well here, even when quite clearly over-specified. When DIF is introduced into the DGP, not allowing for such proves very costly. Once more the DIF-augmented inflation models perform extremely well here, regardless of the precise DGP: thus reinforcing the previous findings that these can be considered very 'safe options' across-the-board.

Table 5
Monte Carlo: absolute percentage partial effect bias by outcome.

DGP	Model	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$
OP	OP	11.0	13.0	5.7	−8.8	−11.6
	$HOPIT$	−42.8	53.5	1.6	4.6	−30.1
	$MIOP$	16.8	11.4	2.6	−7.0	−13.0
	TOP	11.3	15.5	2.8	−7.7	−13.2
	$V/MIOP$	−14.1	34.5	9.9	−8.4	−20.5
	V/TOP	−28.4	40.8	17.3	−2.4	−29.2
$MIOP$	OP	215.9	134.2	−75.9	−601.8	73.1
	$HOPIT$	63.0	107.5	−54.6	−222.1	42.0
	$MIOP$	1.9	2.1	0.7	26.8	−5.5
	TOP	202.0	−11.2	−29.1	406.0	−5.6
	$V/MIOP$	−30.8	18.4	−3.7	48.9	−4.4
	V/TOP	89.1	−3.4	−12.8	222.5	−7.5
TOP	OP	2.1	76.3	−15.9	1.5	−8.6
	$HOPIT$	−71.4	258.7	−19.4	7.3	−40.9
	$MIOP$	11.6	80.1	−11.9	−11.9	7.2
	TOP	8.6	83.9	−10.1	−9.9	−1.4
	$V/MIOP$	1.8	106.0	−9.6	−14.2	3.9
	V/TOP	−18.0	87.6	11.1	−20.2	0.7
$HOPIT$	OP	128.7	3.5	−129.4	17.4	20.5
	$HOPIT$	−0.2	−0.1	0.4	0	−0.1
	$MIOP$	126.8	−13.4	−12.6	−0.2	17.0
	TOP	125.7	1.6	−129.3	23.2	8.8
	$V/MIOP$	3.0	0.3	0.6	−0.1	−1.7
	V/TOP	3.1	0.7	−1.5	−0.1	−1.2
$V/MIOP$	OP	79.3	−14.7	−45.5	−3.4	33.0
	$HOPIT$	−76.1	−0.9	39.7	16.1	−24.7
	$MIOP$	100.5	−14.9	−63.3	5.1	22.4
	TOP	70.2	−15.9	−72.3	23.0	9.0
	$V/MIOP$	1.8	1.1	−17.1	3.8	1.4
	V/TOP	−52.9	2.8	21.5	−1.3	2.0
V/TOP	OP	142.4	−9.9	−98.7	10.1	38.9
	$HOPIT$	−43.5	−0.9	−2.1	1.8	9.2
	$MIOP$	152	−16.9	−46.1	−1.6	33.4
	TOP	142.3	−8.4	−101.3	15	29.6
	$V/MIOP$	7.2	11.4	−38.5	4.1	−2.6
	V/TOP	0.0	1.3	−2.7	−1.6	2.4

5. Conclusion

This paper examines the analysis of self-reports of health, with a focus on reporting behavior brought about through differential item functioning, DIF, (the use of different thresholds that separate the reported levels of the self-assessment) and the adoption of a "box-ticking" strategy leading to inflation of the middle categories of the health variable. We consider and extend various models to address these forms of reporting in isolation, and propose models that allow for both types in combination.

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 remove DIF and increase cross-respondent comparability. For example, Kapteyn et al. (2007) 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. The approach is useful in identifying and correcting for general forms of reporting behavior by linking DIF to observed levels of respondent characteristics. However, the approach excludes more nuanced reporting brought about by "box-ticking" that leads to the artificial inflation of specific categories. In the case of subjective health, this type of reporting is suggested by the observation that the distribution of self-reported health is most often bunched around the middle categories, indicating a more favorable distribution of health than might be inferred from more objective measures. Identifying such behavior is strengthened 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). We extend these models and combine them to provide greater flexibility in modeling reporting behavior of either or both types exists in data.

In an empirical application we find convincing evidence for the existence of both types of reporting behavior in the widely-used generic measure of self-assessed health. After adjusting for these apparent nuances in reporting, we find that health levels on average are rather different than the raw self-reported data suggest. For example, in our preferred specifications, there is evidence that the observed categories of *very good* and *good* health are overestimated by 17% to 22% respectively. When adjusting for reporting behavior, we find a much higher proportion of respondents in the neighboring categories of *excellent* and *fair* health. Monte Carlo results, where we define objective health, clearly imply that the more flexible modeling approaches perform well, and collapse to the simpler versions when over-specified. These findings suggest that it is not simply the case of more flexible approaches fitting the data better. In practice, when no objective measures are available, it is only possible to probabilistically determine the likely magnitude and/or presence of such reporting behavior. Estimation and inference will, however, provide researchers an appropriate level of confidence in their findings. This is supported by the results of our Monte Carlo experiments where effectively zero levels of inaccurate reporting was found when the data was generated with none.

