
Article:
Brown, S., Harris, M.N., Srivastava, P. et al. (1 more author) (2022) Mental health, reporting bias and economic transitions. Oxford Economic Papers, 74 (2). pp. 541-564. ISSN 0030-7653

https://doi.org/10.1093/oep/gpab005

Mental Health, Reporting Bias and Economic Transitions

Sarah Brown,[1] Mark N. Harris,[2] Preety Srivastava[3] and Karl Taylor[4#]

[1] Department of Economics, University of Sheffield, 9 Mappin Street, Sheffield, S1 4DT, and IZA Bonn; email: sarah.brown@sheffield.ac.uk

[2] School of Economics and Finance, Curtin Business School, Curtin University, Kent Street, Bentley, Perth, Australia; email: mark.harris@curtin.edu.au

[3] School of Economics, Finance and Marketing,

Royal Melbourne Institute of Technology, 124 La Trobe Street, Melbourne, Australia; email: pratima.srivastava@rmit.edu.au

[4] Department of Economics, University of Sheffield, 9 Mappin Street, Sheffield, S1 4DT, and IZA Bonn; email: k.b.taylor@sheffield.ac.uk

[#] Corresponding author.

Abstract

Measures of mental health are heavily relied upon to identify at-risk individuals. However, self-reported mental health metrics might be unduly affected by mis-reporting (perhaps stemming from stigma effects). In this paper, we consider this phenomenon by focusing upon the mis-reporting of mental health using UK panel data from 1991-2018. In separate analyses of males and females, we examine how inaccurate reporting of the $GHQ-12$ measure, specifically its sub-components, can adversely affect the distribution of the index. The analysis suggests that individuals typically over report their mental health (especially so for males). The results are then used to adjust the $GHQ-12$ score to take mis-reporting into account. We then compare the effects of the adjusted/unadjusted $GHQ-12$ index when modelling a number of important economic transitions. Using the original index typically leads to an underestimate of the effect of poor mental health on transitions into improved economic states, for example, unemployment to employment.

JEL Classification: C3, D1, I1

1 Introduction

Five of the fifteen leading causes of disability worldwide are psychiatric conditions (Mathers *et al.*, 2008). Mental disorders have become a global public health concern with the World Health Organization (WHO) predicting that one out of four people will endure some kind of mental illness during their life (WHO, 2001), and that the global economic burden of such mental disorders will be of the order of US$16 trillion between 2011 and 2030 (Bloom *et al.*, 2011). Mental illness thus represents an immense psychological, social and economic burden to society and additionally, increases the risk of physical illnesses such as heart disease and diabetes (Stein *et al.*, 2006).

The objective of the current paper is to develop a latent-class type modelling approach to analyse the extent of mis-reporting in mental health instruments that are self-reported. In this study, we focus on the $GHQ-12$ (and its sub-components). Using the GHQ, the relationship between psychological wellbeing and various economic outcomes of interest has been explored widely in the literature, ranging from education (Cornaglia *et al.*, 2015); employment (Boyce and Oswald, 2012; Thomas *et al.*, 2005); financial behaviour (Brown *et al.*, 2005); housing (Ratcliffe, 2015); stock prices (Ratcliffe and Taylor, 2015); transport (Roberts *et al.*, 2011); mortality (Gardner and Oswald, 2004); crime (Dustmann and Fasani, 2015); and income inequality (Wildman, 2003).

The $GHQ-12$ contains 12 items including depression, anxiety, somatic symptoms, feelings of incompetence, difficulty in coping and sleep disturbance, which are either self- or interviewer-administered, with each item measured using a 4-point Likert-type scale (Likert, 1932).[1] The accuracy of the information is dependent on respondents providing reliable and accurate responses. It is very likely the case, however, that because of the social stigma associated with adverse mental health (for example, Hinshaw, 2009), respondents have a perceived incentive to mis-report the true status of their mental health. For example, Bharadwaj *et al.* (2017) find that survey respondents are significantly more likely to under-report mental illnesses (compared to other health conditions) because of the fear of being stigmatized, socially sanctioned or disgraced. The dimensions of psychiatric morbidity that the $GHQ-12$ measures have been subject to a lot of debate in the literature, some indicating that it represents a unidimensional index of severity of psycho-

[1] Likert scale is the most widely used psychometric scale to measure respondents' agreement with various statements.

logical morbidity, and others arguing that it is multidimensional (Shevlin and Adamson, 2005; Hankins, 2008). Hankins (2008) attributes such inconsistency to a response bias on the negatively phrased items in the questionnaire.

Mis-reporting leads to information being mis-classified in survey data, which can mask the incidence of such behaviours and lead to biased and inconsistent estimates in statistical analyses (Hausman *et al.*, 1998). Although very little work has been undertaken in analysing the possible extent and consequences of inaccurate reporting in empirical models, its presence has been well-established in the psychology and related literatures. For example, social desirability has been found to be significantly associated with the over-reporting of physical activity and height, and the under-reporting of weight among women (Adams *et al.*, 2005; Ezzati *et al.*, 2006; Hebert *et al.*, 2002).

