The effects of self-assessed health: Dealing with and understanding misclassification bias*

Linkun Chen† University of Melbourne

Philip M. Clarke‡ University of Oxford

Dennis J. Petrie§ Monash University

Kevin E. Staub¶ University of Melbourne

December 19, 2019

Abstract

Self-assessed health (SAH) is often used in health econometric models as the key explanatory variable or as a control variable. However, there is evidence questioning its test-retest reliability, with up to 30% of individuals changing their response. Building on recent advances in the econometrics of misclassification, we develop ways of obtaining consistent estimates despite misclassification in reported SAH using data from a large representative household survey where SAH was elicited twice, to gain new insights into the nature of SAH misclassification and its potential for biasing health econometric estimates. The results from applying our approach to nonlinear models of long-term mortality and chronic morbidities reveal that there is substantial heterogeneity in misclassification patterns. We find that adjusting for misclassification is important for estimating the impact of SAH. For other variables of interest, we find significant but generally small changes to their estimates when misclassification is ignored.

Keywords: Misreporting; measurement error; multinomial regressor; discrete and limited dependent variables; subjective health; mortality; chronic conditions.
JEL classification: C35; I12.

Acknowledgements: We thank Denzil Fiebig, Bill Griffiths, Mark Harris, Joe Hirschberg, Maarten Lindeboom, Jenny Lye, Frank Windmeijer, Rainer Winkelmann, Eugenio Zuccheli, the participants of the European Workshop on Econometrics and Health Economics (Groningen), the Asian Meeting of the Econometric Society (Hong Kong), the China Meeting of the Econometric Society (Wuhan), the International Association for Applied Econometrics conference (Sapporo), the Australian Health Economics Society conference (Freemantle, WA), the Health and Wellbeing Workshop (Werribee, VIC) and seminar participants at Erasmus University for helpful comments. Petrie acknowledges support from the Australian Research Council through grant DE150100309. Staub acknowledges support from the Australian Research Council through grant DE170100644. Alex Ballantyne and Edwin Chan provided excellent research assistance. The names of the authors are listed in alphabetical order.

†Melbourne School of Population and Global Health, 207 Bouverie Street, The University of Melbourne 3010 VIC, Australia

‡Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK. E-mail: philip.clarke@ndph.ox.ac.uk

§Centre for Health Economics, Monash Business School, 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. E-mail: dennis.petrie@monash.edu

¶Department of Economics, 111 Barry Street, The University of Melbourne, 3010 VIC, Australia. E-mail: kevin.staub@unimelb.edu.au

1 Introduction

Self-assessed health (SAH) is a ubiquitous measure in the health economics literature and, more broadly, in social science research (Au & Johnston, 2014). It is often asked as a simple question, "in general, how would you rate your health?", where respondents select from categories such as excellent, very good, good, fair or poor. SAH is used variously in econometric models as the outcome variable, as the key explanatory variable or as a control variable to prevent health from confounding the effect of interest. However, there is a new and growing literature calling into question the reliability of reported SAH, as up to 30% of individuals change their response when re-asked about their SAH (Crossley & Kennedy, 2002; Clarke & Ryan, 2006; Black *et al.*, 2017). This paper takes advantage of major breakthroughs in the recent econometric literature on misclassification and the fact that SAH was elicited twice in some waves of a prominent household survey—the Household, Income and Labour Dynamics in Australia (HILDA) survey—to gain new insights into the existence and nature of misclassification in reported SAH and on its potential for biasing estimates of the effects of SAH and other explanatory variables in health econometrics models.

In particular, our analysis uses data from the 2001 wave of HILDA, which records the same individual's SAH responses in two different but similar questionnaires in the same wave (face-to-face or over-the-phone interview, and on a self-completion questionnaire), and combines this information with longitudinal data on mortality and the development of chronic health conditions 15 years later. We develop a new likelihood-based nonlinear estimator which uses this data to jointly identify the misclassification in both reported SAH measures as well as the effects of SAH on mortality and morbidity. This framework builds on the insights of Hu (2008, 2017), who showed how two measurements of a misclassified categorical variable, coupled with outcomes that depend on that variable, nonparametrically identify all misclassification probabilities as well as the conditional expectation of the outcomes as a function of the unobserved categorical variable.

This finding makes it possible to go beyond the current literature, which only identifies differences in multiple measures of reported SAH. Our new framework explains such differences in terms of the underlying misclassification probabilities, and therefore answers behavioural questions about the extent, patterns and heterogeneity of individuals' responses. It also makes it possible to assess questions pertaining to survey methodology, such as the type and incidence of response errors associated with each of the two survey instruments—face-to-face interview and self-completion questionnaire—where the previous approaches would only be able to document different response patterns, but would be unable to make statements about the accuracy of these instruments.

A second key contribution of our paper is that it provides a way to estimate the effects of SAH in nonlinear models, while accounting for misclassification in reported SAH, which is a non-classical form of measurement error. By comparing such estimates to naïve estimates which ignore misclassification, this makes it possible to assess how biased conventional estimates of the effects of reported SAH might be by misclassification. Similarly, it makes it possible to assess how biases stemming from misclassification of reported SAH affect the estimates of other regressors of interest.

Finally, a third contribution relates to the estimation procedure we propose. As in the closely related setting of Hu (2008), the identification leads to a finite-mixture type of model. Nonparametric estimators suffer from the curse of dimensionality which can make it difficult to implement them in more complex applications. Our sample size, for instance, is an order of magnitude larger than that in Hu (2008). Therefore, our approach is a flexible parametric specification estimated via a standard expectation-maximisation (EM) algorithm, which offers fast and reliable computation. This makes it straightforward to apply these methods to a large dataset for a model with many covariates and a potentially misclassified SAH variable with many categories, as in our application. Further, our approach lends itself easily to specifications with interaction effects where the impact of unobserved SAH differs depending on other individual characteristics. Such specifications, while common in applied work, have received little attention in the misclassification literature. We also propose to improve the finite sample performance of our estimator by estimating several outcomes jointly and by implementing a penalised likelihood version of the estimator. The first potential improvement takes advantage of the fact that the misclassification system is the same across different outcome models, while penalisation is used to improve the convergence properties of the estimator and avoid extreme estimates in the face of sparse data.

Dealing with measurement errors in self-assessed variables related to health is an active area of research within health econometrics. For instance, in two recent contributions, Greene *et al.* (2018) and Brown *et al.* (2018) adjust for untruthful reporting in discrete dependent variable models. In contrast, our focus lies in the case of discrete health taking the role of a regressor. A related strand of the literature considers justification biases in reported SAH (Currie & Madrian, 1999; Lindeboom & Kerkhofs, 2009), where, e.g., individuals who do not work under-report their health status. Since we consider future outcomes, misreporting due to justification does not bias our estimates. Rather, our approach is agnostic regarding the reasons for variation in reporting behaviour and encompasses such measurement errors. Our paper also contributes to the literature which contrasts "objective" and self-assessed health measures (Mossey & Shapiro, 1982; Butler *et al.*, 1987; Baker *et al.*, 2004; Doiron *et al.*, 2015). We consider substantially longer-term associations between SAH and mortality (and morbidity) than in these studies (15 years vs 3-6 years), and we adjust the association by accounting for misclassification in reported SAH. Most closely related are Crossley & Kennedy (2002), Clarke & Ryan (2006) and Black *et al.* (2017), which also consider the change in an individuals' response when SAH is asked twice; however, none of these papers estimates the impact of misclassification when reported SAH is used as a regressor, nor do they study the underlying misclassification probabilities.

Section 2 introduces our econometric model for predicting long-term mortality and morbidity in the face of potentially misclassified SAH, and in Section 3 we examine its performance in Monte Carlo simulations. The evidence from the simulations suggests that the estimator, especially the penalised version of it, is able to estimate the parameters of interest satisfactorily even under challenging circumstances such as when SAH has many categories, there are interaction effects in SAH, or the estimator's parametric assumptions are violated.

Section 4 presents the results of our application using the HILDA survey. We find strong evidence

for the presence of misclassification, and for heterogeneity in misreporting behaviour across different population subgroups, such as male vs females and low vs high income earners. We also document that there is less measurement error in the SAH question elicited by face-to-face interviews than in the one from the self-completion questionnaire. The results indicate that misclassification leads to statistically significant biases in the parameters of the mortality and morbidity models. While for the coefficients of SAH in the mortality models these range mostly from 10 to 20 percent, for the morbidity models the biases are as high as 100 percent. For the coefficients of other covariates, the biases, while statistically significant, are more moderate and around 10 percent.

We conclude the paper in Section 5. Our findings suggest that when specifying models where SAH is the regressor of interest, it is important to adjust for misclassification. In case this is not possible, SAH measures from face-to-face interviews should be strongly preferred over self-completed SAH measures. On the other hand, our findings also indicate that when specifying models where SAH is used as a key control variable, there is likely to be little contamination of the variables of interest from the msiclassification in SAH.

2 Econometric models with misclassification in SAH

In this section, we discuss the specification (Section 2.1) and estimation (Section 2.2) of econometric models with misclassified SAH. As mentioned, conceptually, we build on the approach of Hu (2008). However, our estimator is parametric. Other related approaches include Gosling & Saloniki (2014) and Kane *et al.* (1999); but these papers do not include regressors and are limited to linear models, respectively. Battistin *et al.* (2014) develop a Bayesian approach. With the exception of Hu (2008), in all these papers the misclassified regressor is binary; see Schennach (2016) and Hu (2017) for overviews of the econometric measurement error literature.

To fix ideas, introduce notation, and give an intuition about the identification of the model, we start Section 2.1 by discussing a minimal example of a logit model with a binary potentially misclassified regressor and constant misclassification probabilities. We then show how this simple model can be extended to accommodate covariates, interaction effects with unobserved health, multinomial health with more than two categories, and heterogeneous misclassification probabilities. As we discuss in Appendix A.2, the framework we propose can be easily extended to other common nonlinear models. In Section 2.2 we present a standard expectation-maximisation (EM) algorithm to estimate these models, and discuss two ways of potentially improving the estimation in finite samples: penalisation and system estimation.

2.1 Specification

2.1.1 A simple logit model

Consider a simple logit model for mortality, an outcome we will use in our application in Section 4. The outcome y_i equals 1 if individual i is dead 15 years after the initial survey, and 0 otherwise. We are interested in how SAH, h_i^*, at the time of the initial survey, is related to mortality y_i. For now, let SAH be a binary variable: $h_i^* = 1$ indicates that individual i is in good health; and $h_i^* = 0$ that i is in bad health. The key feature of the models we consider is that SAH, h_i^*, is unknown; what is known instead is an individual's reported SAH, and this might be misclassified. Each individual reports his or her SAH twice, thus providing two potentially misclassified measures. SAH is related to mortality as follows:

$$y_i = \mathbb{1}(\alpha\, h_i^* + \beta_0 + \varepsilon_i > 0), \quad i = 1, \ldots, N, \tag{1}$$

where $\mathbb{1}(\cdot)$ represents the indicator function, α and β_0 unknown scalars and ε_i an IID logistically-distributed idiosyncratic error. Thus, the probability of mortality as a function of SAH is

$$P(y_i = 1|h_i^*) = \frac{\exp(\alpha h_i^* + \beta_0)}{1 + \exp(\alpha h_i^* + \beta_0)} \equiv \Lambda(\alpha h_i^* + \beta_0). \tag{2}$$

If h_i^* were observed, (2) would serve as the basis for a standard logit estimation; but since h_i^* is unobserved, this is infeasible. Instead, we consider conditions under which we can estimate α and β_0 by using two potentially misclassified SAH measures denoted as h_{1i} and h_{2i}, corresponding to the first and second response of the individuals, respectively. We define the following misclassification probabilities—i.e., conditional probabilities of misreporting SAH—as

$$\delta_{0|1}^m = P(h_{mi} = 0|h_i^* = 1) \quad \text{and} \quad \delta_{1|0}^m = P(h_{mi} = 1|h_i^* = 0), \quad \text{for } m = 1, 2. \tag{3}$$

We denote the distribution of the SAH as

$$P(h_i^* = 1) \equiv \pi. \tag{4}$$

The marginal distributions of the reported SAH measures can then be expressed as functions of the parameters defined in equations (3) and (4):

$$P(h_{mi} = 1) = \pi_i(1 - \delta_{0|1}^m) + (1 - \pi_i)\delta_{1|0}^m. \tag{5}$$

2.1.2 Identification

Here we briefly summarize the intuition behind the identification of the model which has been established in the important work by Hu (2008). Not observing h_i^*, we identify and estimate the parameters of the outcome equation (2) using the structure provided by equations (2) and (3), the data (y_i, h_{1i}, h_{2i}) and a couple of assumptions. Specifically, we use the structure to derive the joint distribution of y_i, h_{1i}, h_{2i} and make assumptions about the relationship between the misclassified health measures and the outcome:

Conditional Independence Assumption (CIA1): Conditional on SAH status h_i^*, the reported measures, h_{1i} and h_{2i}, are independent of each other and of the outcome, y_i.

The joint distribution of the outcome and the two misreported health measures consists of the eight probabilities $P(y_i{=}r_0,\, h_{1i}{=}r_1,\, h_{2i}{=}r_2 \,|\, \boldsymbol{x}_i) \equiv F(r_0, r_1, r_2)$, where $r_0 \in \{0,1\}$, $r_1 \in \{0,1\}$, $r_2 \in \{0,1\}$. Then,

$$\begin{aligned} F(r_0,r_1,r_2) &= \pi\, F(r_0,r_1,r_2|h_i^*{=}1) \quad + \quad (1-\pi)\, F(r_0,r_1,r_2|h_i^*{=}0) \\ &= \pi\, F(r_0|h_i^*{=}1)\, F(r_1|h_i^*{=}1)\, F(r_2|h_i^*{=}1) \\ &\quad + (1-\pi)\, F(r_0|h_i^*{=}0)\, F(r_1|h_i^*{=}0)\, F(r_2|h_i^*{=}0), \end{aligned} \tag{6}$$

where

$$\begin{aligned} F(r_m|h_i^*=1) &= (\delta_{0|1}^m)^{1-r_m}\, (1-\delta_{0|1}^m)^{r_m}, \\ F(r_m|h_i^*{=}0) &= (\delta_{1|0}^m)^{r_m} \quad (1-\delta_{1|0}^m)^{1-r_m}, \\ F(r_0|h_i^*{=}1) &= \Lambda(\alpha+\beta_0)^{r_0}\, (1-\Lambda(\alpha+\beta_0))^{1-r_0}\,, \\ F(r_0|h_i^*{=}0) &= \Lambda(\beta_0)^{r_0}\, (1-\Lambda(\beta_0))^{1-r_0}\,. \end{aligned}$$

The second equality in (6) follows from the conditional independence assumption (CIA1). To see an example of one of the expressions in (6), consider $F(1,1,1)$:

$$\begin{aligned} F(1,1,1) &= P(y_i=1, h_{1i}=1, h_{2i}=1 \,|\, \boldsymbol{x}_i) = \pi F(1,1,1|h_i^*=1) + (1-\pi)F(1,1,1|h_i^*=0) \\ &= \pi\, \Lambda(\alpha+\beta_0)\, (1-\delta_{0|1}^1)\, (1-\delta_{0|1}^2) \quad + \quad (1-\pi)\, \Lambda(\beta_0)\, \delta_{1|0}^1\, \delta_{1|0}^2. \end{aligned}$$

The model fulfils a necessary condition for identification since the data provides seven linearly independent quantities $F(r_0, r_1, r_2)$, which we can map to the seven parameters of the model: α, β_0, π, $\delta_{0|1}^1$, $\delta_{1|0}^1$, $\delta_{0|1}^2$, $\delta_{1|0}^2$.[1] However there are actually two solutions to this problem. To obtain a unique solution and identify the parameters, we can require for each measure that the probability of reporting truthfully be greater than the probability of misreporting:

No Mirror Assumption (NMA1): $\qquad \delta_{1|1}^m > \delta_{0|1}^m \qquad$ and $\qquad \delta_{0|0}^m > \delta_{1|0}^m,$

which amounts to assuming that $\delta_{0|1}^m, \delta_{1|0}^m < 0.5$. With this assumption, we rule out the "mirror solution" in which probabilities of misreporting and correctly reporting are switched and the impact of each health level on the outcome y_i is also switched (i.e., $\tilde{\alpha} = -\alpha$ and $\tilde{\beta}_0 = \beta_0 + \alpha$).[2]

[1] Note that for Eq. (6) to provide seven linearly independent quantities we need the regularity condition that the outcome be informative of the SAH status; that is, we need to assume that $\alpha \neq 0$. Thus, while one can identify if α is equal to 0 or not, it is not possible in the case of $\alpha = 0$ to further estimate the misclassification probabilities because the outcome does not inform us about which SAH group each person falls into.

[2] See Hu (2008) for a discussion of an alternative identifying mirror assumption: that the estimated direction of the impact of each health level on the outcome is known; that is, in this case, that $\hat{\alpha}$ is negative.

Thus, under CIA1 and NMA1, the system is just-identified, paving the way for estimation. If only one health measure, say h_{1i}, was available, the joint distribution (y_i, h_{1i}) would consist of three independent probabilities. However, there would be five parameters to estimate—$\pi, \delta^1_{0|1}, \delta^1_{1|0}, \alpha, \beta_0$— and the system would be under-identified. Similarly, with two health measures but without the outcome y_i it would also be impossible to identify the misclassification probabilities. There would only be the three independent probabilities of the joint distribution of (h_{1i}, h_{2i}) to estimate the four parameters $\delta^1_{0|1}, \delta^1_{1|0}, \delta^2_{0|1}, \delta^2_{1|0}$ (or five, including π).