While our motivating example is to health, self-reported data is ubiquitous in social science research and the methods described are equally applicable to other outcomes where systematic reporting behavior is likely to be present. Applied research should be alert to the presence of reporting behavior when rely on self-reports, especially at an individual level, and the distorting effect that this can have on distribution of reported outcomes. This is likely to be most important in evaluating interventions, where ignoring such behavior could lead to erroneous inference of treatment effects. It will also be relevant in comparative research considering the distribution of subjective outcomes across social groups, regions or countries where systematic heterogeneity in reporting behavior, brought about through differences in preferences, expectations and norms is appreciable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank the Australian Research Council (Discovery Project, DP110101426, 2011-2013) for funding for this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econmod.2023.106277.

References

Akaike, H., 1987. Information measures and model selection. Int. Stat. Inst. 44, 277–291.

Angelini, V., Cavapozzi, D., Corazzini, L., Paccagnella, O., 2014. Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases. Oxf. Bullet. Econ. Stat. 76 (5), 643–666.

Au, N., Lorgelly, P.K., 2014. Anchoring vignettes for health comparisons: an analysis of response consistency. Qual. Life Res. 23 (6), 1721–1731.

Bago d'Uva, T., O'Donnell, O., Van Doorslaer, E., 2008. Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. Int. J. Epidemiol. 37, 1375–1383.

Berg, N., Lien, D., 2006. Same-sex sexual behaviour: US frequency estimates from survey data with Simultaneous Misreporting and Non-Response. Appl. Econ. (ISSN: 0003-6846) 38 (7), 757–769.

Black, N., Johnston, D.W., Suziedelyte, A., 2017. Justification bias in self-reported disability: New evidence from panel data. J. Health Econom. 54, 124–134.

Bound, J., 1991. Self-reported versus objective measures of health in retirement models. J. Hum. Resour. 26 (1), 106–138.

Bozdogan, H., 1987. Model selection and Akaike's information criteria (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370.

Brown, S., Harris, M.N., Spencer, C., 2020. Modelling category inflation with multiple inflation processes: Estimation, specification and testing. Oxf. Bullet. Econ. Stat. 82, 1342–1361.

Brown, S., Harris, M., Srivastava, P., Taylor, K., 2021. Mental health, reporting bias and economic transitions. Oxf. Econ. Pap. (ISSN: 0030-7653) 74 (2), 541–564.

Brown, S., Harris, M.N., Srivastava, P., Zhang, X., 2018. Modelling illegal drug participation. J. R. Stat. Soc. Ser. A (Statistics in Society) 181 (1), 133–154.

Carro, J., Traferri, A., 2014. State-dependence and heterogeneity in health using a bias-corrected fixed-effects estimator. J. Appl. Econometrics 29 (2), 181–207.

Chen, Y., Wang, H., Cheng, Z., Smyth, R., 2023. Education and migrant health in China. Econ. Model. 121, 106223.

Contoyannis, P., Jones, A., Rice, N., 2004. The dynamics of health in the British Household Panel Survey. J. Appl. Econometrics 19 (4), 473–503.

Crossley, T.F., Kennedy, S., 2002. The reliability of self-assessed health status. J. Health Econ. 21 (4), 643–658.

Currie, J., 2000. Child health in developed countries. In: Culyer, A., Newhouse, J. (Eds.), Handbook of Health Economics. UK, Elsevier.

Currie, J., Madrian, B., 1999. Health, health insurance and the labour market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labour Economics. Elsevier Science Publishers BV, Amsterdam, pp. 3309–3416.

Davillas, A., Jones, A.J., Benzeval, M., 2019. The income-health gradient: evidence from self-reported health and biomarkers in Understanding Society. In: Tsionas, M. (Ed.), Panel Data Econometrics: Empirical Applications. Elsevier, pp. 709–741.

Dowd, J.B., Zajacova, A., 2010. Does self-rated health mean the same thing across socioeconomic groups? Evidence from biomarker data. Ann. Epidemiol. 20, 743–749.

Etilé, F., Milcent, C., 2006. Income-related reporting heterogeneity in self-assessed health: evidence from France. Health Econom. 15, 965–981.

Greene, W., Gillman, M., Harris, M., Spencer, C., 2013. The Tempered Ordered Probit (TOP) model with an application to monetary policy. Department of Economics, Loughborough University.

Greene, W., Harris, M., Hollingsworth, B., 2015. Inflated responses in measures of self-assessed health. Am. J. Health Econ. 1, 461–493.

Greene, W., Harris, M.N., Hollingsworth, B., Weterings, T.A., 2014. Heterogeneity in ordered choice models: a review with applications to self-assessed health. J. Econ. Surv. 28 (1), 109–133.