Although survey mis-classification, or mis-reporting, is known to be pervasive and to potentially bias statistical analyses, there is a limited body of research that has explicitly modelled such behaviours. A key study on mis-classified dependent variables is by Hausman *et al.* (1998). They consider a binary choice model with two types of mis-classification: the probability that the true 0 is recorded as a 1; and the probability that the true 1 is recorded as a 0, implying that the mis-classification errors are conditionally independent of covariates. A number of other studies have followed, extending this research in terms of semi-parametric estimation (Abrevaya and Hausman, 1999; Lewbel, 2000), the use of ordered data (Dustmann and Van Soest, 2001) and modelling mis-classification as a function of both observables and unobservables (Meyer and Mittag, 2017).

More recently, researchers such as Mahajan (2006), Hu (2008) and Molinari (2008) have attempted to model mis-classification in discrete dependent variables using a secondary measurement or an instrument to identify a nonlinear model. Their approach is based on the assumption that in the presence of classification errors, the relationship between the true variable and its mis-classified representation is given by a linear system of simultaneous equations in which the coefficient matrix is the matrix of mis-classification probabilities. Mis-classification resulting from anchoring, focal point answers and crude rounding in surveys has also increasingly been a subject of interest to researchers (for example, Van Soest and Hurd, 2008; Manski and Molinari, 2010; Kleinjans and Van Soest, 2014). Lastly, anchoring vignettes have also been used to measure discrepancies in reporting behaviours, particularly in the case of self-reported health and life satisfaction (Kristensen and Johansson, 2008; Van Soest *et al.*, 2011); however, vignettes are very

rare in most large scale data sets.

Our methodology ties in with the literature on latent class type models. Our basic starting hypothesis is that there are inherently two types of individuals in the population with regard to how they respond to particular survey questions of interest: "accurate" and "inaccurate". However, we will never directly observe to which type, or (latent) class, a respondent belongs. Thus, the broad approach we follow is that of latent class or finite mixture models (for a comprehensive review, see McLachlan and Peel (2004)), where our hypothesized classes correspond to these two types of individuals. In latent class modelling, the researcher aims to split the population according to high/low (healthcare, say) users, for example, even with observationally equivalent usage levels. Therefore, a novelty of our approach is to adopt these widely used and accepted techniques to help us identify and quantify any potential inaccurate reporting.

Explicitly, we offer researchers some generic tools with which to account for, and quantify, the effect of any mis-reporting behaviour in large scale surveys. We show how these can be applied to the important area of mental health, and in particular, the commonly used $GHQ-12$ instrument. We then show how these results can be used to identify potential questions of interest that may be particularly subject to mis-reporting, and to also adjust the index so as to obtain a more realistic distribution of the population's mental health over time. Our findings can also assist clinicians and researchers to assess the reliability of the $GHQ-12$ and the validity of the dimensions it measures. Finally, we illustrate how the use of these corrected indices can affect inference regarding the effects of mental health on several important individual economic outcomes, such that one could draw erroneous policy implications by ignoring these mis-reporting behaviours.

2 The 12-item General Health Questionnaire

The General Health Questionnaire (GHQ) is a self-administered psychometric screening tool that was developed with the aim to detect and assess individuals with a diagnosable psychiatric disorder (Goldberg and Hillier, 1979; Goldberg and Williams, 1988; McDowell, 2006). Based on the original 60-item questionnaire, several versions have been constructed. The $GHQ-12$ is a quick and reliable short form version which is often used in research studies (https://www.gl-assessment.co.uk/products/general-health-questionnaire-ghq, last accessed 14/01/2021) and is commonly interpreted as a general

measure of psychological well-being. It has twelve items stemming from the following questions which are asked annually in both the *BHPS* and *UKHLS* surveys: *'Here are some questions regarding the way you have been feeling over the past few weeks. Have you recently...'; (1) 'been able to concentrate on whatever you're doing?'; (2) 'lost much sleep over worry?'; (3) 'felt that you were playing a useful part in things'; (4) 'felt capable of making decisions about things?'; (5) 'felt constantly under strain?'; (6) 'felt you couldn't overcome difficulties?'; (7) 'been able to enjoy your normal day-to-day activities'; (8) 'been able to face up to problems?'; (9) 'been feeling unhappy or depressed'; (10) 'been losing confidence in yourself?'; (11) 'been thinking of yourself as a worthless person?'; and (12) 'been feeling reasonably happy, all things considered'.* The responses to each of the twelve questions lie on a four-point Likert scale ranging from 0 to 3. The Likert scale of each sub-component is scored so that higher values indicate decreased levels of mental health. The $GHQ-12$ score converts valid answers to the 12 items to a dichotomous scale by recoding scores of 0 and 1 ('better than usual' and 'same as usual' on positively worded items; 'not at all' and 'no more than usual' on negatively worded items) on individual sub-items to 0, and scores of 2 and 3 ('less than usual' and 'much less than usual' on positively worded items; 'rather more than usual' and 'much more than usual' on negatively worded items) to 1, which is then summed, giving a scale running from 0 (the least distressed) to 12 (the most distressed).