2.1.3 Full model with covariates and heterogeneous misclassification probabilities

The model discussed so far is quite minimal. Not only does it not include any other regressors apart from the health indicator, but the misclassification probabilities are the same across all individuals. We now consider a model which addresses these shortcomings by adding covariates in the outcome equation and making the misclassification probabilities depend on these covariates as well. First, consider the case of modifying the minimal model by only adding covariates to the outcome equation. The constant β_0 can be replaced by a linear index $\boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is a $K \times 1$ vector of covariates with conforming coefficient vector $\boldsymbol{\beta}$. The joint distribution in (6) and the corresponding expressions are then simply to be taken conditional on $\boldsymbol{x}_i$. The number of parameters to be estimated is now $6 + K$ (the five probabilities $\pi, \delta^1_{0|1}, \delta^2_{0|1}, \delta^1_{1|0}, \delta^2_{1|0}$, the key parameter of interest α, as well as the K elements in $\boldsymbol{\beta}$). In this case, the system is over-identified since there will be at least $(1 + 2^{K-1}) \times 7$ different values of $F(r_0, r_1, r_2|\boldsymbol{x}_i)$, the number $(1 + 2^{K-1}) \times 7$ corresponding to the minimal case of a constant and $K - 1$ linearly independent binary regressors.

With covariates, it is also possible to revisit the conditional independence assumption. The current conditional independence assumption is strong, but it might be reasonable in some contexts.[3] Violation of the current independence assumption can occur, for example, if men and women have different misreporting probabilities. In such a case, the assumption does not hold because the two misreported measures will be dependent through the impact of gender. Thus, a way to weaken this assumption is to explicitly make the misclassification probabilities dependent on x_i and only require independence to hold conditional on some x_i.

> Conditional Independence Assumption (CIA2): Conditional on SAH status h_i^* and on observed variables $\boldsymbol{x}_i$, the reported measures, h_{1i} and h_{2i}, are independent of each other and of the outcome, y_i.

Is the model still identified under CIA2 and the no-mirror-solution assumption? Consider first the case of discrete regressors $\boldsymbol{x}_i$. In this case, we know from above that we could identify the parameters for each subsample defined by one particular set of values of $\boldsymbol{x}_i$. Thus, the identification of the model under CIA2 is equivalent to the identification of each subsample under the constraint that α and $\boldsymbol{\beta}$ are the same across subsamples.

[3] For instance, in the field of health economics, Gosling & Saloniki (2014) use it in an application to misreported binary disability status.

When some of the other regressors are continuous, we cannot directly resort to this simple approach of identification in each subsample. However, the model remains identified: Intuitively, in an infinitely large sample, we could discretize the continuous regressors ever more finely and then apply identification in each subsample. A formal proof of the model's nonparametric identification has been given by Hu (2008), who also proposes a nonparametric estimator. Having such a flexible estimation framework for the misclassification has the tradeoff of increasing small sample bias and becoming computationally intensive when there are numerous regressors over which the level of misclassifications may vary. Instead, we proceed by using a more standard parametric, and hence restrictive, approach for the misclassification, but which has the advantage of reducing the impact of small sample bias and of easily being able to incorporate many regressors in the misclassification equations. With our approach it is also straightforward to increase the flexibility of the parametric form in the misclassification equations to test the sensitivity of the results to a particular functional form. More generally, because of the underlying nonparametric identification of the misclassification, our approach can also serve as the basis of a nonparametric estimation via a series estimation approach (such as by including polynomials or splines of the linear indices, cf. Newey, 1994). As a basic specification, we assume that the misclassification probabilities are known functions of the regressors,

$$\delta^m_{0|1} = \Lambda\big(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}^m_{0|1})\big), \qquad \text{and} \qquad \delta^m_{1|0} = \Lambda\big(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}^m_{1|0})\big). \tag{7}$$

The logistic function coupled with the negative exponential function in (7) enforces the (0,0.5)-bounds under the no-mirror-solution assumption on the misclassification probabilities. Similarly, we assume that SAH is also a known function of the regressors,

$$\pi_i \equiv P(h^*_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\eta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\eta})}. \tag{8}$$

That is, we assume SAH conditional on covariates to be of the logit form.

2.1.4 Categorical SAH

For our application, we are interested in extending the framework to deal flexibly with SAH with more than two categories.

For concreteness, let us assume SAH has five outcomes, $h^*_i \in \{0,1,2,3,4\}$, as this is the case in our application of Section 4. As before, two potentially misclassified measures, h_{1i}, h_{2i} are observed. The model is now

$$y_i = \mathbb{1}(\boldsymbol{d}^{*'}_i\boldsymbol{\alpha} + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0), \quad i = 1,\ldots,N \tag{9}$$

with $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3,\alpha_4)'$ and $\boldsymbol{d}^*_i = (d^*_{1i}, d^*_{2i}, d^*_{3i}, d^*_{4i})'$. The elements of the latter are indicators of a particular SAH status:

$$d^*_{ji} = \mathbb{1}(h^*_i = j), \quad \text{for } j = 1,2,3,4;$$

that is, there are $4+K$ parameters from the outcome equation (9). There are now twenty misreporting probabilities per measure h_{mi}, $m = 1,2$, which we denote as

$$\delta^m_{k|j} = P(h_{mi} = k|h^*_i = j) \quad \forall j,k = 0,1\ldots,4, \text{ and } j \neq k.$$

When parametrised in terms of $\boldsymbol{x}_i$ as in (7), these are $20K$ parameters per measure. In addition, there are four probabilities of the distribution of SAH, $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)'$, where $\pi_j \equiv P(h_i^* = j)$, which parametrised as in (8) adds $4K$ parameters. Thus, the grand total is $45K + 4$ parameters to be estimated.

Without covariates, for instance, that is 49 parameters. As before, we can base identification and estimation on the joint distribution of (y_i, h_{1i}, h_{2i}). The joint probabilities $P(y_i = r_0,\, h_{1i} = r_1,\, h_{2i} = r_2 \,|\, \boldsymbol{x}_i) \equiv F(r_0, r_1, r_2)$ are now defined for $r_0 \in \{0, 1\}$, $r_1 \in \{0, 1, 2, 3, 4\}$, $r_2 \in \{0, 1, 2, 3, 4\}$. Thus, the joint distribution has $2 \times 5 \times 5 = 50$ support points, of which the last one is not linearly independent. The other 49 points will provide the necessary equations to identify the 49 parameters in the case without covariates. With covariates, similar arguments as before can be made. The corresponding equation to (6) in the categorical case and other details of the general model are given in Appendix A.1.

A condition to avoid "mirror solutions" in the case with multiple categories of SAH[4] is that the probability of truthfully reporting the SAH level j ($\delta^m_{j|j}$) is larger than the probability of reporting any other level:

$$\text{No Mirror Assumption (NMA2):} \qquad \delta^m_{j|j} > \delta^m_{k|j}, \quad \forall j, k,$$

which is a generalisation of NMA1 for the two category case. To implement this constraint in the estimation, we use multinomial logit-based expressions similar to the ones above (see Appendix A.1).

The model with a categorical regressor (9) is an important generalisation with respect to the binary case, which the only similar application of a model for categorical regressors in the literature being Hu (2008).

2.1.5 Interaction effects

The impact of unobserved SAH may also differ based on an individual's characteristics. Our flexible approach allows us to go further and also accommodate interaction terms between all or some of the regressors $\boldsymbol{x}_i$ and unobserved health. The model with interactions in categorical SAH status is

$$y_i = \mathbb{1}\left(\sum_{j=1}^{J} d^*_{ji}\alpha_j + \sum_{j=1}^{J} d^*_{ij} x_{ki}\alpha_{j,x} + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0\right), \quad i = 1, \ldots, N, \tag{10}$$

for some variable of interest x_{ki} such as education. To gain an intuition for the identification of the interaction effect, we imagine again that the regressors are discrete and that $\boldsymbol{x}_i$ is fully saturated. In that case, the interaction effects are obtained by simply estimating the model separately by each subsample (where each is already identified as before) without imposing the restriction that the slopes on $\boldsymbol{x}_i$ be the same across subsamples.

[4] There are 120 sets of solutions here; i.e., 120 ways to order the 5 groups that correspond to each health level.

2.2 Estimation

2.2.1 EM algorithm

The model can be estimated in a number of ways based on the joint distribution function (6), which takes the form of a finite mixture (FM) or latent class model. One option is GMM estimation, which is presented in Appendix A.3. Another option is maximum likelihood estimation, which takes the form

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}; y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \sum_{r_0} \sum_{r_1} \sum_{r_2} I_i^{r_0 r_1 r_2} \ln(F(r_0, r_1, r_2)), \tag{11}$$

where $\boldsymbol{\theta}$ collects all the parameters: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\gamma}^m_{k|j}$ for $m = 1, 2$ and $j \neq k$. $I_i^{r_0 r_1 r_2}$ is an indicator variable equal to one if $y_i = r_0$, $h_{1i} = r_1$ and $h_{2i} = r_2$. Maximisation can be implemented, in principle, directly via a standard Newton-Raphson procedure based on (11). However, we found that, especially in models with categorical health and several regressors, maximum likelihood estimation via the Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977) was substantially faster and more stable, than either GMM or standard maximum likelihood, making it our only viable estimator. The EM algorithm iterates between the maximisation or M-step, and the expectation or E-step. The nth iteration of the M-step is

$$\hat{\boldsymbol{\theta}}^n = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \tilde{\ell}_i(\boldsymbol{\theta}; y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i, \hat{w}_i^n), \tag{12}$$

where

$$\tilde{\ell}_i(\cdot) = \sum_{j=0}^{4} \hat{w}_{ji}^n \Big(\ln F(y_i|h_i^*=j, \boldsymbol{x}_i) + \ln F(h_{1i}|h_i^*=j, \boldsymbol{x}_i) + \ln F(h_{2i}|h_i^*=j, \boldsymbol{x}_i) + \ln \pi_{ji} - \ln \hat{w}_{ji}^n \Big), \tag{13}$$

and all $F(\cdot|\cdot)$ correspond to terms like those defined in (6), and the $\hat{w}_{ji}^n$ are estimates of the posterior probabilities $P(h^* = j|y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i)$. In the $(n{+}1)$th iteration of the E-step, we update these posterior probabilities as follows:

$$\hat{w}_{ji}^{n+1} = \frac{\hat{\pi}_{ji}^n \, \hat{F}^n(y_i|h_i^*=j) \, \hat{F}^n(h_{1i}|h_i^*=j) \, \hat{F}^n(h_{2i}|h_i^*=j)}{\sum_{j=0}^{4} \hat{\pi}_{ji}^n \, \hat{F}^n(y_i|h_i^*=j) \, \hat{F}^n(h_{1i}|h_i^*=j) \, \hat{F}^n(h_{2i}|h_i^*=j)}, \tag{14}$$

where all $\hat{F}^n(\cdot|\cdot)$ correspond to terms similar to the ones in (6) and are evaluated at $\hat{\boldsymbol{\theta}}^n$.

The increased stability and speed of EM comes from the fact that, first, as opposed to the likelihood $\ell_i(\cdot)$ from (11), in $\tilde{\ell}_i(\cdot)$ of the M-step, the logarithm goes through the sum of the finite mixture components of the joint distribution $F(y, h_1, h_2)$; and, second, these components depend on separate sets of parameters (since the w_{ji} are fixed in the M-step), meaning that each can be maximised separately: the first term in the parentheses, $F(y_i|h_i^*=j, \boldsymbol{x}_i)$ is a function only of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$; the second and third are functions of all the $\boldsymbol{\gamma}^1_{k|j}$ and $\boldsymbol{\gamma}^2_{k|j}$ vectors (with $j \neq k$), respectively; and $\pi_i = \pi(\boldsymbol{x}_i)$ is a function only of $\boldsymbol{\eta}$.

2.2.2 Penalisation

A potential issue in the estimation of models with a flexible parametrisation of the misclassification system as proposed here is that in finite samples there might be low statistical power to estimate the misclassification probabilities given that (i) they depend on potentially many parameters (if the dimension of $\boldsymbol{x}_i$ is large) and (ii) are potentially small. These two issues imply that in practice the misclassification probabilities may be identified from potentially low frequency cells of the joint distribution of (y_i, h_{1i}, h_{2i}).

In the most extreme case, the likelihood function for the sample at hand may be maximised for a value of a misclassification probability equal to zero, which may manifest itself as a convergence problem in the maximum likelihood procedure since parameters will tend to infinity. But even in less extreme cases, where estimates are finite, they might be severely biased. These issues, while originating in the estimates of the misclassification probabilities, may spill over to the parameter estimates of the outcome equation.

To overcome such convergence issues and reduce the small sample/low statistical power bias we suggest implementing a penalised likelihood estimation, which rules out infinite estimates and reduces extreme misclassification probabilities. For each of the ten components in (13) related to the misclassification probabilities (that is, $\ln F(h_{mi}|h_i^* = j, \boldsymbol{x}_i)$ for $m = 1, 2$ and $j = 0, \ldots, 4$), we add a ridge penalty to their objective function:

$$\hat{\boldsymbol{\gamma}}_j^m = \arg\max_{\boldsymbol{\gamma}_j^m} \sum_i \ln F(h_{mi}|h_i^* = j, \boldsymbol{x}_i) - \frac{t}{N}\, \boldsymbol{\gamma}_j^{m\prime} \boldsymbol{\gamma}_j^m \tag{15}$$

$$= \arg\max_{\boldsymbol{\gamma}_j^m} \sum_i \hat{w}_{ji} \left(\sum_k \mathbb{1}(h^m = k) \ln \delta_{k|j}^m(\boldsymbol{x}_i, \boldsymbol{\gamma}_{k|j}^m) \right) - \frac{t}{N}\, \boldsymbol{\gamma}_j^{m\prime} \boldsymbol{\gamma}_j^m; \tag{16}$$

where $\boldsymbol{\gamma}_j^m$ contains all the parameter vectors $\gamma_{k|j}^m$ for a given j and m such that $j \neq k$. The scalar t is a tuning parameter which determines the weight given to the penalty. However, as with all penalised likelihood estimations, it can introduce bias: if the penalty is too harsh (that is, if t is too large), the overall bias of the estimator may increase. Since our primary objective with the penalisation is to ensure finiteness of all estimates, we choose $t > 0$ to be as small as possible in our estimations. With increasing N, the weight of the penalisation decreases and the penalised estimator converges towards the unpenalised one.

2.2.3 System estimation

Apart from penalisation, a second possible avenue for reducing low power issues in the estimation of the misclassification is using more than one outcome variable, say $\boldsymbol{y}_i = (y_{1i}, y_{2i})$. If more than one possible outcome is available which is dependent on SAH and conditionally independent of misclassification, then the joint estimation of the outcomes can be beneficial for the accuracy of the estimation and minimising bias in small samples. We propose pooling outcomes and treating them both as con-

ditionally independent of misclassification but potentially correlated with each other. The connection between the two (or more) models is that the unobserved SAH is obviously the same for each observation across both outcome models and thus the misclassification parameters are also the same, which can be imposed as a restriction to reduce the loss of degree of freedoms relative to the case of separate estimation.[5] Adapting the EM algorithm is straightforward. In equations (12)-(14), the terms $F(y_i|h_i^* = j, \boldsymbol{x}_i)$ are simply replaced by $F(y_{1i}|h_i^* = j, \boldsymbol{x}_i)F(y_{2i}|h_i^* = j, \boldsymbol{x}_i)$.

3 Monte Carlo experiments

While the estimators presented in the last section are consistent if the conditional independence assumption is met, they are biased in finite samples. The estimation of the many parameters relating to the misclassification probabilities, for instance, can pose a challenge in practice. We examine the finite sample performance of these estimators in a Monte Carlo simulation study. We are particularly interested in the possibility of improving our proposed finite mixture (FM) estimator's performance by using the penalised FM (PFM) variant. To benchmark the performance of the estimators, we compare their performance to the ideal estimator that uses the unobserved SAH status, which is infeasible in practice. On the other end of the spectrum, we compare the performance of the proposed estimators to the naïve estimator that just uses the first observed reported SAH measure, treating it as if it was the SAH status.

The baseline design we use is a simple data generating process (DGP) with a single regressor and a binary health indicator. Survival status y_i (=1 if alive) is generated as

$$y_i = \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x + \varepsilon_i > 0\right). \tag{17}$$

Details of the parameter specification choices and the drawing procedure are given in Appendix A.4. Similar to the survey data used in our application, the reported SAH measures in our chosen simulation DGP have distributions which are broadly but not exactly similar to each other— $P(h_1 = 1) = 0.61$ and $P(h_2 = 1) = 0.57$—while at the same time there is a substantial share of conflicting answers: $P(h_1 \neq h_2)$=0.37. We use two sample sizes, $N = \{1000; 10000\}$ and replicate the estimations 500 times.

The results in Table 1 show that the infeasible estimator that uses the unobserved SAH status (in columns "h^*") is, as expected, virtually unbiased. The naïve estimator which uses the misreported SAH measure h_1 (depicted in columns "h_1"), is severely biased. The average estimate of α is about 45 percent below its true value of 1 in both sample sizes, illustrating the pernicious effects of misclassification. The remaining columns of the table present estimates from the proposed finite mixture ("FM") estimator, as well as its variant, the penalised finite mixture ("PFM") estimator. The FM estimator in samples of N=1,000 is able to greatly reduce the bias from h_1 from 46 to 4 percent for α. In samples of N=10,000,

[5] In the EM algorithm both outcomes are also used to estimate the posterior probabilities of the unobserved SAH category. Note, we are not proposing a seemingly-unrelated-regression-type approach that exploits efficiency gains through correlated errors in the outcomes.