Greene, W.H., Harris, M.N., Knott, R.J., Rice, N., 2020. Specification and testing of hierarchical ordered response models with anchoring vignettes. J. R. Stat. Soc. Ser. A 184, 31–64.

Greene, W., Hensher, D., 2010. Modeling Ordered Choices. Cambridge University Press.

Groot, W., 2000. Adaptation and scale of reference bias in self-assessments of quality of life. J. Health Econ. 19 (3), 403–420.

Hannan, E., Quinn, B., 1979. The determination of the order of an autoregression. J. R. Stat. Soc. Ser. B Stat. Methodol. 41, 190–195.

Harris, M.N., Knott, R.J., Lorgelly, P.K., Rice, N., 2020. Using externally collected vignettes to account for reporting heterogeneity in survey self-assessment. Econom. Lett. 194, 109325.

Harris, M., Zhao, X., 2007. A zero-inflated ordered probit model, with an application to modelling tobacco consumption. J. Econometrics 141 (2), 1073–1099.

Hausman, J., 2001. Mismeasured variables in econometric analysis: problems from the right and problems from the left. J. Econ. Perspect. 15 (4), 57–67.

Hausman, J.A., Abrevaya, J., Scott-Morton, F.M., 1998. Misclassification of the dependent variable in a discrete-response setting. J. Econometrics 87 (2), 239–269.

Jones, A., Rice, N., Roberts, J., 2010. Sick of work or too sick to work? Evidence on self-reported health shocks and early retirement from the BHPS. Econ. Model. 27, 866–880.

Kapteyn, A., Smith, J., Van Soest, A., 2007. Vignettes and self-reports of work disability in the United States and the Netherlands. Am. Econ. Rev. 97 (1), 461–473.

Kerkhofs, M., Lindeboom, M., 1995. Subjective health measures and state dependent reporting errors. Health Econ. 4 (3), 221–235.

Kesavayuth, D., Poyago-Theotoky, J., Tran, D.B., Zikos, V., 2020. Locus of control, health and healthcare utilization. Econ. Model. 86, 227–238.

King, G., Murray, C., Salomon, J., Tandon, A., 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. Am. Political Sci. Rev. 98 (1), 191–207.

Knott, R.J., Lorgelly, P.K., Black, N., Hollingsworth, B., 2017. Differential item functioning in quality of life measurement: An analysis using anchoring vignettes. Soc. Sci. Med. 190, 247–255.

Kraus, L., Augustin, R., 2001. Measuring alcohol consumption and alcohol-related problems: comparison of responses from self-administered questionnaires and telephone interviews. Addiction 96 (3), 459–471.

Kristensen, N., Johansson, E., 2008. New evidence on cross-country differences in job satisfaction using anchoring vignettes. Lab. Econ. 15 (1), 96–117.

Lalji, C., Pakrashi, D., Smyth, R., 2018. Can eating five fruit and veg a day really keep the doctor away? Econ. Model. 70, 320–330.

Lindeboom, M., 2006. Health and work of older workers. In: Jones, A. (Ed.), Elgar Companion to Health Economics. Edward Elgar, Cheltenham.

Lindeboom, M., van Doorslaer, E., 2004. Cut-point shift and index shift in self-reported health. J. Health Econ. 23 (6), 1083–1099.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. In: Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, ISBN: 0-471-00626-2, p. xxii+419, 1789474(2002b:62025).

Mensch, B., Kandel, D., 1988. Underreporting of substance use in a national longitudinal youth cohort. Publ. Opin. Quart. (ISSN: 0033-362X) 52 (1), 100–124.

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

Rice, N., Robone, S., Smith, P.C., 2012. Vignettes and health systems responsiveness in cross-country comparative analyses. J. R. Stat. Soc. Ser. A (Statistics in Society) 175 (2), 337–369.

Sadana, R., Mathers, C., Lopez, A., Murray, C., Iburg, K., 2000. Comparative analysis of more than 50 houshold surveys on health status. Technical report GPE Discussion Paper No 15, EIP/GPE/EBD, World Health Organisation, Geneva.

Schwarz, G., 1978. Estimating the dimensions of a model. Ann. Statist. 6 (2), 461–464.

Sirchenko, A., 2020. A model for ordinal responses with heterogeneous status quo outcomes. Stud. Nonlinear Dyn. Econom. 24 (1), 20180059.

Terza, J., 1985. Ordered probit: A generalization. Commun. Stat. A. Theory Methods 14 (1), 1–11.

UKHLS, 2022. Understanding Society, Waves 1-8 2009–2017 and Harmonised BHPS: Waves 1-18 1991–2009 [data collection]. 17th Edition. Technical report, University of Essex, Institute for Social and Economic Research. UK Data Service. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-13.

Wilson, P., 2015. The misuse of the vuong test for non-nested models to test for zero-inflation. Econom. Lett. 127, 51–53.