3 Econometric framework

The main purpose of this study is to determine if there is any bias in the composite $GHQ-12$ measure, which has been used in numerous important economics studies; for example, Clark (2003), Roberts *et al.* (2011) and Cornaglia *et al.* (2015). Since this is a simple construct from the 12 underlying items or components (by summing the 12 individual 0/1 scores as described above), obvious (related) questions are: *are any of these 12 questions in particular, subject to mis-reporting bias?* And *what is the extent of any mis-reporting bias across these 12 questions?* Hence, any bias in the overall index, must arise from mis-reporting or bias in some, or all, of the composite $GHQ-12$ components. Explicitly, the hypothesis is that, due to stigma and related effects (see, for example, Bharadwaj *et al.* (2017)), a proportion of individuals will erroneously over report zero scores in the components (corresponding to an original value of 0 or 1, in the original

Likert index; see above). Then, due to the summary composition of the overall index, this hypothesized behaviour in the components will lead to item inflation in the composite score, and most likely at 0 in the 0-12 score.

Accordingly, the econometric framework developed here consists of modelling the individual components that, in sum, describe the composite measure. The aim here is to model any potential mis-reporting in these individual components. In doing this, it will be possible to identify particular questions that are more likely to be adversely affected by mis-reporting behaviours. It will also allow us, post-estimation, to construct a new composite $GHQ-12$ index by systematically correcting the 12 individual components if we find the probability of mis-reporting to be high.

A casual inspection of the distribution of the composite (original) $GHQ-12$ measure (see Figures 1 and 2 for males and females, respectively) clearly illustrates, as expected, a marked spike at zero; and indeed at a magnitude (apparently) completely at odds with the remainder of the distribution. Indeed, zero values in this composite instrument are important as "a score of zero on the $GHQ-12$ questionnaire can, in contrast (*to a score of more than 4*), be considered to be an indicator of psychological wellbeing" (Scottish-Government, 2013). It is our contention that such a large relative representation of a lack of mental health issues, may be an over-representation of the true state of affairs. As noted, the hypothesis is that there is a subset of the population who erroneously identify themselves into this (favourable) category by reporting a 0 or 1 score (on the Likert scale) in all 12 individual components. The reasons for this will presumably be wide and varied across this sub-population, but may result from a desire to appear aligned with social norms and to avoid any associated stigma effects of being identified as having either an actual, potential, or perceived, psychological disorder.

For these reasons, we require an econometric model that allows for an "inflation" of the zero outcome in each sub-component. That is, we wish to (probabilistically) distinguish "true" zero responses from the "false" ones; or equivalently, to allow for two different types of zero observations, following the latent class literature. We wish to model binary outcome variables for each of the separate 12 questions, where 1 relates to a score of 2 or 3 on the Likert scale and a value of 0 relates to a score of 0 or 1 on the Likert scale, whereby we believe that, in some cases at least, an excess of zeros is recorded.

In such a set-up, there are two equations driving the eventual observed outcome. Firstly, a latent variable, $\widetilde{y}_q^*$, is specified that represents the true mental health status

related to the q^{th} question for each of the $q = 1, \ldots, 12$ questions. $\widetilde{y}_q^*$ is a function of variables $\mathbf{z}$ with unknown weights $\boldsymbol{\gamma}_q$, and a standard-normally distributed error term (as is commonly assumed in the literature), u_q, such that

$$\widetilde{y}_q^* = \mathbf{z}'\boldsymbol{\gamma}_q + u_q. \tag{1}$$

This translates into a discrete variable $\widetilde{y}_q$, where $\widetilde{y}_q = 1$ for $\widetilde{y}_q^* > 0$ and $\widetilde{y}_q = 0$ for $\widetilde{y}_q^* \leq 0$. Secondly, there is a equation which relates to the individual's propensity to report accurately, represented by r_q^* (where $q = 1, 2, \ldots, 12$). Again, this is specified as a function of variables $\mathbf{x}$ with unknown weights $\boldsymbol{\beta}_q$, where there may be some overlap between $\mathbf{x}$ and $\mathbf{z}$, and an error term ε_q, such that

$$r_q^* = \mathbf{x}'\boldsymbol{\beta}_q + \varepsilon_q. \tag{2}$$

The observability criterion for observed y_q is now (where $r_q = 1 \times \left[r_q^* > 0\right]$)

$$y_q = \widetilde{y}_q \times r_q. \tag{3}$$

Allowing for the likely correlation between ε_q and u_q $\left(\rho_q\right)$, the full probabilities are given by

$$\Pr(y_q) = \begin{cases} \Pr\left(y_q = 0 \left|\mathbf{x}\right.\right) = \left[1 - \Phi\left(\mathbf{z}'\boldsymbol{\gamma}_q\right)\right] + \Phi_2\left(\mathbf{z}'\boldsymbol{\gamma}_q, -\mathbf{x}'\boldsymbol{\beta}_q; -\rho_q\right) \\ \Pr\left(y_q = 1 \left|\mathbf{x}\right.\right) = \Phi_2\left(\mathbf{x}'\boldsymbol{\beta}_q, \mathbf{z}'\boldsymbol{\gamma}_q; \rho_q\right). \end{cases} \tag{4}$$