Table 1: Simulation results: Baseline DGP

		$N = 1,000$				$N = 10,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}$	Bias	0.004	-0.459	0.041	0.018	0.002	-0.457	0.008	0.005
	RMSE	0.152	0.482	0.286	0.267	0.050	0.460	0.085	0.083
$\hat{\beta}$ const	Bias	-0.007	0.258	0.004	0.073	-0.002	0.260	-0.010	-0.000
	RMSE	0.167	0.303	0.349	0.247	0.049	0.265	0.118	0.098
$\hat{\beta}$ slope	Bias	0.014	0.169	0.017	-0.012	0.003	0.156	0.012	0.021
	RMSE	0.271	0.317	0.457	0.306	0.084	0.177	0.154	0.124
$\hat{\eta}$ const	Bias			-0.127	-0.313			0.036	0.003
	RMSE			1.142	0.591			0.393	0.289
$\hat{\eta}$ slope	Bias			-0.013	-0.002			-0.064	-0.120
	RMSE			1.560	0.552			0.517	0.363
$\hat{\gamma}^1_{1\|0}$ const	Bias			-0.005	-0.064			0.039	0.034
	RMSE			1.649	0.349			0.373	0.252
$\hat{\gamma}^1_{1\|0}$ slope	Bias			-0.272	-0.624			0.010	-0.199
	RMSE			5.787	0.735			0.612	0.424
$\hat{\gamma}^2_{1\|0}$ const	Bias			-0.218	0.149			-0.012	0.048
	RMSE			1.560	0.351			0.323	0.227
$\hat{\gamma}^2_{1\|0}$ slope	Bias			-0.027	-0.538			0.022	-0.156
	RMSE			9.004	0.672			0.547	0.380
$\hat{\gamma}^1_{0\|1}$ const	Bias			0.109	0.224			-0.010	0.007
	RMSE			0.964	0.395			0.249	0.195
$\hat{\gamma}^1_{0\|1}$ slope	Bias			0.063	-0.162			0.028	0.031
	RMSE			1.342	0.379			0.337	0.246
$\hat{\gamma}^2_{0\|1}$ const	Bias			0.024	0.388			-0.029	0.044
	RMSE			0.770	0.482			0.235	0.194
$\hat{\gamma}^2_{0\|1}$ slope	Bias			0.100	-0.342			0.049	-0.016
	RMSE			1.056	0.489			0.281	0.221

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are $\alpha = 1$, β const=0, β slope=1, η const=-0.1342, η slope=1.5, $\gamma^m_{k|j}$slope $= 1$ for all m, k, and $\gamma^1_{0|1}$const=-0.25, $\gamma^2_{0|1}$const=-0.75 , $\gamma^1_{1|0}$const=0, and $\gamma^2_{1|0}$const=-0.5. See Appendix A.4 for more details on the DGP.

the bias is less than 1 percent. The RMSE in the DGP with N=1,000 is about twice as large as that of the infeasible estimator. The other parameters of the outcome model, β_0 and β_1, are estimated similarly well.

However, for the parameters of the misclassification system, at N=1,000, there are larger biases, ranging up to about 20 percent; and even when the biases are small, the RMSE can still be substantial. It is for this issue that we see the advantages of the PFM estimator most clearly. It achieves reductions in the RMSE of these parameters that range from 50 to almost 90 percent. This improvement in the estimation of the misclassification parameters also translates into uniformly lower RMSE in the estimates of the outcome parameters, and sometimes also in bias reductions. For the estimate of α, for instance, PFM reduces FM's bias of 4 percent to less than 2 percent.

Table 2: Simulation results: DGP with interaction effect in health

		$N = 1,000$				$N = 10,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}$ const	Bias	-0.004	-0.606	0.234	-0.020	0.008	-0.596	0.020	0.002
	RMSE	0.284	0.669	0.966	0.560	0.090	0.602	0.186	0.177
$\hat{\alpha}$ slope	Bias	0.025	-0.374	-0.251	0.117	-0.011	-0.409	-0.018	0.012
	RMSE	0.560	0.660	1.392	0.966	0.178	0.442	0.295	0.286
$\hat{\beta}$ const	Bias	-0.001	0.326	-0.138	0.055	-0.003	0.324	-0.015	-0.009
	RMSE	0.212	0.380	0.820	0.349	0.063	0.330	0.149	0.131
$\hat{\beta}$ slope	Bias	-0.001	0.508	0.215	-0.007	0.006	0.514	0.018	0.024
	RMSE	0.426	0.643	1.161	0.583	0.129	0.528	0.222	0.203

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are α const=1, α slope=1, β const =-0.375, β slope=1; all misclassification parameters are kept at their baseline values (see notes of Table 1); see Appendix A.4 for more details on the DGP.

3.1 Interaction effect

The ability to easily specify interaction effects is a hallmark of our approach, and in this section we simulate from a DGP where the impact of SAH on the outcome varies with x:

$$y_i = \mathbb{1}\left(\alpha\, h_i^* + \alpha_x h_i^* x_i + \beta_0 + \beta_1 x + \varepsilon_i > 0\right), \tag{18}$$

where α_x is the coefficient on the new interaction between health and x. Table 2 shows the results from this DGP.[6] That this is a more challenging DGP can be clearly seen by observing the RMSE at N=1,000 for the infeasible estimator, which almost doubles for the constant in α (and quadruples for the slope in α, i.e. the interaction coefficient) relative to RMSE of α in the baseline case from Table 1. The FM estimator, while still improving substantially over the naïve approach, displays visible biases. The estimate of both main and interaction effect of SAH have biases of about 25 percent with N=1,000. However, the PFM estimator is able to obtain improved estimates, with biases of about 2 and 12 percent for main effect and interaction, yielding reductions in RMSE of about 50 and 40 percent relative to FM. At N=10,000, however, the FM estimator works well and the advantages of PFM over FM in this DGP are only marginal.

3.2 Multivariate outcome

Next, we consider the case of a multivariate outcome. In Table 3 we present results from estimations with two outcomes, simulated from the specification:

$$\begin{aligned} y_{1i} &= \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x + \varepsilon_{1i} > 0\right) \\ y_{2i} &= \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x + \varepsilon_{2i} > 0\right). \end{aligned}$$

[6] For space reasons, only the parameters of the outcome model are depicted.

Table 3: Simulation results: Multivariate DGP for $\boldsymbol{y} = (y_1, y_2)'$, $N = 1,000$

$\rho =$		1.00	0.75	0.50	0.25	0.00
FM						
$\hat{\alpha}$	Bias	0.059	0.039	0.015	0.007	0.001
	RMSE	0.309	0.289	0.283	0.286	0.281
$\hat{\beta}$ const	Bias	0.007	0.010	0.027	0.033	0.033
	RMSE	0.354	0.326	0.313	0.312	0.304
$\hat{\beta}$ slope	Bias	0.004	0.017	0.008	0.003	0.004
	RMSE	0.474	0.439	0.423	0.423	0.420
PFM						
$\hat{\alpha}$	Bias	0.045	0.029	0.009	0.001	-0.001
	RMSE	0.284	0.265	0.262	0.266	0.263
$\hat{\beta}$ const	Bias	0.044	0.044	0.059	0.064	0.065
	RMSE	0.249	0.233	0.228	0.227	0.230
$\hat{\beta}$ slope	Bias	0.000	0.011	0.002	0.002	-0.000
	RMSE	0.338	0.314	0.318	0.315	0.322

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the System Finite Mixture (FM) and System Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The DGP consists of two outcome equations. The parameters in the table are for the first outcome, y_1. The parameter ρ indicates the correlation between the idiosyncratic errors in the two outcomes models. The true values of the parameters in the DGP are equal to those indicated in the notes of Table 1; see Appendix A.4 for more details on the DGP.

This is a setup in the vein of "seemingly unrelated regressions". The true coefficients have been specified as having the same values across the two outcome equations, but this is merely for convenience and the estimated coefficients are allowed to vary in estimation (i.e. they are not constrained to be the same across equations). As explained previously, the gain from considering y_1 and y_2 jointly is that, since the parameters of the misclassification probabilities are the same across both outcomes, we are increasing the information (statistical power) available to estimate these parameters. The extent to which pooling both outcomes adds information depends on the degree of the dependence between the two errors, ε_1 and ε_2 (though this dependence is not estimated with our method). In the worst case, $\varepsilon_1 = \varepsilon_2$ and joint estimation will bring no advantage. Since the DGP is symmetric for y_1 and y_2, we only present estimates for equation y_1. The table presents results for N=1,000 for the cases where the correlation between the errors ε_1 and ε_2 is equal to 1, 0.75, 0.50, 0.25, and 0.

The case ρ=1 is the same as the baseline, and indeed we get very similar results. For both FM and PFM, as the correlation decreases, the estimators in general become progressively more successful at reducing the biases, although not uniformly (the bias in $\hat{\beta}_0$ increases, for instance). However, the RMSE is reduced in all cases, with the magnitude of the reduction for FM ranging from about 10 to 20 percent. Similar although often somewhat larger reductions in RMSE are achieved for the parameters of the misclassification system (see results in Appendix Table A4).

3.3 Misspecification

So far we have evaluated the performance of the FM and PFM estimators in DGPs where they correctly specify the misclassification system. We conclude this section by evaluating these proposed parametric estimators in a DGP where the misclassification probabilities are misspecified. We use the same DGP of Hu (2008), and also compare our estimator against the nonparametric instrumental variables (NPIV) estimator introduced in that paper. We have argued that the FM/PFM estimators may have two potential advantages despite the drawback of fully specifying the functional form of the misclassification probabilities and the unobserved health distribution. First, by using flexible specifications of the linear indices $\boldsymbol{x}_i'\boldsymbol{\gamma}_{k|j}^m$, many functional forms may be approximated well. Second, compared to more nonparametric approaches, even if FM/PFM might be inconsistent due to misspecified functional forms, they might still be preferable in terms of RMSE for finite samples. Here, we give some evidence of the second point. That is, we do not explore potential further improvements by specifying polynomials of $\boldsymbol{x}_i$ in the linear indices.

Details for the DGP from Hu (2008) are given in Appendix A.6. Importantly, misclassification in this DGP does not follow our logit-based functional forms. Rather, some misclassification probabilities, for instance, are partially linear functions with kinks. Table 4 shows our results for FM and PFM from this DGP, with N=500 and 200 replications as in the original Hu (2008) paper, next to the h_i^*, h_1 and Hu (2008) NPIV results from their paper. Scenarios 1 and 2 depicted in the table correspond to two variants of the DGP in which the probabilities $\delta_{0|1}^m$ depend negatively (Scenario 1) or positively (Scenario 2) on the regressor x_i. While NPIV substantially reduces the bias of the naïve estimator, for instance from about 50 percent to 12 percent for $\hat{\alpha}$ in Scenario 1, FM and PFM reduce the bias even further, and they also have the lowest RMSE of the feasible estimators presented. It could be that the good results of FM and PFM were achieved by chance: small sample bias and misspecification bias could be offsetting each other, yielding the low biases observed in the table. To check whether this was the case, we repeated the simulations for N=10,000 for PFM in Scenario 2. This resulted in biases of -0.001, 0.001, and -0.016 for $\hat{\alpha}$, '$\hat{\beta}$ const' and '$\hat{\beta}$ slope', thus dispelling the concern that an equal and opposite small sample bias might be concealing what potentially could be large misspecification biases. It also highlights that the bias due to incorrect functional form of the misclassification in this case is already small.

3.4 Further results and conclusion

Having a binary outcome is the most difficult case for correcting misclassification, as the additional information stemming from the outcome that identifies the whole system is sparse. Table A2 in the Appendix explores other nonlinear outcome models where there is more information in the dependent variable: counts and durations. In both these cases, the results indicate that FM and PFM perform even better. In Table A3 in the Appendix we present results for a DGP where categorical health is not binary, but multinomial with five categories, showing that FM and PFM also perform well in such a case. Finally, we also present a detailed look at alternative ad-hoc approaches intended to adjust

Table 4: Simulation results: Misspecified functional form of misclassification, N=500

		Scenario 1					*Scenario 2*				
		h^*	h_1	NPIV	FM	PFM	h^*	h_1	NPIV	FM	PFM
$\hat{\alpha}$	Bias	0.012	-0.520	-0.124	0.087	0.070	0.015	-0.491	-0.108	0.062	0.070
	RMSE	0.157	0.538	0.409	0.309	0.375	0.160	0.509	0.318	0.253	0.233
$\hat{\beta}$ const	Bias	0.000	0.275	0.061	-0.012	0.008	-0.001	0.263	0.052	0.006	-0.016
	RMSE	0.104	0.290	0.238	0.138	0.129	0.104	0.279	0.205	0.132	0.137
$\hat{\beta}$ slope	Bias	-0.011	-0.150	-0.138	0.014	0.020	-0.014	-0.094	-0.071	0.015	0.017
	RMSE	0.165	0.210	0.332	0.220	0.230	0.165	0.176	0.307	0.234	0.197

Notes: Cell entries show bias and root mean square error for parameters estimated over 200 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), the Nonparametric IV estimator from Hu (2008) (NPIV), the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators. For the PFM, the tuning parameter is set to $t = 0.5$. For NPIV, the results are taken from Hu (2008). The DGP is that from Hu (2008, Table 1, p.45) and given in Appendix A.6. The true values of the parameters in the DGP are α=1, β const=0.5, β slope=1. In Scenario 1, the misclassification probabilities depend negatively on x; in Scenario 2, positively.

for misclassification in Appendix A.7 and Appendix Table A1. These approaches include procedures sometimes used by practitioners, such as averaging over the two responses, keeping only observations where $h_{1i} = h_{2i}$, and mimicking instrumental variables and control function approaches from linear models. The results show that none of these inconsistent estimators can be recommended.

To summarise, the simulation results in this section illustrated a number of issues which inform our application of the estimation to real world data. First, the performance of the FM estimator can often be substantially improved, especially in smaller samples, by using the penalised version. Second, the performance might also be improved by combining outcomes and estimating them jointly. Third, our estimator is able to estimate the effects of interest reliably even under challenging circumstances such as many health categories, interaction effects, and severe misreporting, in samples of about 10,000 observations. In the next section, we will estimate a joint logit-logit model for mortality and morbidity using two five-category reported SAH measures and a sample of close to 13,000 individuals.

4 Estimating SAH misclassification and the effects of SAH on mortality and morbidity

In this section we present our estimates of SAH misclassfication and how misclassification impacts on the association between SAH and two outcomes measured 15 years later: mortality (whether the individual is deceased) and, if the individual is still alive, whether they developed any chronic conditions in the 15-year period. We first outline the HILDA data and describe the categorical self-reported health measures, for which repeated measures are available in some HILDA waves (Section 4.1). We then estimate our joint model and discuss the estimates from the misclassification system in Section 4.2 and the estimates from the outcome equations in Section 4.3.

Table 5: Joint distribution of reported SAH measures from Personal Questionnaire (h_1) and Self-Completion Questionnaire (h_2) in HILDA

Wave 1, $N = 12,908$	h_2					
h_1	0	1	2	3	4	Total
0	2.65	1.03	0.16	0.04	0.02	3.90
1	0.55	8.94	2.11	0.30	0.02	11.92
2	0.13	2.36	21.64	3.97	0.52	28.62
3	0.04	0.53	7.23	25.67	2.03	35.50
4	0.02	0.09	0.90	5.74	13.33	20.07
Total	3.38	12.95	32.04	35.71	15.92	100.00
Wave 9, $N = 11,110$	h_2					
h_1	0	1	2	3	4	Total
0	2.66	1.12	0.16	0.07	0.01	4.02
1	0.32	8.53	3.31	0.23	0.05	12.44
2	0.08	2.20	23.96	5.25	0.23	31.72
3	0.00	0.23	6.24	27.55	2.05	36.08
4	0.00	0.05	0.53	4.28	10.89	15.74
Total	3.06	12.12	34.20	37.38	13.23	100.00
Wave 13, $N = 14,993$	h_2					
h_1	0	1	2	3	4	Total
0	2.31	1.21	0.19	0.04	0.01	3.75
1	0.54	9.38	3.72	0.34	0.01	14.00
2	0.11	2.40	24.16	4.84	0.36	31.86
3	0.03	0.26	6.71	27.36	2.06	36.42
4	0.00	0.02	0.42	3.93	9.60	13.97
Total	2.98	13.27	35.20	36.50	12.05	100.00

Notes: Source: HILDA waves 1, 9 and 13. Cell entries show relative frequencies in per cent for joint and marginal distribution of the reported SAH measures h_1 and h_2. Labels: 0="poor"; 1="fair"; 2="good"; 3="very good"; 4="excellent".

4.1 Descriptive statistics

HILDA is an annual Australian household-based longitudinal survey that began in 2001 (Summerfield *et al.*, 2014). The survey covers social and economic topics such as household structure, income, work and health. Individuals aged 15 or over are asked to respond and reasons for non-response are recorded where known. Wave 1 (2001) covered a total of 7,682 households and 13,969 responding individuals. These individuals were followed up in the later waves and new household members joining the original sample were also included. A further 2,153 household and 4,009 individuals were added as a top-up sample in 2011. Overall, there are roughly 13,000 respondents in each wave of the HILDA Survey from 2001 to 2016. While non-response due to death is recorded annually where known, the survey sample was also linked in 2014 to the National Death Index so that details of individuals' year and age of death are available for all those originally in the survey, including the subsequent non-responders.