So here, the probability of a zero observation has been "inflated" as it is a combination of the probability of a "true" 0 score from the mental health equation - $\left[1 - \Phi\left(\mathbf{z}'\boldsymbol{\gamma}_q\right)\right]$ - plus the probability of an "inaccurately" reported one from the splitting probit model; $\Phi_2\left(\mathbf{z}'\boldsymbol{\gamma}_q, -\mathbf{x}'\boldsymbol{\beta}_q; -\rho_q\right)$. We refer to this as an *inflated probit* model (see Brown *et al.* (2018)). Once the assumed form of the probabilities is known and observations on $y_{i,q}$ are available in an $i.i.d.$ sample of size N from the population, the parameters of the full model $\boldsymbol{\theta}_q = \left(\boldsymbol{\beta}_q', \boldsymbol{\gamma}_q', \rho_q\right)'$ can be consistently and efficiently estimated using maximum likelihood (ML) techniques. The likelihood function for a *single component* (q) is therefore

$$L_i\left(\boldsymbol{\theta}\right) = \prod_{j=0}^{j=1} \Pr\left(y_i = j \left|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}\right.\right), \; j = 0, 1 \tag{5}$$

$$= P_i \tag{6}$$

As argued in Brown *et al.* (2018), it is generally preferable to have exclusion restrictions across both $\mathbf{x}$ and $\mathbf{z}$, which we return to below.

In the analysis that follows, we analyse panel data: that is, for each individual i, we have repeated observations over time periods $t = 1, \ldots, T_i$. Formulating the above model in this context allows one to account for the very likely unobserved individual heterogeneity in both underlying equations, $\boldsymbol{\alpha}$ (in each of the q components). As is standard in the literature, it is assumed that $\boldsymbol{\alpha} \sim N(0, \Sigma)$; and we denote the individual elements of Σ by $\widetilde{y}_q^*$ and r_q^*, respectively. Since the presence of such unobserved effects complicates evaluation of the resulting likelihood function, we utilize the method of maximum simulated likelihood. Dropping the q subscript for ease of notation, we can define $\mathbf{v}_i$ as a vector of standard normal random variates, which enter the model generically as $\boldsymbol{\Gamma}\mathbf{v}_i$, such that for a single draw of $\mathbf{v}_i$, $\boldsymbol{\Gamma}\mathbf{v}_i = (\alpha_{i,\widetilde{y}^*}, \alpha_{i,r^*})$. $\boldsymbol{\Gamma}$ is the $chol(\Sigma)$ such that $\Sigma = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$. Conditioned on $\mathbf{v}_i$, the sequence of T_i outcomes for individual i are independent, such that the contribution to the likelihood function for a group of t observations is defined as the product of the sequence P_{it} - see equation (6) - which we denote e_i, corresponding to the observed outcome of y_i, $e_i \mid \mathbf{v}_i$,

$$e_i \mid \mathbf{v}_i = \prod_{t=1}^{T_i} (P_{it} \mid \mathbf{v}_i)^{d_{it}} \tag{7}$$

where d_{it} is the indicator function, $1 \times [y_{it} = j]$. The unconditional log-likelihood function is found by integrating out the $\mathbf{v}_i$ as

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int_{v_i} \prod_{t=1}^{T_i} (P_{it} \mid \boldsymbol{\Gamma}\mathbf{v}_i) \, f(\mathbf{v}_i) d\mathbf{v}_i, \tag{8}$$

where all parameters of the model are contained in $\boldsymbol{\theta}$. Using the usual assumption of multivariate normality for $\mathbf{v}_i$ yields

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int_{v_i} \prod_{t=1}^{T_i} (P_{it} \mid \boldsymbol{\Gamma}\mathbf{v}_i) \prod_{k=1}^{K} \phi(\mathbf{v}_{ik}) d\mathbf{v}_{ik}, \tag{9}$$

where k indexes the different unobserved effects in the model (so here, $K = 2$ per q). The expected values in the integrals can be evaluated by simulation by drawing R observations on $\mathbf{v}_i$ from the multivariate standard normal population. The following is the resulting simulated log-likelihood function

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} (P_{it} \mid \boldsymbol{\Gamma}\mathbf{v}_{ir}) \,. \tag{10}$$

Halton sequences of length $R = 1000$ were used (see Train, 2009), and this now feasible function is maximized with respect to $\boldsymbol{\theta}$.

As is common in the non-linear panel data literature, given that these unobserved heterogeneity terms are (potentially) correlated with observed heterogeneity terms, the correction proposed by Mundlak (1978) is applied. Consequently, we include averages of the continuous covariates of individual i in the set of explanatory variables, $\overline{x_i} = \frac{1}{T_i}\sum_{t=1}^{T_i} x_{it}$.

4 Data

We use the British Household Panel Survey (*BHPS*), a survey conducted by the Institute for Social and Economic Research, which is a large scale representative longitudinal study collecting data on individuals over the period 1991 to 2008.[2] It is household-based and interviews every adult member of sampled households. In 1991, the sample comprized around 5,500 households and over 10,000 individuals living in 250 areas of Great Britain. We also employ the successor to the *BHPS*, Understanding Society - the UK Household Longitudinal Study (*UKHLS*) - which is a nationally representative longitudinal study of the UK population which started in 2009, see University of Essex (2019).[3] In the first wave of the *UKHLS*, over 50,000 individuals were interviewed over the period 2009 to 2011 and, correspondingly, in the latest wave available (at the time of writing), wave 9, around 36,000 individuals were interviewed between 2017 and 2019 (hereafter referred as 2018). Both the *BHPS* and *UKHLS* contain detailed information on economic and socio-demographic characteristics. It is possible to track individuals from the *BHPS* into the *UKHLS* hence making a relatively long panel dataset.

We focus upon two unbalanced panels over the period 1991 to 2018 split by gender, where the total number of observations for males is 122,247 comprising 14,531 individuals aged 18 or above, and the respective figures for females are 148,056 observations comprising 16,382 individuals. Males are observed, on average, 14 times over a quarter of a century whilst the corresponding figure for females is 15 times. The percentage of individuals, by gender, present in all periods is 6.1% (7,384 males) and 6.8% (10,114 females).

[2] https://www.iser.essex.ac.uk/bhps (last accessed 14/01/2021).
[3] https://www.understandingsociety.ac.uk (last accessed 14/01/2021).

In part of the interview, respondents are asked to complete the self-completion $GHQ-12$ questionnaire. This measure of mental health is available in both the *BHPS* and the *UKHLS* and has been used to examine a range of policy-relevant areas such as education and employment (as discussed above). Throughout the *BHPS* and in the first two waves of the *UKHLS,* the self-completion component of the questionnaire, which includes the $GHQ-12$, was a paper instrument handed to the participant to fill in. From wave 3 onwards in the *UKHLS*, the interviewer handed a laptop to the participant to complete the self-completion component of the questionnaire for themselves. From wave 7 onwards, the full interview including the self-report component of the questionnaire could be completed online with no interviewer involvement. In sensitivity analysis discussed below, we explore the effect of the mode of interview.

Figures 1 and 2 show the distribution of the $GHQ-12$ for males and females, respectively, and Table 1 provides summary statistics for the $GHQ-12$ and its sub-components, by gender. From Figures 1 and 2, looking at the original $GHQ-12$ score, it is clear that there is around a 10 percentage point differential across gender in reporting a score of 0, with it being lower for females. It is also apparent from Figures 1 and 2 that around 60% of males and 50% of females report none of the above (component) problems, whilst Table 1 reveals that the average number of problems is 1.5 for males compared with 2 for females. Around 13% of males and 19% of females in the sample report in excess of four problems over the period 1991-2018. Considering the elements of the $GHQ-12$, the most common problem faced by individuals is feeling constantly under strain, *i.e.*, 23% for males and just under 30% for females, followed by around 17% of males and 24% of females feeling unhappy or depressed. Interestingly, Table 1 reveals that, across each of the $GHQ-12$ sub-components, problems are more prevalent for females.

The variables used to model the sub-components of the $GHQ-12$, given in the vector $\mathbf{z}$, essentially follow the received literature (for example, Metcalfe *et al.*, 2011). In terms of the explanatory variables in both $\mathbf{x}$ and $\mathbf{z}$, we control for: the age of the individual (entered as a quadratic); married or cohabiting (other states constitute the reference group); white; highest educational attainment, specifically a degree, teaching or nursing qualification, A levels, GCSE (or O level), other qualifications (no education is the omitted group); the natural logarithm of labour income last month; the natural logarithm of non-labour income last month; employment status (employed, self-employed or unemployed; other states make up the reference group); housing tenure, specifically

whether the home is owned outright, via a mortgage, or rented (other tenure states form the reference category); the number of dependent children in the household; the number of adults in the household (excluding the respondent); region of residence; and year of interview.

In addition, we control for the general health of the individual in $\mathbf{z}$. The *BHPS* and *UKHLS* both contain a question on self-assessed health (SAH): '*Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?*' However, due to reporting bias and measurement error, the reported SAH may be endogenous in the subsequent analyses. To accommodate this possibility, we follow the standard approach in the literature, see, for example, Stern (1989) and Bound (1991), and condition SAH on a set of instruments, namely whether the individual reports specific health problems.[4] The logic here is that more objective measures are used to instrument the endogenous and potentially error ridden subjective health measure. Following the literature, we estimate the health stock of an individual by employing a Generalized Ordered Probit (GOP) model, which allows for the fact that people with the same underlying level of health may apply different thresholds when reporting SAH and hence different ordered categories for similar positions on the assumed underlying continuous scale (Rice *et al.*, 2010; Lindeboom and Van Doorslaer, 2004; Kerkhofs and Lindeboom, 1995). We then take the linear prediction from the GOP model as a measure of an individual's health stock, where higher values denote worse health.