In waves 1, 9 and 13 of the survey, the SAH question is asked twice for each individual. The question is first asked as a part of the Person Questionnaire that is conducted by an interviewer face-to-face or over the phone. The SAH question is the first question in the health section, and is followed by a number of other health-related questions such as long-term conditions and disabilities. We designate this SAH variable as h_1. Respondents are asked to choose their health on a 5-option scale "Poor", "Fair", "Good", "Very Good", and "Excellent" which we label as $0, \ldots, 4$. Respondents are then issued with the self-completion Questionnaire, which is to be filled in by themselves and collected by the interviewer that day or posted back after completion. In this questionnaire, the same SAH question is asked again at the beginning. We designate this SAH variable as h_2, and label it in the same way as h_1. The dates of completing both questionnaires are only available for waves 9 and 13. The median time between completion of the two questionnaires is 1 day in both waves and on average, the questionnaires were completed only 4.8 days and 4.6 days apart in 2009 and 2013, respectively. Since the surveys were taken close together, the likelihood of an actual meaningful change in health is fairly low. As a result, we believe that the majority of differential responses to the SAH question are random and unlikely to be related to changes in an individual's underlying health status.

The top panel of Table 5 reports the joint-distribution (in percent of respondents) from the two SAH questions in wave 1. About 27.8 percent of respondents changed their health status between h_{1i} and h_{2i}, similar to that reported by Clarke & Ryan (2006) and similar to Crossley & Kennedy (2002) where SAH was asked twice in a different survey. It could be that this pattern is specific to the first wave, however, the joint distributions of h_{1i} and h_{2i} in waves 9 (N=11,110) and 13 (N=14,993) are very similar (middle and bottom panels of Table 5), and so is the share of respondents giving different answers for h_1 and h_2: 26.4 and 27.2 percent for waves 9 and 13, respectively.

Although there is a consistent percentage of individuals who revised their health status in each wave, this was not driven by the same individuals switching in each wave. The correlation of switchers (individuals who revised their response) in wave 1 and switchers in wave 9 is only 0.03 while the correlation of switchers in wave 9 and switchers in wave 13 is only 0.05, which means the vast majority of the switchers are actually new switchers from one wave to another. This increases our confidence that switching displays a large amount of randomness.

Given the two questionnaires were completed around the same time, and the percentage of switchers stays consistent over time, we conjecture that at least one of the SAH measures, if not both, is measured with some error. The marginal distributions of h_1 and h_2 given in Table 5 also reveal that individuals are more likely to select the extreme categories—"poor" (0) and "excellent" (4)—when responding to an interview (h_1) than a written questionnaire (h_2). This may suggest that compared to the self-completion mode the interviewing mode increases the chance that individuals misclassify into more extreme categories; or, alternatively, that compared to the self-completion mode the interviewing mode reduces the chance that individuals misclassify into the middle categories. Either or both cases could produce the observed joint distribution.

From here on, we focus on h_{1i} and h_{2i} for the first wave in 2001 because we want to study the implications of SAH for long-term (15-year) mortality and morbidity. There are 12,908 individuals

Table 6: Descriptive statistics

Variable	N	Mean	Std.Dev.
Covariates (Wave 1)			
age/10 (years/10)	12,908	0.438	0.176
male (=1, if yes)	12,908	0.470	0.499
education/10 (years/10)	12,908	1.272	0.203
log HH income	12,908	3.135	0.654
chronic condition (=1 if any chronic conditions in 2001)	12,908	0.233	0.423
married (=1, if married or in a relationship)	12,908	0.642	0.479
overseas (=1, if born overseas)	12,908	0.243	0.429
not in labour force (=1, if out of labour force)	12,908	0.344	0.475
unemployed (=1, if unemployed)	12,908	0.042	0.201
smoker (=1, if current or former smoker)	12,908	0.493	0.500
Outcomes (Wave 16)			
dead (=1, if deceased by 2016)	12,908	0.109	0.312
cond (=1, if any new chronic conditions in 2016 since 2001)	7,340	0.161	0.368

*N*otes: Source: HILDA waves 1 and 16.

with responses on h_{1i} and h_{2i}. Descriptive statistics for selected demographic and socio-economic characteristics of these individuals are given in Table 6. We begin our empirical investigation by applying a reduced form strategy used in previous literature to characterise individual misclassification behaviour (Black *et al.*, 2017). In Table 7, we present estimates of logit models where the dependent variable is an indicator that an individual gave two conflicting reports of SAH, $\mathbb{1}(h_{1i} \neq h_{2i})$ (Column 1), an indicator that they gave a higher SAH in the self-completion questionnaire, $\mathbb{1}(h_{1i} < h_{2i})$ (Column 2), and that they gave a higher SAH in the personal questionnaire, $\mathbb{1}(h_{1i} < h_{2i})$ (Column 3). Explantory factors include age, sex, education, income, whether individuals suffered from any chronic conditions in 2001 (*chronic condition*), whether they were married or in a relationship (*married*), whether they were born overseas (*overseas*), whether they were not in the labour force (*not in labour force*), whether they were unemployed (*unemployed*) and whether they were currently smokers or had been smokers in the past (*smoker*).

The presence of some statistically significant estimates suggest that misclassification is related to covariates. In particular, consistent with the previous literature, low education and income are strongly predictive of giving conflicting reports. However, insignificant estimates are harder to interpret. There could be different types of misclassification which 'average out', resulting in small insignificant effects on a change in reported SAH ($\mathbb{1}(h_{1i} \neq h_{2i})$). For instance, the regressor *male*, has an effect on the dependent variable '$\mathbb{1}(h_{1i} < h_{2i})$ despite not having a significant effect on $\mathbb{1}(h_{1i} \neq h_{2i})$. In addition, more complex misclassification patterns can be completely undetectable with these dependent variables and they also tell us nothing about which of the two measures are more likely to be misclassified. By estimating our finite mixture model, we can go beyond these reduced-form patterns in misclassification

Table 7: Estimation results: Logit models for changes in SAH response

Dep. var.	$\mathbb{1}(h_1 \neq h_2)$ (1)	$\mathbb{1}(h_1 > h_2)$ (2)	$\mathbb{1}(h_1 < h_2)$ (3)
age/100	-0.17	0.88	-0.79
	(0.65)	(0.94)	(0.76)
age^2/100	1.01	-0.83	1.91**
	(0.67)	(0.98)	(0.78)
male	0.05	0.23**	-0.08
	(0.04)	(0.06)	(0.05)
education/10	-0.59**	-0.44**	-0.55**
	(0.11)	(0.16)	(0.13)
log HH income	-0.13**	-0.16**	-0.07*
	(0.03)	(0.05)	(0.04)
chronic condition	-0.14**	0.27**	-0.39**
	(0.05)	(0.07)	(0.06)
married	0.13**	-0.02	0.19**
	(0.05)	(0.07)	(0.06)
overseas	0.21**	0.22**	0.15**
	(0.05)	(0.07)	(0.05)
not in labour force	0.03	0.11	-0.03
	(0.05)	(0.08)	(0.06)
unemployed	0.05	0.10	0.01
	(0.10)	(0.14)	(0.12)
smoker	0.03	-0.03	0.06
	(0.04)	(0.06)	(0.05)
mean dep. var.	0.278	0.102	0.176
N	12,908	12,908	12,908

Notes: Source: HILDA wave 1, own calculations. Cells represent estimated coefficients, and robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$

and instead examine the underlying misclassification probabilities which generate these patterns.

A joint model for mortality and development of new chronic conditions adjusting for misreported SAH

We estimate models for mortality and, conditional on survival, whether individuals report the development of any new chronic condition in the 15 years after the initial survey, where these probabilities could be affected by SAH, age, gender, basic indicators of socio-economic status (education, household income) along with whether they were in a relationship, born overseas, in the labour force, unemployed and either were or had been a smoker. There were 12,908 individuals in the 2001 survey of which 10.9% were deceased by 2016 and we can obtain information on the possible development of new chronic conditions in 2016 for 7,340 individuals. Means and standard deviations for these outcome variables are reported in Table 6, along with those of the covariates. We estimate the two outcomes jointly using the penalised finite mixture estimator, using the same $\boldsymbol{x}_i$ specification to parametrise the

misclassification probabilities according to equation (21) and the same low tuning parameter (t=0.5). We first examine the estimated misclassification patterns and then explore what bias these patterns cause for the estimated outcome equations when misclassification is ignored.

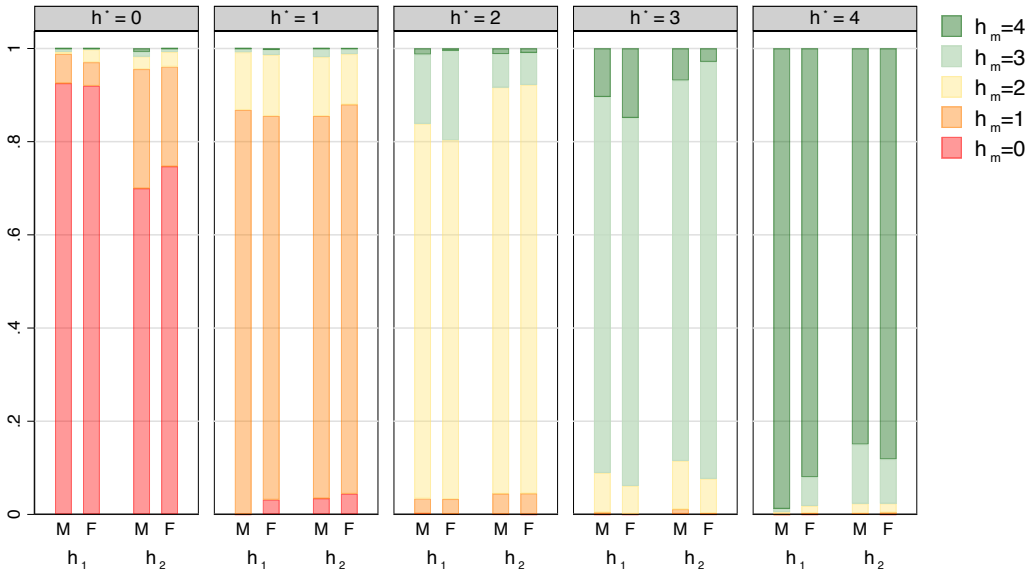
4.2 SAH and misclassification

Figure 1(a) shows the average predicted probabilities of reporting behaviour by gender. Each of the five panels illustrates how males (M) and females (F) in each health state ($h^* = j$) are likely to respond when answering h_1 and h_2. We find there is substantial misclassification in both reported SAH measures but slightly more so for the self-completion questionnaire h_2 with this being concentrated in those with excellent and poor health. As expected we see that most misclassification is by only one category with other larger misclassifications rare. As already noted there was a discrepancy between the two SAH measures in those reporting being in "excellent" health, with "excellent" health being reported more often in face-to-face interviews (h_1) than when filling out the questionnaire privately in the self-completion questionnaire (h_2). The results seen in (a) suggest that this is because males in excellent health ($h^* = 4$) are more likely to under-report their health in the self-completion questionnaire (very few males in excellent health under-report their health in the face-to-face case) and males and females in very good health ($h^* = 3$) are more likely to over-report their health in the face-to-face questionnaire. For males and females in good health ($h^* = 2$) we also observe similar over-reporting patterns with nearly 20 percent over-reporting their health in the face-to-face case compared to just under 10 percent over-reporting in the privately answered questionnaire. Conversely, for males and females in poor health ($h^* = 0$) the share truthfully reporting their health status is much higher in the face-to-face interviews with over 20 percent over-reporting their health in the self-completion questionnaire. Perhaps poor health individuals are more honest in the face-to-face case because their poor health is more evident to the interviewer or they can verbally justify claiming they have poor health. For those in fair and good health we see very little evidence of under-reporting.

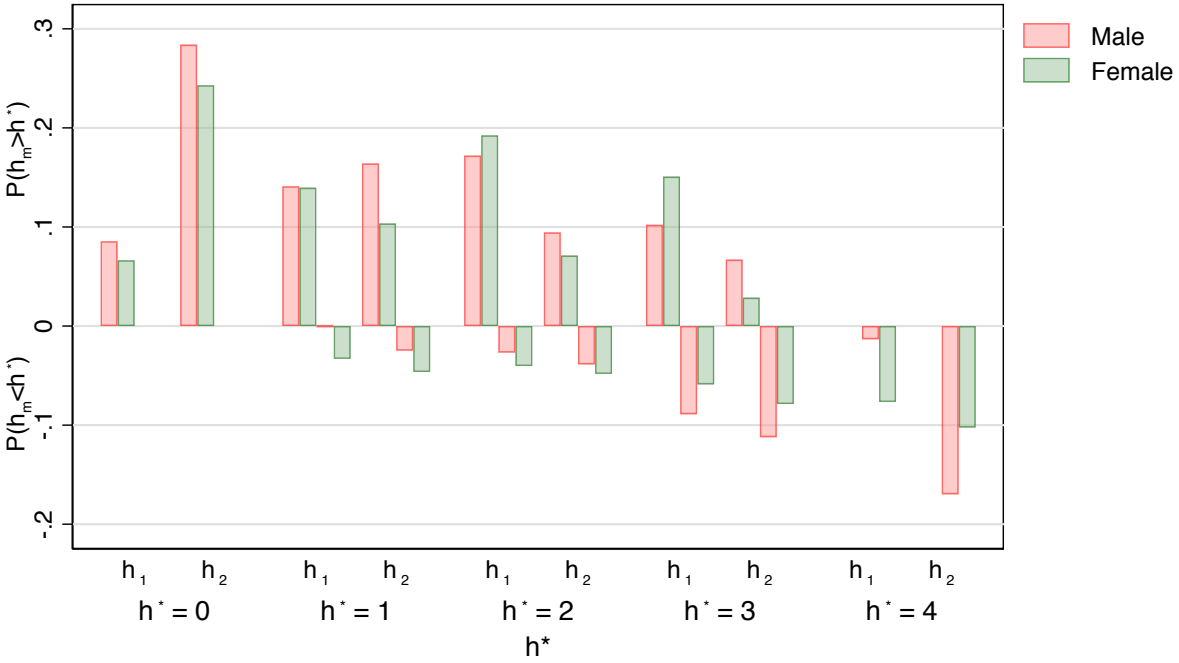
Figure 1(b) shows the average predicted posterior misclassification probabilities grouped into upward and downward misclassification when those in each SAH status is assumed to be male and female respectively (i.e. the difference between the two gives the average marginal effect of gender on misclassification). Here we see that the patterns are largely similar to those observed when we look at the misclassification for each subgroup in the population i.e. the partial role of gender is not being masked by the role of other individual characteristics on misclassification in the subgroup analysis seen in (a). While Figure 1 considers misclassification probabilities for males and females there is also considerable heterogeneity in the individual probabilities of reporting health status truthfully by other individual characteristics. In the appendix we present the equivalent figures but where we consider the role of income, age and education on misclassification Figures A1, A2 and A3. Of note we see that after controlling for other characteristics low income individuals are more likely to over-report their health compared to high income individuals and older people are more likely to misclassify in general compared to younger people while there are limited differences in misclassification related to education apart from the extreme categories in the self-completion questionnaire.

Figure 1: Misclassification in SAH for males and females

(a) Reporting for males (M) and females (F)



(b) Average predicted upward and downward misreporting



Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SRH questions in wave 1 (N=12,908). In Panel (a), weighted average predicted probabilities are presented for the separate samples of males and females. In Panel (b), weighted average predicted probabilities are presented when probabilities are evaluated at male=0 and then male=1. While in (a) differences in reporting may also be due to other charateristics which differ across males and females, (b) attempts to isolate the role of gender itself on reporting behaviour such that the differences give the average marginal effects of gender on misreporting.