In the vector $\mathbf{x}$, we include a number of additional covariates to identify mis-reporting. Firstly, we control for the percentage of compulsory questions (*i.e.*, those asked to everyone completing the survey) not answered in the individual questionnaire. The idea here is that those individuals who complete a smaller proportion of questions, perhaps because they have a lower level of trust in the survey, will *a priori* be more likely to answer less accurately. This is consistent with the approach of Brown *et al.* (2018) and is based on existing literature which suggests that the longer a respondent spends time with the interviewer the more trusting they are of both them and the survey in general; see, for example, Corbin and Morse (2003). Secondly, we condition on whether there is a change

[4]Individuals are asked whether they have any of the following health problems: arms, legs or hands; sight; hearing; skin conditions or allergies; chest or breathing; heart or blood pressure; stomach or digestion; diabetes; anxiety or depression; alcohol or drugs; epilepsy or migraine; any other problem.

in interviewer over time (*i.e.*, between waves) following Niccoletti and Peracchi (2005) and Jenkins *et al.* (2008). The logic behind the use of this control is similar to the above, in that interviewer continuation is associated with respondent trust, interviewer reputation and rapport with the respondent, and hence continued survey participation over time (for example, Schrapler (2004) and Vassallo *et al.* (2015)).[5] We also control for whether the respondent has an optimistic personality and for interview conditions. Specifically, for the latter, we control for whether the individual was highly cooperative during the interview and whether other individuals were present when the questionnaire was completed, which might capture the effects of social stigma.[6] The literature to date has found that having a third person present during interviews typically results in biased responses (for example, Hartmann (1994) and Zipp and Toth (2002)).

5 Results

We estimate the random-effects inflated probit models for each sub-component of the $GHQ-12$, separately for males and females. Whilst we have argued that the variables used to model the sub-components of the $GHQ-12$, given in the vector $\mathbf{z}$, follow the existing literature, in order to explore the robustness of our findings, three alternative specifications are estimated which vary in the identifying variables used to model misreporting (*i.e.* those covariates in the vector $\mathbf{x}$). Specifically, specification 1 conditions on the percentage of compulsory questions not answered in the survey and whether there has been a change in interviewer between waves. Specification 2 in addition includes a control for whether the respondent is optimistic and specification 3 additionally incorporates controls for interview conditions. To select between the alternative specifications, we refer to the Akaike and Bayesian Information Criteria (AIC and BIC, respectively), where the minimum AIC and BIC are highlighted in bold.[7] The results are shown in Table A.1 in the Appendix for males and females. Clearly, the favoured model is specification 3

[5] Note that interviewers in the *BHPS* and *UKHLS* are randomly allocated to respondents the first time that a household appears in the survey and are, hence, independent of respondent characteristics.

[6] A general lack of cooperation in the survey is an individual decision related to the perceived cost of completing the interview, which is also related to a person's past survey experience, see, for example, Niccoletti and Peracchi (2005), whilst interviewer experience and skill are likely to influence the respondent's cooperation in face-to-face surveys, see, for example, Jackle *et al.* (2013).

[7] These results are also confirmed by likelihood ratio tests. Note the bias is relatively stable across the specifications (within a range of ± 5 percentage points) implying that including alternative controls does not significantly affect the results.

across-the-board.[8]

The full set of coefficients for our preferred specification is shown in Tables A.3 and A.4 for males (Tables A.5 and A.6 for females) in the Appendix. Due to the number of results, we focus here on the variables used to identify inaccurate reporting behaviours.[9] In general, across each sub-component of the $GHQ-12$ and gender, a number of the variables employed to identify mis-reporting are individually statistically significant. Moreover, in the majority of instances, the percentage of (compulsory) questions left unanswered in the questionnaire is positively associated with the respondent's propensity to report inaccurately, which is also generally true of changes in the interviewer over time (where statistically significant), which is consistent with our *a priori* expectations.[10] Conversely, interview conditions, in particular being highly cooperative during the interview, are inversely associated with the likelihood of inaccurate reporting. In general, the correlation between the mental health and mis-reporting equations, ρ_q, is statistically significant for each sub-component ($q = 1, \ldots, 12$), justifying the inclusion of this additional parameter in estimation. So, in summary, the variables in our reporting behaviour equations are generally statistically significant, especially in regard to our identifying variables, suggesting that these are performing well, which lends further support to the modelling approach.

Of particular importance to the current study, Tables 2 and 3 present summary probabilities for males and females, respectively. These provide insights into the extent of mis-reporting (or reporting bias). Column 1 presents the sample proportion reported for each $GHQ-12$ sub-component as indicated by survey responses. Using the estimated models, the predicted rates for each sub-component are presented in Column 2 and the resulting estimated "reporting bias" in Column 3. To be specific, the elements in Column 2 are obtained by evaluating the expression $\Phi\left(\mathbf{z}_{it}'\widehat{\boldsymbol{\gamma}}_q\right)$ in the first line of equation (4),

[8] For a sub-sample of individuals present from wave 7 onwards in the *UKHLS,* a group of respondents were invited to take part online as well as others doing the interview face-to-face and by phone. Hence, for this sub-sample, we compare specification 3 to specification 4, which in addition includes whether the interview was completed online. The AIC and BIC across the two specifications are shown in Table A.2, where again specification 3 is preferred. Interestingly, the mode of interview was found to be statistically insignificant in most sub-components of the $GHQ-12$.

[9] Moreover, the results from modelling the sub-components of the $GHQ-12$ are generally in accordance with those found in the literature.

[10] Whilst the choice of identifying variable is always open to discussion, the results which follow are generally robust to alternative instruments. Following the broad suggestions of Angrist and Pischke (2009), we explore whether the identifying variables are directly associated with the mental health outcome. These tests support the use of these controls and lends support to our identification strategy. Furthermore, the use of these identifying variables has precendence in the existing literature, as discussed above.

which corresponds to the "true" probability of experiencing that problem, in the absence of any reporting effects, and averaged over individuals. It should be noted that the standard errors of these probability estimates (obtained via the Delta Method) are all very small, giving us confidence in their estimated magnitudes. Comparing the numbers in Columns 1 to 2 provides the reporting bias numbers in Column 3, given as a percentage. For example, the reported rate of a score of 1 for $GHQ1$ (*concentration*) is 0.154 versus the model predicted rate of 0.329; this results in what we are calling a 'reporting bias' of 114% [*i.e.*, $(0.329 - 0.154)/0.154$].

Importantly, these results generally indicate significant under-reporting in most of the 12 sub-components of the $GHQ - 12$, with the most significant bias of 189% estimated for $GHQ6$ (*overcoming difficulties*) for males and 181% estimated for $GHQ3$ (*usefulness*) for females. The predicted rates more than doubled for several other sub-components amongst males, such as $GHQ1$ (*concentration*), $GHQ7$ *(enjoying activities*) and $GHQ10$ (*confidence*), and only $GHQ7$ *(enjoying activities)* in females. In general, reporting biases are lower among females.[11]

In Column 4, the marginal probabilities of mis-reporting are presented for the 12 elements, which generally reflect the results in Column 3, with the highest probability of mis-reporting for $GHQ3$ (*usefulness*), $GHQ6$ (*overcoming difficulties*), $GHQ7$ *(enjoying activities*) and $GHQ11$ (*worthless* - with a positive bias) for males, and $GHQ3$ (*usefulness*) and $GHQ7$ *(enjoying activities)* for females. Finally, we present two sets of posterior probabilities in Columns 5 and 6. As noted above, zero observations come from two sources: mis-reporters; and accurate reporters with a true 0 score. Using posterior probabilities that are conditional on knowing what outcome the individual chooses (we re-visit this below), we can also make a prediction on what percentage of the zeros come from mis-reporters and accurate reporters with a true 0 score, respectively, using all the information we have on the individual. All the posterior probabilities again appear to be accurately estimated with respect to their very small standard errors, with the sub-elements $GHQ5$ *(strain), GHQ6* (*overcoming difficulties*) and $GHQ7$ (*enjoyment*) being subject to the greatest amount of mis-reporting in males, and the sub-elements $GHQ3$

[11] We have also explored whether the extent of the bias varies year-by-year by estimating yearly cross sectional models based upon specification 3. We found that across-the-board, bias did not appear to be constant over time, implying that it would not be appropriate to control for it by inclusion of simple individual fixed effects. It does though, suggest that mis-reporting should be explicitly modelled through covariates and, hence, endorses our modelling approach. Moreover, these yearly models were found to be statistically inferior to our preferred panel variants.

(usefulness), $GHQ5$ (*strain*) and $GHQ7$ *(enjoyment)*, in females.[12]

6 Adjusting the $GHQ-12$ index

As a natural extension of the above analyses, in this section we show how the results can be used to adjust the $GHQ-12$ index in light of the estimated amount of mis-reporting. We do this on the basis of the estimated posterior probabilities. We favour these, as opposed to prior probabilities, because they use all the information available on an individual, and should therefore provide more accurate predictions.

On the basis of these posterior probabilities, as noted above, we can make a prediction on what percentage of the reported zeros are related to a true zero-outcome and to mis-reporting, respectively. These are similar to probabilities estimated in latent class models (Greene, 2012) and essentially attempt to answer the question: *given that an individual recorded a zero, what is the probability that they are a mis-reporter versus an accurate reporter with a genuine* 0*-score (given their observed characteristics)?* The posterior probabilities for the two types of zeros for each sub-component q ($q=1,\ldots,12$) are given as

$$\begin{aligned} Pr\left(\widetilde{y}_q=0|\mathbf{x},y_q=0\right) &= \frac{f(\widetilde{y}_q=0|\mathbf{x})}{f(y_q=0|\mathbf{x})} \\ &= \frac{1-\Phi\left(\mathbf{z}'\boldsymbol{\gamma}_q\right)}{\left[1-\Phi\left(\mathbf{z}'\boldsymbol{\gamma}_q\right)\right]+\Phi_2\left(\mathbf{z}'\boldsymbol{\gamma}_q,-\mathbf{x}'\boldsymbol{\beta}_q;-\rho\right)} \end{aligned} \tag{11}$$

and

$$\begin{aligned} Pr\left(\widetilde{y}_q=1,r_q=0|\mathbf{x},y_q=0\right) &= \frac{f(\widetilde{y}_q=1,r_q=0|\mathbf{x})}{f(y_q=0|\mathbf{x})} \\ &= \frac{\Phi_2\left(\mathbf{z}'\boldsymbol{\gamma}_q,-\mathbf{x}'\boldsymbol{\beta}_q;-\rho\right)}{\left[1-\Phi\left(\mathbf{z}'\boldsymbol{\gamma}_q\right)\right]+\Phi_2\left(\mathbf{z}'\boldsymbol{\gamma}_q,-\mathbf{x}'\boldsymbol{\beta}_q;-\rho\right)}, \end{aligned} \tag{12}$$

which necessarily sum to unity.