4.3 Impact of misclassification on the outcome model

Table 8 contains the estimated parameters of our penalised finite mixture models for mortality and morbidity (Columns 1 and 4) along with the difference compared to what the estimates would have

Table 8: Estimation results: System Penalised Finite Mixture (PFM) models for mortality (dead: yes/no) and morbidity (chronic condition: yes/no)

Dep. var.	Dead			Chronic cond.		
	PFM	Diff. to naïve		PFM	Diff. to naïve	
		h_1	h_2		h_1	h_2
	(1)	(2)	(3)	(4)	(5)	(6)
α_1	-0.80**	-0.06	-0.17	-0.15	0.01	-0.23*
	(0.14)	(0.04)	(0.11)	(0.17)	(0.08)	(0.12)
α_2	-1.13**	-0.10**	-0.23**	-0.41**	0.03	-0.21**
	(0.15)	(0.05)	(0.09)	(0.17)	(0.08)	(0.10)
α_3	-1.46**	-0.12*	-0.21**	-0.71**	-0.02	-0.25**
	(0.16)	(0.06)	(0.09)	(0.18)	(0.08)	(0.10)
α_4	-1.77**	-0.24**	-0.26**	-1.11**	-0.16*	-0.37**
	(0.20)	(0.09)	(0.12)	(0.20)	(0.09)	(0.11)
age/100	-3.90**	-0.11	-0.50**	6.28**	-0.08	-0.28**
	(1.56)	(0.08)	(0.12)	(1.39)	(0.09)	(0.09)
age^2/100	13.20**	0.04	0.56**	-3.08**	0.00	0.29**
	(1.44)	(0.08)	(0.14)	(1.49)	(0.10)	(0.11)
male	0.58**	-0.00	-0.02**	-0.12*	0.01	-0.01
	(0.08)	(0.00)	(0.01)	(0.07)	(0.00)	(0.00)
education/10	-0.12	0.03**	-0.01	-0.52**	0.01	-0.00
	(0.22)	(0.01)	(0.02)	(0.18)	(0.01)	(0.01)
log HH. income	-0.10*	0.01**	0.02**	-0.17**	0.01**	0.02**
	(0.06)	(0.00)	(0.01)	(0.06)	(0.00)	(0.00)
chronic condition	0.27**	-0.03*	-0.07**	0.39**	-0.00	-0.05**
	(0.09)	(0.02)	(0.02)	(0.09)	(0.02)	(0.02)
married	-0.38**	-0.01**	0.01	-0.14*	0.00	0.01*
	(0.08)	(0.00)	(0.01)	(0.08)	(0.00)	(0.00)
overseas	-0.25**	0.01**	0.02**	-0.05	0.01**	0.01**
	(0.09)	(0.00)	(0.01)	(0.08)	(0.00)	(0.00)
not in labour force	0.07	-0.02**	-0.05**	0.13	-0.00	-0.02**
	(0.11)	(0.01)	(0.01)	(0.09)	(0.01)	(0.01)
unemployed	0.07	0.01	0.03**	0.31*	0.01	0.01*
	(0.25)	(0.01)	(0.01)	(0.17)	(0.01)	(0.01)
smoker	0.60**	0.00	0.02**	0.29**	0.00	0.01
	(0.08)	(0.01)	(0.01)	(0.07)	(0.00)	(0.00)
N	12,908	12,908	12,908	7,340	7,340	7,340

*N*otes: Source: HILDA waves 1 and 16, own calculations. Bootstrap standard errors in parentheses. Columns (1) and (4) present the coefficients of the joint PFM model estimated for the binary dependent variables 'Dead' and 'Chronic cond.'. Columns (2), (3) and (5), (6) show the PFM coefficient estimate minus the coefficient estimate from models using h_1 or h_2 (and the corresponding bootstrapped standard error for this difference). The independent variable 'chronic condition' refers to whether the individual had a chronic condition in 2001, while the dependent variable 'Chronic cond.' refers to whether they had developed an additional chronic condition by 2016. * $p < 0.10$, ** $p < 0.05$

been using naïve estimators h_1 and h_2 (Columns 2,3,5 and 6). For the key parameters of interest, the health coefficients α_j, the differences to the naïve approaches are often significant and range from small to large. For mortality the PFM estimates differ by about 10 percent to 20 percent but some PFM estimates of α_j are more than twice as big in the chronic condition equation. In most cases we find that by using our PFM approach there are larger difference in health outcomes between SAH categories than when misclassification is ignored. The biases in estimates of α_j using the face-to-face responses h_1 are smaller than when using the self-completion responses. There are also significant biases in the estimated coefficients on other explanatory variables, but in most cases these biases are small (in relative terms for those factors that are highly associated with health outcomes or in absolute terms of those that are not strongly associated with the outcomes) and in most cases less than 10 percent in relative terms. For example, in our PFM model for both mortality and future chronic conditions we now find that income has a significantly smaller effect after we take into account that low income individuals are more likely to over-report their SAH (i.e. some of their poorer future health outcomes are actually due to their current poorer SAH that they do not always disclose). Some of the largest effects in absolute terms of ignoring misclassification in the outcome equation are for chronic conditions in 2001 and whether they are in the labour force. The direction of the bias in both these cases suggest those with chronic conditions and those not in the labour force are more likely to under-report their SAH.

As a sensitivity analysis, we also estimated specifications where we replaced the continuous variables age, education and household income by sets of dummy variables. The estimation results can be found in Appendix Table A5 (and additional descriptive statistics for the discretised variables in Table A6). We find broadly similar results to the ones in our baseline specification with continuous regressors, although differences tend to be somewhat larger.

Finally, we also estimated specifications with interaction effects in SAH. We run four separate specifications where we interacted health with education, household income, sex, and age (results for these specifications are in Table A7 in the appendix). We found little evidence for significant differences to the naïve approach for either income and age. And while there were significant differences in both mortality and chronic conditions for education, and in mortality for sex when misclassification was taken into account in the PFM model, in terms of odds of mortality and developing future chronic conditions these differences did not substantially change the patterns of the odds ratios for those in each SAH category (see Appendix Figure A4).

5 Conclusions

While previous literature has documented that a large share of individuals report different SAH when asked twice, several important questions raised by this issue have so far remained unanswered. Because many forms of misclassification are compatible with given observed differences in reported SAH, questions such as whether reported SAH is inherently unreliable or whether observed differences stem from one particular deficient measure could not be addressed. Similarly, it was not possible to know

what patterns of misclassification occur nor how these vary with individual characteristics. Given that SAH is arguably one of the most widely used variables to measure health status or health capital in health economics, a key question is also how this misclassification in reported SAH translates to biases in estimates of models where reported SAH is used as a regressor. The question is not only if the effects of SAH are biased by misclassification, but also whether these biases spill over to the effects of other regressors.

Making use of recent advances in the econometrics of misclassification, we provide answers to these questions with data from a prominent household survey where SAH was reported twice in the same wave. Thanks to the setup of two measurements and at least one outcome, it is possible to identify the entire system of misclassification probabilities and the effects of SAH on the outcome without the need for additional arbitrary instruments or exclusion restrictions (Hu, 2008). Another advantage of the approach is that it specifies the effects of the categorical SAH variable directly by including dummy variables for each category of SAH in the outcome model, which is the standard way in which the applied literature includes categorical SAH variables as regressors in models. This avoids the additional modelling step of linking categorical SAH to some latent underlying continuous health, which would require additional assumptions.

Our results showed that there is substantial misclassification in both reported SAH measures, face-to-face interviews and self-completion questionnaires, and that misclassification patterns further vary by individual characteristics. When considering the role of current SAH in predicting future mortality and chronic conditions, comparisons between our proposed PFM approach and the naïve approaches using the misclassified responses showed that the naïve approaches were affected by statistically significant biases that ranged in magnitude from small to large. One result worth noting is that there were smaller biases in the outcome equations when the face-to-face questionnaire responses were used compared to the self-completion responses. We found significant but small biases for the role of explanatory variables other than SAH when misclassification is ignored. The small magnitude of these differences suggest that the use of SAH as a control variable might not be badly compromised despite the large shares of inconsistent answers in SAH. This is a result which might be useful to researchers who do not have multiple measure of SAH available but rely on including SAH in their empirical analysis as a useful way of addressing omitted variable bias from health status. However, when trying to estimate the role of SAH on outcomes it is likely to be more important to account for misclassification, especially if SAH is obtained through a self-completion questionnaire.

In this paper, we focussed on explaining the large share of different responses by the same individuals when asked to report SAH twice. This type of misclassification, which is virtually uncorrelated over time as could be seen from the data, fits our assumption of conditional independence and our modelling of misclassification as conditionally random. However, if there is additional misclassification that is not conditionally independent/random then this remaining misclassification remains hidden. So while adjusting for conditionally random misclassification will improve estimates, it might not fully account for all misclassification.

The proposed method for nonlinear regression models where the key regressor is a categorical health

variable and two misclassified measures of the regressor are available, can naturally be applied to contexts other than SAH. Our simulation results on its finite sample performance provide guidance to other practitioners working with misclassified categorical regressors. The use of *ad-hoc* methods such as averaging the responses or restricting the sample to individuals with the same responses in both measures cannot be recommended in most cases; nor can the use of two-stage prediction inclusion or residual inclusion. Sample sizes in the order of 10,000 observations seem to be necessary to achieve reliable estimates when using a misclassified regressor with many categories. Finite sample bias is also expected to be smaller with dependent variables with more possible outcomes, such as counts or durations, compared to binary dependent variables. Using a penalised estimator can visibly improve the performance of the estimator, especially when there is limited statistical power. Using several dependent variables jointly can also help reduce finite sample bias. Thus, our parametric approach with penalisation may have advantages over nonparametric approaches in finite samples where there is limited statistical power (e.g small sample size or many explanatory factors over which the misclassification probabilities may vary). Finally, in principle, estimates of the misclassification parameters from one study can be used to adjust key outcome parameters from another study using the assumption that the nature of misclassification is constant across both studies. This might be especially useful for exploring the sensitivity to misclassification in studies which only have one mis-measured SAH variable or have small sample sizes available.

References

[1] Au, Nicole & David W Johnston, 'Self-assessed health: What does it mean and what does it hide?' *Social Science & Medicine*, **121**, pp. 21–28, *2014*.

[2] Baker, Michael, Mark Stabile, & Catherine Deri, 'What do self-reported, objective, measures of health measure?' *Journal of Human Resources*, **39** (4), pp. 1067–1093, *2004*.

[3] Basu, Anirban & Norma Coe, '2SLS vs 2SRI: Appropriate methods for rare outcomes and/or rare exposures.' *Unpublished mansucript, University of Washington, Seattle*, *2015*.

[4] Battistin, Erich, Michele De Nadai, & Barbara Sianesi, 'Misreported schooling, multiple measures and returns to educational qualifications.' *Journal of Econometrics*, **181** (2), pp. 136–150, *2014*.

[5] Black, Nicole, David W Johnston, Michael A Shields, & Agne Suziedelyte, 'Who provides inconsistent reports of their health status? The importance of age, cognitive ability and socioeconomic status.' *Social Science & Medicine*, **191**, pp. 9–18, *2017*.

[6] Brown, Sarah, Mark N Harris, Preety Srivastava, & Xiaohui Zhang, 'Modelling illegal drug participation.' *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181** (1), pp. 133–154, *2018*.

[7] Butler, Joseph S, Richard V Burkhauser, Jean M Mitchell, & Theodore P Pincus, 'Measurement error in self-reported health variables.' *Review of Economics and Statistics*, pp. 644–650, *1987*.

[8] Clarke, Philip M & Chris Ryan, 'Self-reported health: reliability and consequences for health inequality measurement.' *Health Economics*, **15** (6), pp. 645–652, *2006*.

[9] Crossley, Thomas F & Steven Kennedy, 'The reliability of self-assessed health status.' *Journal of Health Economics*, **21** (4), pp. 643–658, *2002*.

[10] Currie, Janet & Brigitte C Madrian, 'Health, health insurance and the labor market.' *Handbook of labor economics*, **3**, pp. 3309–3416, *1999*.

[11] Dempster, Arthur P, Nan M Laird, & Donald B Rubin, 'Maximum likelihood from incomplete data via the EM algorithm.' *Journal of the royal statistical society. Series B*, pp. 1–38, *1977*.

[12] Doiron, Denise, Denzil G Fiebig, Meliyanni Johar, & Agne Suziedelyte, 'Does self-assessed health measure health?' *Applied Economics*, **47** (2), pp. 180–194, *2015*.

[13] Gosling, Amanda & Eirini-Christina Saloniki, 'Correction of misclassification error in disability rates.' *Health Economics*, **23** (9), pp. 1084–1097, *2014*.

[14] Greene, William, Mark N Harris, Preety Srivastava, & Xueyan Zhao, 'Misreporting and econometric modelling of zeros in survey data on social bads: An application to cannabis consumption.' *Health economics*, **27** (2), pp. 372–389, *2018*.

[15] Hu, Yingyao, 'Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution.' *Journal of Econometrics*, **144** (1), pp. 27–61, *2008*.

[16] ——— , 'The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics.' *Journal of Econometrics*, **200** (2), pp. 154–168, *2017*.

[17] Kane, Thomas J, Cecilia Elena Rouse, & Douglas Staiger, 'Estimating returns to schooling when schooling is misreported.' *NBER Working Paper Series, Paper 7235*, *1999*.

[18] Lindeboom, Maarten & Marcel Kerkhofs, 'Health and work of the elderly: subjective health measures, reporting errors and endogeneity in the relationship between health and work.' *Journal of Applied Econometrics*, **24** (6), pp. 1024–1046, *2009*.

[19] Mossey, Jana M & Evelyn Shapiro, 'Self-rated health: a predictor of mortality among the elderly..' *American Journal of Public Health*, **72** (8), pp. 800–808, *1982*.

[20] Newey, Whitney K, 'Series estimation of regression functionals.' *Econometric Theory*, **10** (1), pp. 1–28, *1994*.

[21] Schennach, Susanne M, 'Recent advances in the measurement error literature.' *Annual Review of Economics*, **8**, pp. 341–377, *2016*.

[22] Staub, Kevin E, 'Simple tests for exogeneity of a binary explanatory variable in count data regression models.' *Communications in Statistics: Simulation and Computation*, **38** (9), pp. 1834–1855, *2009*.

[23] Summerfield, Michelle, Simon Freidin, Markus Hahn, Peter Ittak, Ning Li, Ninette Macalalad, Nicole Watson, Roger Wilkins, & Mark Wooden, 'HILDA User Manual–Release 13.' *Melbourne Institute of Applied Economic and Social Research, University of Melbourne*, *2014*.

[24] Terza, Joseph V, Anirban Basu, & Paul J Rathouz, 'Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling.' *Journal of Health Economics*, **27** (3), pp. 531–543, *2008*.

[25] Wooldridge, Jeffrey M, 'Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables.' *Journal of Econometrics*, **182** (1), pp. 226–234, *2014*.

Appendix

A.1 General model with categorical health

The pendant to equation (6) in the case of an ordinal regressor h_i^* is

$$\begin{aligned} F(r_0, r_1, r_2) &= \sum_{j=0}^{4} p_j^* \, F(r_0, r_1, r_2 | h_i^* = j) \\ &= \sum_{j=0}^{4} p_j^* \, F(r_0 | h_i^* = j) F(r_1 | h_i^* = j) F(r_2 | h_i^* = j), \end{aligned} \tag{19}$$

where, naturally, $\pi_0 = 1 - \sum_{j=1}^{4} \pi_j$.

For $F(r_0|h_i^*)$ we have:

$$\begin{aligned} F(r_0|h_i^* = j) &= \Lambda\left(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}\right)^{r_0} \left(1 - \Lambda\left(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-r_0} \quad \text{for } j = 1, 2, 3, 4 \\ F(r_0|h_i^* = 0) &= \Lambda\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)^{r_0} \left(1 - \Lambda\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-r_0} \end{aligned}$$

And for $F(r_m|h^*)$:

$$F(r_m|h_i^* = j) = (\delta_{0|j}^m)^{\mathbb{1}(r_m=0)} (\delta_{1|j}^m)^{\mathbb{1}(r_m=1)} (\delta_{2|j}^m)^{\mathbb{1}(r_m=2)} (\delta_{3|j}^m)^{\mathbb{1}(r_m=3)} (\delta_{4|j}^m)^{\mathbb{1}(r_m=4)}, \quad \text{for } j = 1, 2, 3, 4.$$

In this formula, there is always one $\delta_{k|j}^m$ with $j = k$. These are defined as

$$\delta_{j|j}^m = P(h_{mi} = j | h_i^* = j) = 1 - \sum_{k \neq j} \delta_{k|j}^m. \tag{20}$$

NMA2, the condition to avoid mirror solutions in the case with multiple categories of SAH is that the probability of truthfully reporting a health level j is larger than any probability of misreporting it:

$$\delta_{j|j}^m > \delta_{k|j}^m, \quad \forall j, k.$$

To implement this constraint in the estimation, we use the following parametrisation of misreporting probabilities:

$$\delta_{k|j}^m = \frac{\exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{k|j}^m))}{1 + \sum_{k:k\neq j} \exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{k|j}^m))}, \tag{21}$$

which in turn is a generalisation of (7). With covariates, the multinomial logit model is the natural generalisation of (8), the model for unobserved health conditional on $\boldsymbol{x}_i$:

$$\pi_{j,i} \equiv P(h_i^* = j | \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\eta}_j)}{1 + \sum_{j=0}^{4} \exp(\boldsymbol{x}_i'\boldsymbol{\eta}_j)}. \tag{22}$$

A.2 Counts and durations: Poisson and Weibull models

The proposed approach, which we have presented for a logit binary outcome, can be extended to many common nonlinear models that follow the form

$$f(y_i|h_i^*, \boldsymbol{x}_i) = g(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta};\ \omega), \tag{23}$$

where $f(\cdot|\cdot)$ is a functional of the conditional distribution of y_i given health status h_i^* and a $K\times 1$ vector of covariates $\boldsymbol{x}_i$, and $g(\cdot)$ is a known nonlinear function, which might include ancillary parameters ω. To avoid notational clutter, h_i^* is binary. Typical examples for $f(y_i|h_i^*, \boldsymbol{x}_i)$ include it being a survival rate (probability), the time until developing a health condition (hazard rate), the number of doctor visits (count), or expenditures for health care (nonlinear expectation).

For instance, if y_i follows a Poisson distribution we might use the specification

$$P(y_i|h_i^*, \boldsymbol{x}_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad \lambda_i = \exp(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{24}$$

where the left-hand-side of (24) corresponds to $f(\cdot|\cdot)$ and the right-hand-side to $g(\cdot)$. We can use the EM algorithm described in (12)-(14) to estimate this model directly, simply by replacing $F(y_i|h_i^*)$ in those equations by $P(y_i|h_i^*, \boldsymbol{x}_i)$ from (24). Alternatively, one could also base estimation of the Poisson model on its expectation $E(y_i|h_i^*, \boldsymbol{x}_i) = \lambda_i$ and use the GMM approach based on moment conditions

$$E\Big((y_i - \lambda_i)\boldsymbol{x}_i\Big) = 0, \tag{25}$$

where, here, $f(y_i|h_i^*, \boldsymbol{x}_i) = E(y_i|h_i^*, \boldsymbol{x}_i)$ and $g(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta}) = \lambda_i$.

Similarly, if y_i was a duration and followed a Weibull distribution with parameters λ_i and ω, we could estimate the model using the EM algorithm. The corresponding $F(y_i|h_i^*)$ term in this case would simply be the probability density function

$$f(y_i|h_i^*, \boldsymbol{x}_i) = \lambda_i\omega y_i^{\omega-1}\exp(-\lambda y_i^{\omega}), \qquad \lambda_i = \exp(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta}). \tag{26}$$

A.3 GMM estimator

To estimate the model by GMM, we use the indicator variables $I_i^{r_0r_1r_2}$, defined as

$$I_i^{r_0r_1r_2} \equiv \mathbb{1}(y_i = r_0,\ h_{1i} = r_1,\ h_{2i} = r_2),$$

and which are equal to one if all their arguments are true, and equal to zero otherwise. We then base estimation on the $7 \times K$ moment conditions of the form

$$E\big(\ \big[I_i^{r_0r_1r_2} - F_i(r_0, r_1, r_2)\big]\,\boldsymbol{x}_i\ \big) = 0, \tag{27}$$

for seven unique values of the triplet (r_0, r_1, r_2) —e.g., (0,0,0), (0,0,1), (0,1,0), etc.—, and where K is the number of regressors in $\boldsymbol{x}_i$ including a constant. (The eighth variable, say I_i^{111}, is linearly dependent of the other seven; as is $F(1,1,1)$ of the other seven $F(r_0, r_1, r_2)$. Thus, the eighth

equation provides no additional information and is discarded.) We obtain, $\hat{\boldsymbol{\theta}}$, an estimate for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', p^*, \delta^1_{0|1}, \delta^2_{0|1}, \delta^1_{1|0}, \delta^2_{1|0})$, by solving the GMM minimisation problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \boldsymbol{Q}_i(\boldsymbol{\theta})' \boldsymbol{W}_N \boldsymbol{Q}_i(\boldsymbol{\theta}), \tag{28}$$

where the $[7K \times 1]$-vector of moment conditions is

$$\boldsymbol{Q}_i(\boldsymbol{\theta}) = \begin{pmatrix} \left[I_i^{000} - F_i(0,0,0)\right] \boldsymbol{x}_i \\ \left[I_i^{001} - F_i(0,0,1)\right] \boldsymbol{x}_i \\ \left[I_i^{010} - F_i(0,1,0)\right] \boldsymbol{x}_i \\ \left[I_i^{011} - F_i(0,1,1)\right] \boldsymbol{x}_i \\ \left[I_i^{100} - F_i(1,0,0)\right] \boldsymbol{x}_i \\ \left[I_i^{101} - F_i(1,0,1)\right] \boldsymbol{x}_i \\ \left[I_i^{110} - F_i(1,1,0)\right] \boldsymbol{x}_i \end{pmatrix},$$

and $\boldsymbol{W}_N$ is a $[7K \times 7K]$ positive definite weighting matrix with plim $\boldsymbol{W}$. The weighting matrix $\boldsymbol{W}_N$ may be specified as the identity matrix, or estimated in an optimal two-step approach. Note that the i subscript for the joint probabilities $F_i(r_0, r_1, r_2)$ stems from the dependence of these terms on $\boldsymbol{x}_i$.

A.4 Details of the baseline simulation DGP

In the baseline design, $\boldsymbol{x}_i = (1, x_i)$, where $x_i \sim U(0,1)$; health status h_i^* is drawn from a Bernoulli distribution with probability π_i; ε_i, from a logistic distribution. Survival status y_i (=1 if alive) is generated as

$$y_i = \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x + \varepsilon_i > 0\right).$$

We use the four misreporting probabilities $\delta^1_{0|1}$, $\delta^2_{0|1}$, $\delta^1_{1|0}$ and $\delta^2_{1|0}$ to generate the two reported health measures h_{1i} and h_{2i}. Specifically, for observations with $h_i^* = 1$ we draw h_{mi} from a Bernoulli distribution with probability $1 - \delta^m_{0|1}$; and for observations with $h_i^* = 0$ we draw h_{mi} from a Bernoulli distribution with probability $\delta^m_{1|0}$. Thus, jointly, the four misreporting probabilities, the parameter governing the distribution of unobserved health, and the parameters of the outcome equation α, β_0, β_1 determine endogenously the distribution of the survival outcome y_i, and the distribution of the reported health measures h_{mi}. The parameter values are specified as $\alpha = 1$, $\beta_0 = 0$ and $\beta_1 = 1$. Misreporting probabilities are parametrised as

$$\delta^m_{k|j} = \Lambda\big(-\exp(\gamma^m_{k|j}\text{const} + \gamma^m_{k|j}\text{slope}\; x_i)\big), \quad m = 1,2, \quad j \neq k = 0,1,$$

with all four slope parameters $\gamma^m_{k|j}$slope $= 1$, and the four constants $\gamma^1_{0|1}$const=-0.25, $\gamma^2_{0|1}$const=-0.75 , $\gamma^1_{1|0}$const=0, and $\gamma^2_{1|0}$const=-0.5. The distribution of h_i^* is given by

$$\pi_i = \Lambda(\eta_0 + \eta_1 x_i),$$

with η_1=1.5 and η_0=-0.1342.

The simulation DGP implies that the marginal probability of being in good health is $P(h^* = 1) = 0.7$. The reported health measures have marginal distributions $P(h_1 = 1) = 0.61$ and $P(h_2 = 1) = 0.57$. The share of conflicting answers is $P(h_1 \neq h_2)$=0.37. The average misreporting probabilities are about 0.21 ($\delta^1_{0|1}$), 0.31 ($\delta^2_{0|1}$)), 0.16 ($\delta^1_{1|0}$)) and 0.26 ($\delta^2_{1|0}$)).

Sample sizes are $N = \{1000; 10000\}$ and the number of replications is 500.

A.5 Simulation DGP for multinomial health with five categories

Here we present simulation results for models with a discrete SAH measure with five categories, $h^* = 0, \ldots, 4$. We simulate from the following DGP:

$$y_i = \mathbb{1}\left(\alpha_1 h^*_{1i} + \alpha_2 h^*_{2i} + \alpha_3 h^*_{3i} + \alpha_4 h^*_{4i} + \beta_0 + \beta_1 x + \varepsilon_i > 0\right), \tag{29}$$

where we specify $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)' = (0.5, 1.0, 1.5, 2.0)'$. The parameters β_0 and β_1 are set to -1 and 1. We specify the misreporting probabilities as

$$\delta^m_{k|j,i} = \frac{\exp(-\exp(\gamma^m_{k|j}\text{const} + \gamma^m_{k|j}\text{slope } x_i))}{1 + \sum_{k \neq j} \exp(-\exp(\gamma^m_{k|j}\text{const} + \gamma^m_{k|j}\text{slope } x_i))}, \qquad \text{for } j \neq k,$$

and set all slope parameters equal to 1, $\gamma^m_{k|j}$slope=1, and specify the constants as $\gamma^m_{k|j}\text{const} = 0.25|j-k|$. The marginal distribution of unobserved health is specified as $\boldsymbol{\pi} = (0.10, 0.15, 0.20, 0.25, 0.30)$ by setting

$$\pi_{ji} = \frac{\exp(\eta_j\text{const} + \eta_j\text{slope } x_i)}{1 + \sum_{j=1}^{4} \exp(\eta_j\text{const} + \eta_j\text{slope } x_i)}, \quad j = 1, 2, 3, 4,$$

with slopes equal to 1.0, 2.0, 2.0 and 2.5, and constant chosen such as to yield the marginal distribution specified above. This DGP is more challenging not only in that it has more parameters, but also in that misreporting is much more prevalent. About 61 percent of individuals report different values for h_1 and h_2. For roughly half of these, 31 percent, the discrepancy between the first and second SAH measure is 1. Discrepancies of 2, 3, and 4 occur in 18, 9, and 3 percent of individuals. The $\delta^m_{k|j,i}$ vary between about 2 and 20 percent. To the best of our knowledge, this is the first simulation evidence of this type of DGP of a categorical regressor with flexible effects.

The results of the simulation for the parameters of the outcome model are collected in Table A3. Again, that this is a more challenging DGP can be seen in the biases and RMSE that are apparent in the infeasible estimator. We see that at N=1,000, FM and PFM show some visible biases, in the order of about 8 to 16 percent. However, in the larger sample size these biases have all but disappeared, with the maximum bias in FM being less than 4 percent and that in PFM less than 2 percent.

A.6 Details of the simulation DGP in Hu (2008)

We use the setup of Hu (2008) as reported in Hu (2008, Table 1, p.45).

The DGP is for a probit outcome y_i, a binary misclassified regressor h_i^*, and a normally distributed covariate x_i. We adjust our outcome model to be a probit, but leave the misclassification probabilities and π_i as logistic, while in the DGP they are not.

Specifically, the DGP is

$$P(y_i = 1|h_i^*, x_i) = \Phi(\alpha h_i^* + \beta_0 + \beta_1 x_i),$$

where $\Phi(\cdot)$ denotes the standard normal CDF, α=1, β_0=0.5 and $\beta_1 = 1$, and $x_i \sim N(0, 0.25)$. Health status is defined as $h_i^* = \mathbb{1}(\epsilon < 0.6)$, where $\epsilon \sim Uniform(0, 1)$. The reported health measures h_1 and h_2 are defined as follows: $h_{2i} = \mathbb{1}(\epsilon + \delta < 0.6)$, where $\delta \sim N(0.0.04)$. For h_{1i},

$$P(h_{1i} = 0|h_i^* = 1, x_i) = \min(1, \max(0, p_i)), \quad P(h_{1i} = 1|h_i^* = 0, x_i) = \min(1, \max(0, q_i)).$$

In Table 4, for results in Panel "*Scenario 1*",

$$p_i = 0.3 - 0.1x_i,$$
$$q_i = 0.2 + 0.1x_i;$$

and for results in Panel "*Scenario 2*",

$$p_i = 0.3 + 0.1x_i,$$
$$q_i = 0.2 + 0.1x_i.$$

A.7 Ad-hoc approaches to adjust for misclassification

Here we examine four potential competitor estimators, which address the misclassification in an ad-hoc way and are sometimes encountered in the literature.

First, we experiment by using the average of the two SAH measures as the regressor in the models (in tables, we denote this estimator as"$\bar{h}$"). If the measurement error were classical, this approach would produce an (unbiased) SAH measure with less measurement error, thus mitigating some of the bias. A second simple ad-hoc way of addressing the misclassification is to drop all individuals from the estimation sample whose second response to the SAH question is different from the first ("$\bar{\bar{h}}$"). This leaves a sample of individuals with what sometimes is called "consistent responses". It is clear that this is also a procedure leading to biased estimates, since some of the individuals in such a sample may have misreported their SAH status twice. Moreover, this procedure results in a reduced sample size and, therefore, less precise estimates. Nevertheless, similar to the averaging of the SAH responses, the severity of the misclassification problem might be mitigated by this approach.

The last two estimators included in the simulation correspond to approaches that mimic two-stage least squares in linear models. They consist of using one SAH measure as an instrument for the other. Both estimators use the same first stage in which one SAH measure is regressed on the other. The first of

Table A1: Simulation results: Ad-hoc misclassification approaches

		h^*	h_1	$\bar{h}$	$\bar{\bar{h}}$	$\hat{h}_1$	$\hat{e}_1$
$N = 1,000$							
$\hat{\alpha}$	Bias	0.004	-0.459	-0.259	-0.243	0.632	0.675
	RMSE	0.152	0.482	0.321	0.307	0.953	0.989
$\hat{\beta}$ const	Bias	-0.007	0.258	0.162	0.163	-0.262	-0.276
	RMSE	0.167	0.303	0.234	0.265	0.452	0.465
$\hat{\beta}$ slope	Bias	0.014	0.169	0.152	0.054	-0.138	-0.134
	RMSE	0.271	0.317	0.308	0.346	0.359	0.359
$N = 10,000$							
$\hat{\alpha}$	Bias	0.002	-0.457	-0.259	-0.245	0.602	0.643
	RMSE	0.050	0.460	0.268	0.253	0.647	0.686
$\hat{\beta}$ const	Bias	-0.002	0.260	0.165	0.158	-0.245	-0.258
	RMSE	0.049	0.265	0.172	0.171	0.272	0.284
$\hat{\beta}$ slope	Bias	0.003	0.156	0.141	0.057	-0.144	-0.140
	RMSE	0.084	0.177	0.164	0.120	0.179	0.176

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), the average of h_1 and h_2 ($\bar{h}$), h_1 in the sample restricted to i with $h_{1i} = h_{2i}$, predicted h_1 ($\hat{h}_1$) and the residual from a prediction of h_1 ($\hat{e}_1$). The true values of the parameters in the DGP are α=1, β const=0, β slope=1. See Appendix A.4 for more details on the DGP.

these estimators then includes the first-stage predictions as the regressor in the outcome model ("$\hat{h}_1$"). This approach is inconsistent, in general, for nonlinear models, but it is often applied by practitioners. The second estimator includes the first-stage residuals as an additional regressor along the mismeasured SAH response in the outcome model ("$\hat{e}_1$"). This is a version of the control function approach and is valid for nonlinear models under certain conditions. In general, for instance, the endogenous regressor (here, SAH) needs to be continuous. There are, however, specific forms of endogeneity under which the control function approach is consistent with a discrete endogenous regressor (see, for instance, the setup used in Terza *et al.*, 2008). And even when it is inconsistent, the control function approach has been advocated as a potentially useful remedy that might not cure the problem but reduce it in some circumstances (Basu & Coe, 2015; Wooldridge, 2014).[7]

The results in Table A1 show that the infeasible estimator that uses the unobserved SAH status (in column "h^*") is as expected virtually unbiased. The naïve estimator which uses the misreported SAH measure h_1 (depicted in column "h_1"), is severely biased. The average estimate of α is about 45 percent below its true value of 1 in both sample sizes, illustrating the pernicious effects of misreporting. The following two columns show the results obtained by using the two common *ad hoc* fixes for reducing misreporting bias, averaging the two available measures, and keeping only observations with the same reported SAH across both measures. The bias in the estimated α is about -25 percent for both estimators. Thus, these procedures improve over the estimation using a single reported measure, but the bias is still very large.

[7]The control function approach might also be useful if the focus is on testing rather than estimation. Some tests might be valid even when the estimator is inconsistent (Wooldridge, 2014; Staub, 2009).

The columns "$\hat{h}_1$" and "$\hat{e}_1$" report the results for the possible *ad hoc* methods related to IV estimation. All estimated parameters, including the slope of x, are very distorted overestimating the true value on average by about 63 and 67 percent. Thus, such approaches, while well-suited to measurement error in linear models, cannot be recommended as solutions to the measurement error problem at hand. We see that for all these four *ad-hoc* approaches the estimated root mean squared error (RMSE) is driven primarily by the bias. As these biases do not vanish with larger sample sizes, the RMSE approaches the bias as variances shrink with increasing N.

A.8 Additional results

Table A2: Simulation results: Counts (Poisson) and Durations (Weibull) DGPs

		Poisson, $N = 1,000$				Weibull, $N = 1,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}$	Bias	-0.004	-0.469	0.001	-0.008	-0.000	-0.469	0.005	-0.007
	RMSE	0.049	0.474	0.073	0.073	0.072	0.470	0.117	0.113
$\hat{\beta}$ const	Bias	0.001	0.197	-0.002	0.005	0.005	0.191	0.008	0.019
	RMSE	0.055	0.210	0.101	0.095	0.076	0.193	0.141	0.117
$\hat{\beta}$ slope	Bias	0.005	0.126	-0.001	0.005	-0.002	0.129	-0.000	0.015
	RMSE	0.058	0.173	0.085	0.077	0.113	0.134	0.198	0.155
$\hat{\eta}$ const	Bias			0.003	0.011			-0.034	-0.065
	RMSE			0.316	0.278			0.514	0.375
$\hat{\eta}$ slope	Bias			0.009	-0.043			0.036	-0.021
	RMSE			0.459	0.403			0.773	0.522
$\hat{\gamma}^1_{1\|0}$ const	Bias			-0.028	0.113			-0.014	0.115
	RMSE			0.483	0.287			0.629	0.317
$\hat{\gamma}^1_{1\|0}$ slope	Bias			0.118	-0.259			0.137	-0.366
	RMSE			0.880	0.482			1.234	0.572
$\hat{\gamma}^2_{1\|0}$ const	Bias			-0.028	0.209			-0.124	0.198
	RMSE			0.448	0.326			0.728	0.338
$\hat{\gamma}^2_{1\|0}$ slope	Bias			0.043	-0.368			0.211	-0.417
	RMSE			0.722	0.549			1.317	0.621
$\hat{\gamma}^1_{0\|1}$ const	Bias			-0.022	0.063			0.024	0.131
	RMSE			0.262	0.217			0.410	0.287
$\hat{\gamma}^1_{0\|1}$ slope	Bias			0.034	-0.084			-0.002	-0.139
	RMSE			0.363	0.299			0.580	0.363
$\hat{\gamma}^2_{0\|1}$ const	Bias			-0.042	0.156			-0.045	0.209
	RMSE			0.366	0.286			0.442	0.334
$\hat{\gamma}^2_{0\|1}$ slope	Bias			0.053	-0.209			0.072	-0.249
	RMSE			0.485	0.390			0.584	0.433

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are $\alpha = 1$, β const=0, β slope=1, η const=-0.1342, η slope=1.5, $\gamma^m_{k|j}$slope $= 1$ for all m, k, and $\gamma^1_{0|1}$const=-0.25, $\gamma^2_{0|1}$const=-0.75 , $\gamma^1_{1|0}$const=0, and $\gamma^2_{1|0}$const=-0.5. For the Weibull DGP, the true value of ω=1.5. Poisson and Weibull models were parametrised as described in Appendix A.2. See Appendix A.4 for more details on the simulation DGP.

Table A3: Simulation results: DGP with multinomial health ($h^* = 0, 1 \ldots, 4$)

		$N = 1,000$				$N = 10,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}_1$	Bias	0.056	-0.298	0.166	0.105	0.017	-0.293	0.039	0.013
	RMSE	0.392	0.379	0.786	0.738	0.094	0.302	0.175	0.169
$\hat{\alpha}_2$	Bias	0.033	-0.521	0.095	0.057	0.005	-0.534	0.028	0.012
	RMSE	0.301	0.570	0.381	0.574	0.093	0.539	0.157	0.149
$\hat{\alpha}_3$	Bias	0.055	-0.741	0.161	0.147	0.003	-0.754	0.019	0.001
	RMSE	0.307	0.772	0.533	0.612	0.082	0.758	0.155	0.145
$\hat{\alpha}_4$	Bias	-0.123	-0.926	0.078	0.127	0.003	-0.937	0.027	0.010
	RMSE	0.285	0.951	0.453	0.571	0.087	0.940	0.130	0.131
$\hat{\beta}$ const	Bias	-0.026	0.648	-0.123	-0.082	-0.011	0.660	-0.030	-0.009
	RMSE	0.241	0.670	0.419	0.517	0.077	0.662	0.128	0.124
$\hat{\beta}$ slope	Bias	0.110	0.133	0.150	0.039	0.005	0.165	0.005	0.019
	RMSE	0.266	0.264	0.275	0.278	0.071	0.180	0.073	0.077

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are $\alpha = 0.5$, $\alpha = 1$, $\alpha = 1.5$, $\alpha = 2$, β const=-1, β slope=1. See Appendix A.5 for more details on the DGP.

Table A4: Simulation results: Full results—Multivariate DGP $\boldsymbol{y} = (y_1, y_2)'$, $N = 1,000$

		FM					PFM				
$\rho =$		1.00	0.75	0.50	0.25	0.00	1.00	0.75	0.50	0.25	0.00
$\hat{\alpha}$	Bias	0.059	0.039	0.015	0.007	0.001	0.045	0.029	0.009	0.001	-0.001
	RMSE	0.309	0.289	0.283	0.286	0.281	0.284	0.265	0.262	0.266	0.263
$\hat{\beta}$ const	Bias	0.007	0.010	0.027	0.033	0.033	0.044	0.044	0.059	0.064	0.065
	RMSE	0.354	0.326	0.313	0.312	0.304	0.249	0.233	0.228	0.227	0.230
$\hat{\beta}$ slope	Bias	0.004	0.017	0.008	0.003	0.004	0.000	0.011	0.002	0.002	-0.000
	RMSE	0.474	0.439	0.423	0.423	0.420	0.338	0.314	0.318	0.315	0.322
$\hat{\eta}$ const	Bias	-0.145	-0.142	-0.153	-0.151	-0.118	-0.242	-0.223	-0.229	-0.229	-0.216
	RMSE	1.157	1.093	1.038	1.009	0.987	0.652	0.587	0.578	0.554	0.544
$\hat{\eta}$ slope	Bias	-0.009	0.013	0.028	0.059	0.035	-0.047	-0.047	-0.040	-0.035	-0.034
	RMSE	1.562	1.509	1.446	1.400	1.395	0.676	0.643	0.627	0.611	0.613
$\hat{\gamma}^1_{1\|0}$ const	Bias	-0.093	-0.087	-0.078	-0.121	-0.099	-0.072	-0.060	-0.070	-0.071	-0.061
	RMSE	1.601	1.525	1.384	1.385	1.321	0.444	0.412	0.394	0.378	0.372
$\hat{\gamma}^1_{1\|0}$ slope	Bias	0.019	0.025	0.057	0.163	0.214	-0.479	-0.467	-0.454	-0.442	-0.435
	RMSE	2.811	2.756	2.594	2.519	2.477	0.741	0.726	0.706	0.696	0.684
$\hat{\gamma}^2_{1\|0}$ const	Bias	-0.201	-0.267	-0.257	-0.297	-0.216	0.114	0.116	0.117	0.114	0.118
	RMSE	1.645	1.939	1.598	1.998	1.466	0.455	0.414	0.406	0.394	0.387
$\hat{\gamma}^2_{1\|0}$ slope	Bias	0.297	0.368	0.339	0.417	0.312	-0.436	-0.426	-0.434	-0.435	-0.424
	RMSE	2.525	2.767	2.500	2.842	2.371	0.717	0.687	0.685	0.679	0.660
$\hat{\gamma}^1_{0\|1}$ const	Bias	0.094	0.114	0.108	0.096	0.073	0.148	0.131	0.130	0.131	0.125
	RMSE	0.960	0.964	0.894	0.871	0.846	0.454	0.403	0.387	0.380	0.371
$\hat{\gamma}^1_{0\|1}$ slope	Bias	0.066	0.015	0.026	0.027	0.054	-0.048	-0.041	-0.042	-0.045	-0.047
	RMSE	1.310	1.287	1.235	1.209	1.214	0.489	0.459	0.451	0.447	0.443
$\hat{\gamma}^2_{0\|1}$ const	Bias	0.068	0.048	0.038	0.030	0.012	0.291	0.276	0.281	0.276	0.265
	RMSE	0.774	0.710	0.691	0.701	0.640	0.474	0.447	0.440	0.427	0.413
$\hat{\gamma}^2_{0\|1}$ slope	Bias	0.050	0.058	0.063	0.062	0.070	-0.225	-0.219	-0.227	-0.226	-0.220
	RMSE	1.044	0.960	0.950	1.014	0.872	0.497	0.481	0.478	0.470	0.461

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. See Appendix A.4 for more details on the DGP.

Table A6: Descriptive statistics for additional discretised variables

Variable	*N*	Mean	Std.Dev.
Covariates (Wave 1)			
age: 30s (=1 if 30 years≤age< 40 years)	12,908	0.209	0.407
age: 40s (=1 if 40 years≤age< 50 years)	12,908	0.200	0.400
age: 50s (=1 if 50 years≤age< 60 years)	12,908	0.150	0.358
age: 60s (=1 if 60 years≤age< 70 years)	12,908	0.102	0.303
age: 70 plus (=1 if age≥ 70 years)	12,908	0.101	0.301
education: year 12 (=1 if highest education Year 12)	12,908	0.145	0.353
education: certificate (=1 if highest education certificate)	12,908	0.256	0.437
education: bachelor (=1 if highest education bachelor or higher)	12,908	0.178	0.382
HH income, 2nd quint. (=1 if HH income in 2nd quintile)	12,908	0.200	0.400
HH income, 3rd quint. (=1 if HH income in 3rd quintile)	12,908	0.200	0.400
HH income, 4th quint. (=1 if HH income in 4th quintile)	12,908	0.200	0.400
HH income, 5th quint. (=1 if HH income in 5th quintile)	12,908	0.200	0.400

*N*otes: Source: HILDA waves 1.

Table A5: Estimation results: Specification with discretised continuous variables

Dep. var.	Dead			Chronic cond.		
	PFM	Diff. to naïve		PFM	Diff. to naïve	
		h_1	h_2		h_1	h_2
	(1)	(2)	(3)	(4)	(5)	(6)
α_1	-0.78**	-0.08	-0.17*	-0.16	-0.00	-0.23**
	(0.13)	(0.06)	(0.09)	(0.17)	(0.10)	(0.11)
α_2	-1.12**	-0.14**	-0.22**	-0.44**	0.00	-0.23**
	(0.14)	(0.06)	(0.08)	(0.17)	(0.09)	(0.10)
α_3	-1.45**	-0.16**	-0.20**	-0.74**	-0.03	-0.26**
	(0.15)	(0.07)	(0.08)	(0.17)	(0.09)	(0.10)
α_4	-1.72**	-0.25**	-0.25**	-1.14**	-0.17*	-0.37**
	(0.20)	(0.11)	(0.10)	(0.20)	(0.10)	(0.11)
age: 30s	1.27**	-0.00	0.01	0.56**	0.01	0.01**
	(0.26)	(0.01)	(0.01)	(0.13)	(0.00)	(0.01)
age: 40s	1.66**	-0.01*	-0.02*	0.86**	-0.01	-0.01
	(0.25)	(0.01)	(0.01)	(0.12)	(0.01)	(0.01)
age: 50s	2.41**	-0.03**	-0.02	1.08**	-0.00	0.01
	(0.24)	(0.01)	(0.01)	(0.13)	(0.01)	(0.01)
age: 60s	3.61**	-0.01	0.01	1.43**	-0.02**	0.01
	(0.24)	(0.01)	(0.01)	(0.14)	(0.01)	(0.01)
age: 70 plus	5.16**	-0.04**	0.03**	1.72**	-0.05**	0.00
	(0.24)	(0.01)	(0.01)	(0.17)	(0.01)	(0.01)
male	0.58**	0.00	-0.02**	-0.15**	0.00	-0.01**
	(0.08)	(0.01)	(0.01)	(0.07)	(0.00)	(0.00)
education: year 12	0.14	0.04**	0.04**	-0.10	0.02**	0.02**
	(0.14)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
education: certificate	-0.13	0.00	-0.00	-0.03	-0.01	-0.01
	(0.09)	(0.01)	(0.01)	(0.08)	(0.01)	(0.00)
education: bachelor	-0.07	0.01	-0.01	-0.31**	0.00	-0.01
	(0.13)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 2nd quint.	-0.06	-0.01	-0.00	-0.25**	-0.01*	-0.00
	(0.10)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 3rd quint.	-0.13	-0.00	0.01	-0.22**	-0.00	0.01*
	(0.12)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 4th quint.	-0.03	0.02**	0.03**	-0.23**	0.00	0.01**
	(0.12)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 5th quint.	-0.26*	0.04**	0.04**	-0.29**	0.03**	0.03**
	(0.14)	(0.01)	(0.01)	(0.12)	(0.01)	(0.01)
chronic condition	0.30**	-0.06**	-0.08**	0.40**	-0.02	-0.06**
	(0.09)	(0.02)	(0.02)	(0.09)	(0.02)	(0.01)
married	-0.53**	-0.01*	0.00	-0.11	-0.01*	-0.00
	(0.08)	(0.01)	(0.01)	(0.08)	(0.00)	(0.00)
overseas	-0.24**	0.00	0.02**	-0.02	0.01*	0.01**
	(0.08)	(0.01)	(0.01)	(0.08)	(0.01)	(0.00)
not in labour force	0.20*	-0.03**	-0.05**	0.11	-0.00	-0.02**
	(0.11)	(0.01)	(0.01)	(0.09)	(0.01)	(0.01)
unemployed	0.05	-0.03*	-0.01	0.27	-0.02	-0.01
	(0.26)	(0.02)	(0.01)	(0.17)	(0.01)	(0.01)
smoker	0.49**	-0.00	0.01	0.30**	-0.00	0.00
	(0.08)	(0.01)	(0.01)	(0.07)	(0.00)	(0.00)
N	12,908	12,908	12,908	7,340	7,340	7,340

*N*otes: Source: HILDA waves 1 and 16, own calculations. See notes in Table 8 for more information.
* $p < 0.10$, ** $p < 0.05$

Table A7: Estimation results: System PFM specifications with interactions in health (and difference to naïve estimator using h_1)

	Interaction w. education			
	Dead	*diff.*	Cond.	*diff.*
educ	-0.43	-0.11	-3.15**	-1.06*
	(0.76)	(0.14)	(1.07)	(0.55)
α_1: educ	0.28	-0.07	3.49**	1.32*
	(0.90)	(0.26)	(1.16)	(0.68)
α_1: cons	-1.12	0.03	-4.37**	-1.57*
	(1.09)	(0.32)	(1.41)	(0.81)
α_2: educ	0.43	0.46*	2.66**	1.17**
	(0.83)	(0.24)	(1.10)	(0.58)
α_2: cons	-1.64	-0.65**	-3.61**	-1.37*
	(1.02)	(0.30)	(1.34)	(0.70)
α_3: educ	0.38	-0.05	2.57**	0.97*
	(0.85)	(0.26)	(1.10)	(0.58)
α_3: cons	-1.91*	-0.04	-3.80**	-1.15
	(1.05)	(0.34)	(1.34)	(0.71)
α_4: educ	-0.02	0.05	2.49**	1.12*
	(1.03)	(0.37)	(1.18)	(0.64)
α_4: cons	-1.70	-0.27	-4.09**	-1.49*
	(1.29)	(0.48)	(1.45)	(0.78)
N	12,908	12,908	7,340	7,340

	Interaction w. log HH income			
	Dead	*diff.*	Cond.	*diff.*
lnehi	-0.16	0.05	-0.19	-0.07
	(0.18)	(0.04)	(0.24)	(0.11)
α_1: lnehi	0.14	0.05	0.07	0.08
	(0.21)	(0.07)	(0.27)	(0.13)
α_1: cons	-1.18**	-0.19	-0.35	-0.22
	(0.58)	(0.18)	(0.79)	(0.36)
α_2: lnehi	0.01	-0.06	0.01	0.09
	(0.20)	(0.06)	(0.26)	(0.12)
α_2: cons	-1.15**	0.09	-0.43	-0.24
	(0.56)	(0.17)	(0.75)	(0.34)
α_3: lnehi	0.03	-0.03	0.01	0.08
	(0.21)	(0.08)	(0.26)	(0.12)
α_3: cons	-1.51**	-0.02	-0.74	-0.24
	(0.60)	(0.23)	(0.76)	(0.34)
α_4: lnehi	0.18	-0.09	-0.05	0.08
	(0.28)	(0.12)	(0.29)	(0.13)
α_4: cons	-2.31**	0.04	-0.95	-0.39
	(0.87)	(0.38)	(0.88)	(0.40)
N	12,908	12,908	7,340	7,340

	Interaction w. male			
	Dead	*diff.*	Cond.	*diff.*
male	0.55**	-0.04	-0.17	-0.04
	(0.24)	(0.05)	(0.30)	(0.12)
α_1: male	-0.10	-0.03	0.18	0.07
	(0.28)	(0.09)	(0.34)	(0.16)
α_1: cons	-0.74**	-0.04	-0.24	-0.03
	(0.21)	(0.06)	(0.23)	(0.12)
α_2: male	0.26	0.17**	-0.20	0.01
	(0.27)	(0.09)	(0.32)	(0.13)
α_2: cons	-1.27**	-0.19**	-0.34	0.01
	(0.22)	(0.07)	(0.23)	(0.11)
α_3: male	0.03	-0.05	0.11	0.08
	(0.29)	(0.11)	(0.32)	(0.13)
α_3: cons	-1.46**	-0.08	-0.77**	-0.05
	(0.23)	(0.09)	(0.23)	(0.11)
α_4: male	-0.63*	-0.20	0.55	0.25
	(0.38)	(0.16)	(0.37)	(0.16)
α_4: cons	-1.43**	-0.13	-1.37**	-0.29**
	(0.28)	(0.12)	(0.27)	(0.13)
N	12,908	12,908	7,340	7,340

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

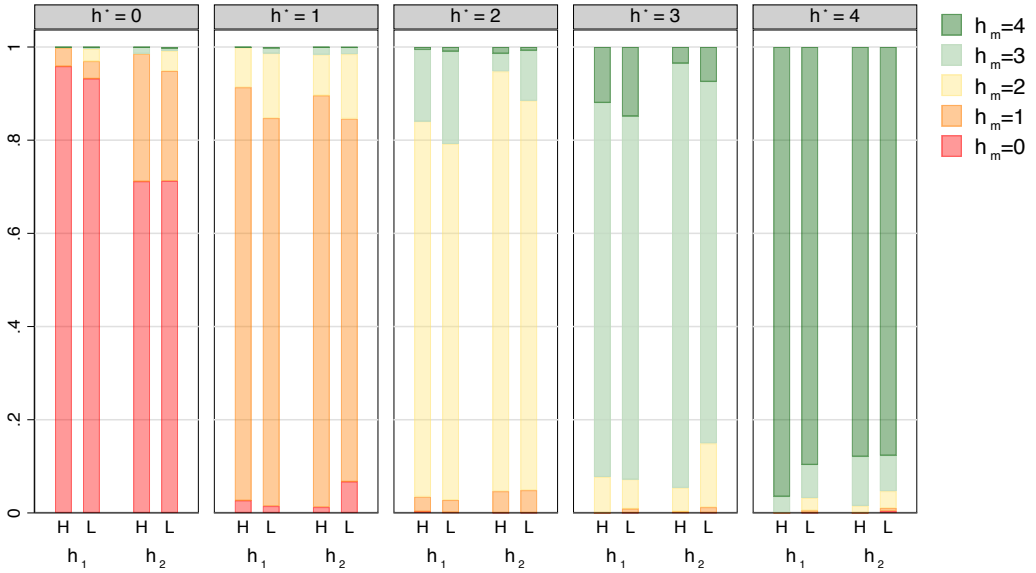
	Interaction w. age			
	Dead	*diff.*	Cond.	*diff.*
age	-3.18	-4.19	9.04	1.76
	(6.44)	(3.30)	(7.38)	(3.96)
agesq	13.17**	3.48	-8.47	-1.92
	(5.70)	(2.66)	(7.61)	(3.78)
α_1: age	0.86	4.90	-3.62	-1.56
	(7.33)	(4.52)	(7.96)	(5.22)
α_1: agesq	-1.47	-4.12	4.72	1.56
	(6.43)	(3.67)	(8.22)	(5.07)
α_1: cons	-0.76	-1.46	0.38	0.35
	(2.06)	(1.36)	(1.88)	(1.30)
α_2: age	1.83	3.00	-2.47	-2.73
	(6.99)	(3.57)	(7.71)	(5.50)
α_2: agesq	-2.18	-2.44	5.19	3.01
	(6.19)	(2.96)	(7.96)	(5.34)
α_2: cons	-1.38	-0.97	-0.52	0.60
	(1.95)	(1.06)	(1.82)	(1.36)
α_3: age	-4.55	3.51	-3.01	-1.55
	(6.95)	(3.61)	(7.73)	(6.82)
α_3: agesq	3.63	-2.92	6.65	1.81
	(6.25)	(3.05)	(8.00)	(6.66)
α_3: cons	-0.10	-1.10	-0.91	0.31
	(1.90)	(1.05)	(1.82)	(1.68)
α_4: age	-0.43	7.64*	-6.67	-3.73
	(8.05)	(4.08)	(8.19)	(10.13)
α_4: agesq	-0.13	-6.61*	9.83	3.76
	(7.31)	(3.51)	(8.53)	(10.14)
α_4: cons	-1.44	-2.28*	-0.37	0.70
	(2.17)	(1.17)	(1.91)	(2.39)
N	12,908	12,908	7,340	7,340

Standard errors in parentheses

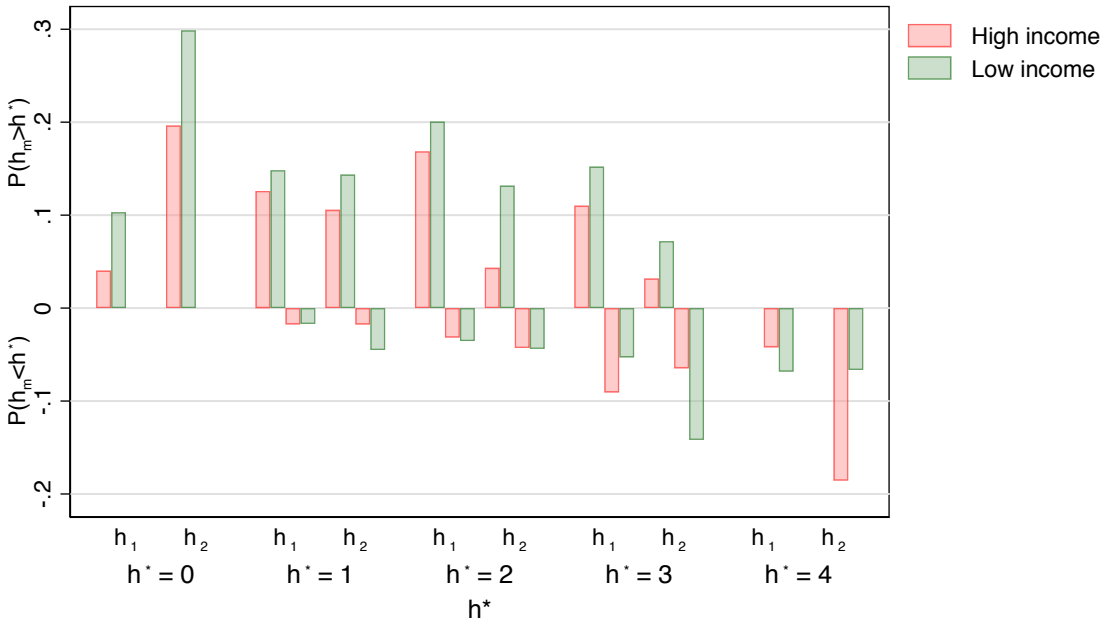
* $p < 0.10$, ** $p < 0.05$

Figure A1: Misclassification in SAH for low and high income individuals

(a) Reporting for high (H) and (L) income individuals



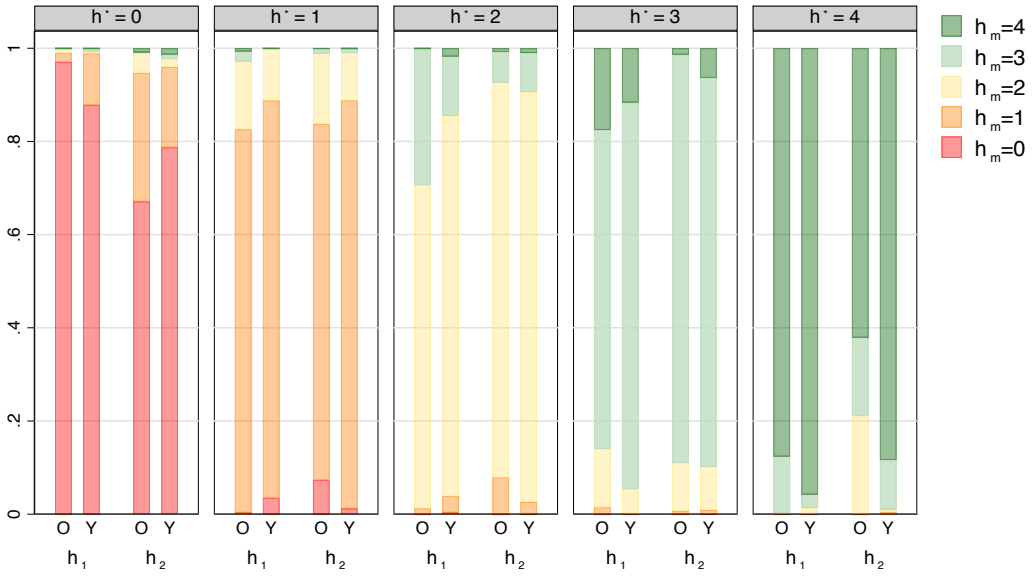
(b) Average predicted upward and downward misreporting



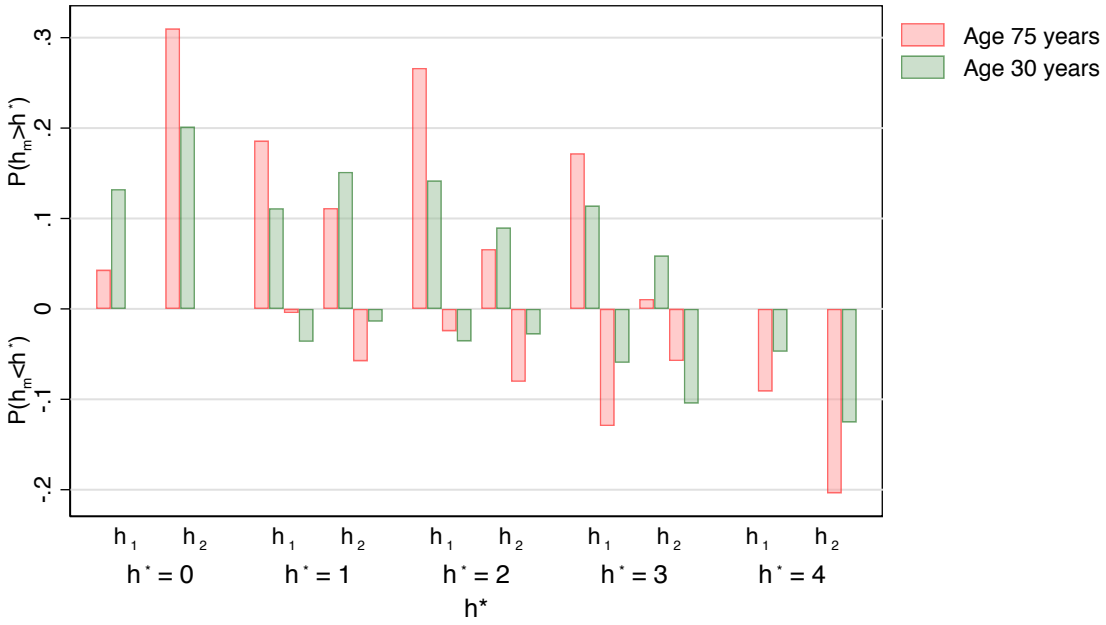
Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SAH questions in wave 1. In Panel (a), high income individuals are those in the top quintile and low income individuals those in the bottom quintile of the distribution of equivalised yearly household income. In Panel (b), predicted probabilities are averaged over the whole sample and evaluated at the mean income of the top quintile (High income) and the mean income of the bottom income (Low income).

Figure A2: Misclassification in SAH for old and young individuals

(a) Reporting for old (O) and young (Y) individuals



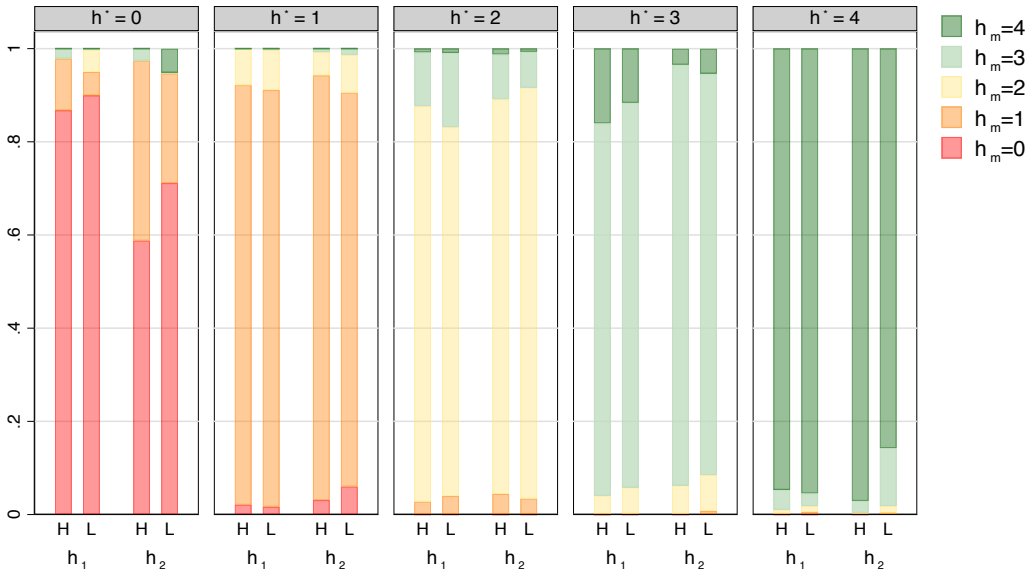
(b) Average predicted upward and downward misreporting



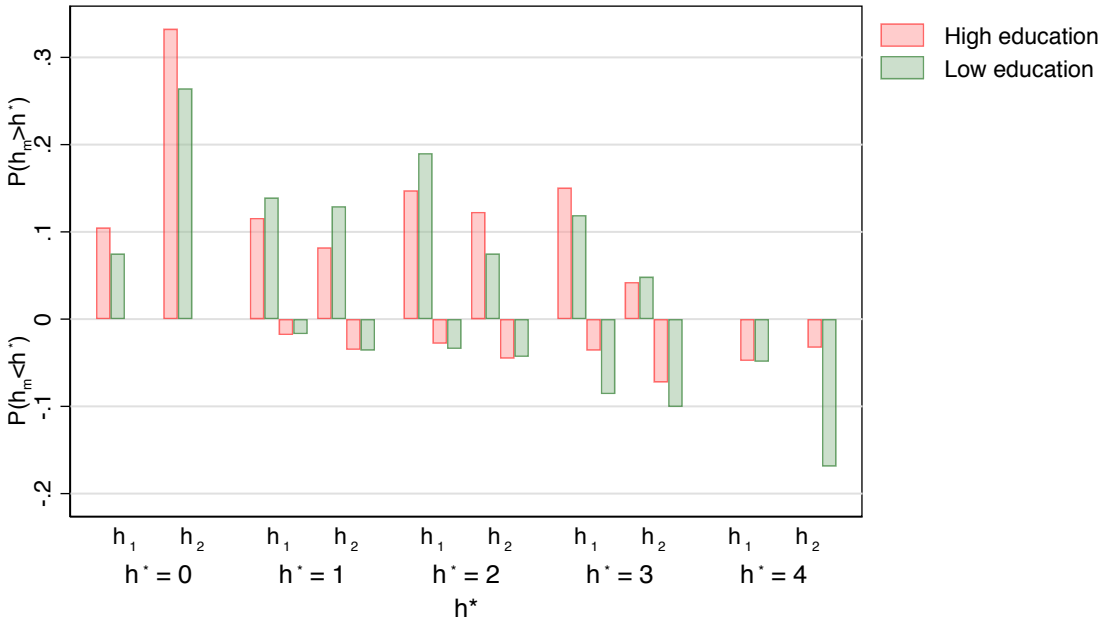
Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SAH questions in wave 1. In Panel (a), old individuals are those over the age of 70 years and young individuals those 40 years of age or younger. In Panel (b), predicted probabilities are averaged over the whole sample and evaluated at age 75 years and 30 years.

Figure A3: Misclassification in SAH for individuals with high and low education

(a) Reporting for high (H) and low (L) education individuals

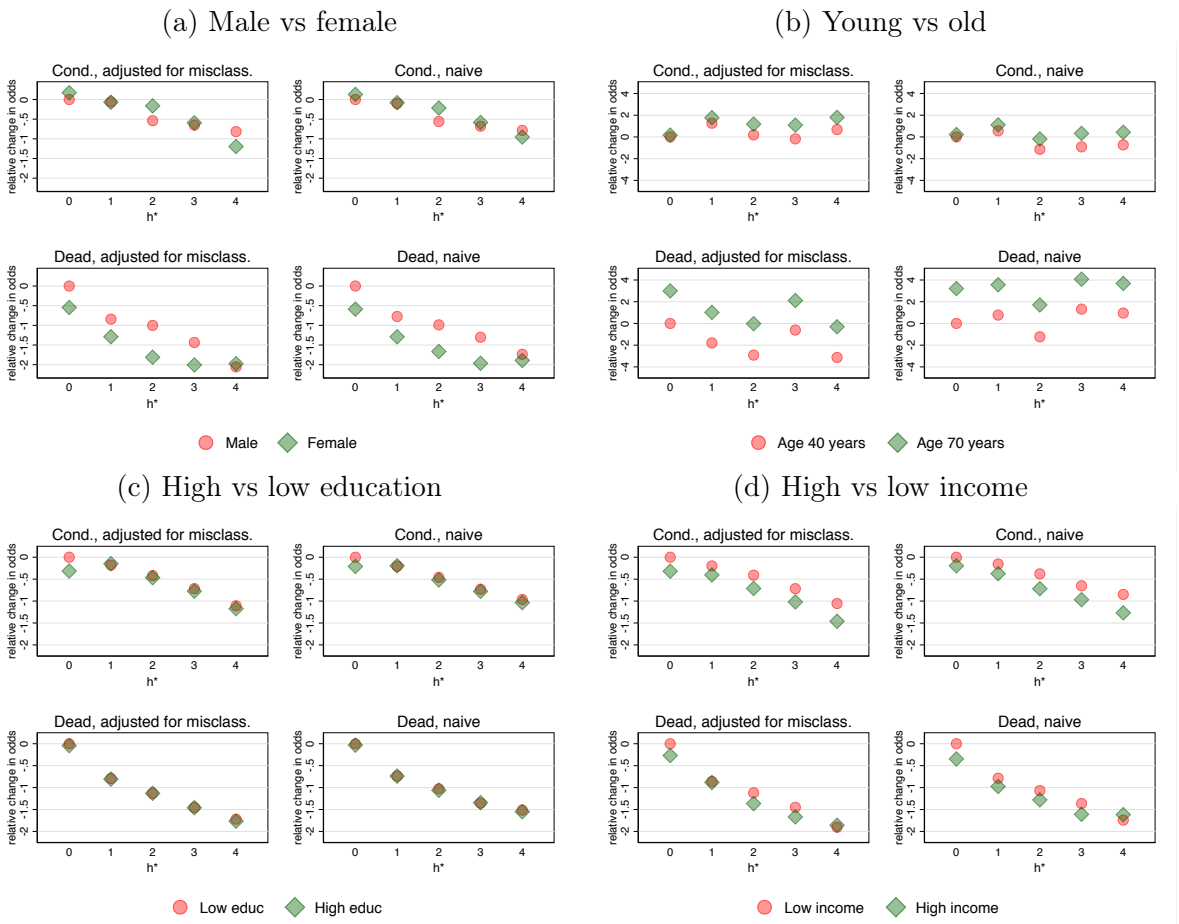


(b) Average predicted upward and downward misreporting



Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SAH questions in wave 1. In Panel (a), high education individuals are those whose highest education degree is a bachelor (education=16) and low education individuals those whose highest degree is Year 12 (education=12). In Panel (b), predicted probabilities are averaged over the whole sample and evaluated at education=16 (High education) and education=12 (Low education).

Figure A4: Heterogeneity in the effect of health on morbidity (cond) and mortality (dead): Relative change in odds



Notes: Data from HILDA waves 1 and 16 for individuals who responded to SAH questions in wave 1.