[12] Our findings relate to the existing literature, which has found that the $GHQ-12$ sub-components measure both positive and negative mental health dimensions. In particular, Hu *et al.* (2007) explore whether interdependence exists between these two domains. Indeed, our results suggest that mis-reporting bias is generally smaller (larger) in the case of positively worded questions (namely $GHQs$ 1, 3, 4, 7, 8 and 12) for males (females). Considering the third column of Tables 2 and 3, the average reporting bias for males (females) across positively worded questions is 82% (75%) compared to 94% (38%) for negatively worded components. Hence, the contrast in the bias between positively and negatively worded sub-components is unambiguous in the case of females. This implies that the phrasing of questions is potentially an important factor in determining the extent of the reporting bias.

We estimate the posterior probability of mis-reporting (at an individual level) for each of the twelve components of the $GHQ-12$ (that is, evaluating equation 12). Next, we assign the estimated probabilities to individuals who reported a zero to the respective questions and were estimated to have a high posterior probability of mis-reporting. Following the convention with predicted success and failure in empirical work, we use the 0.5 cutoff rule. For example, if individual i's posterior probability of mis-reporting for a sub-component (say, $GHQ5$) is 0.61 (which is > 0.5), we contend that there is a (high) 61% chance that the zero recorded by individual i is mis-reported as against a 39% chance that it is a genuine zero-outcome. Thus, we adjust the zero in $GHQ5$ to 0.61 for individual i. Instead, if we estimate a (low) posterior probability of mis-reporting of 0.29 (which is $\leqslant 0.5$) for individual i, we treat the reported zero as a genuine outcome that does not require any adjustment. After so-adjusting the observed zeros, we then sum all of the 12 sub-components to construct an adjusted $GHQ-12$ index. To make this adjusted measure comparable to the original index, we simply round the adjusted sum to the nearest integer.

The resulting indices for males and females are illustrated in Figures 1 and 2, respectively, in Panel A labelled "*adjusted*". While the adjusted $GHQ-12$ indices clearly mimic the overall shape of the original indices, we can see a significant reduction in the frequency of the zeros, which have been predominantly reallocated to the neighbouring outcomes of $1-5$. We next explore the robustness of our adjusted index with a slightly different approach. Using the same rule as before, here, where appropriate, we replace the zeros with a 1 instead of the respective posterior probabilities. We notice quite similar patterns in the respective adjusted $GHQ-12$ index, albeit a larger shift to outcome 6 for males, lending confidence to our approach (shown in Panel B labelled *"robust"* in Figures 1 and 2). As a final exercise, we use the observed sample proportions of the respective sub-components as the cutoff rule to adjust the index (as opposed the usual/previous 0.5). This could be regarded as an upper bound of the adjusted index and is shown in Panel C in Figures 1 and 2 (labelled *"upper bound"*), where for both males and females this measure clearly mimics the original $GHQ-12$, and so would appear to be the least effective approach out of the three alternatives discussed at correcting for mis-reporting.

As highlighted above, scores in excess of 4 on the $GHQ-12$ scale are taken to be possibly symptomatic of a mental health issue, in contrast to a score of 4 or below (Scottish-Government, 2013). For the original $GHQ-12$ composite measures, 12.7% of

the males sample and 18.8% of the females sample reported a score greater than 4. Hence, females appear to have higher levels of mental health issues, which is also evident after conditioning upon covariates. The comparable figures once the composite index has been adjusted using the posterior probabilities are 19.8% and 21.6%, respectively (closer to the 25% lifetime prediction by the WHO (WHO, 2001)). Thus, the resulting distribution of the composite metric has a larger tail reporting states in excess of 4, with a narrower gender difference than before.

7 Applications using the adjusted metrics

In this section, we consider applications of the adjusted $GHQ-12$ index to modelling transitions in some key economic outcomes, by focusing on how the mental health instrument is associated with changes in education, labour market status and savings between time $t-1$ and t. We examine increases in educational attainment ($t-1$ to t); transitions from being unemployed or out of the labour force ($t-1$) to paid employment or self-employment (t), for individuals of working age; and changes in the incidence of saving, *i.e.*, from being a non-saver to a saver, ($t-1$ to t).

The change in the state of each outcome (s_{it}) from $t-1$ to t (Δ) is modelled as a binary outcome, equal to unity if the state improves over time, *i.e.,* an increase in educational attainment, moving out of unemployment into employment, switching from a non-saver to saver. Each outcome is conditioned on a quadratic in age, marital status, total income, housing tenure, year of interview and region of residence, given in vector $\mathbf{z}_{it-1}$. We also control for whether the individual gave a $GHQ-12$ score different to zero at $t-1$. That is, for each economic outcome, we compare the effect of not reporting a zero for the composite $GHQ-12$ and the three alternative adjusted metrics detailed above. This is included as a binary variable, $g_{it-1}=1$, if $GHQ-12\neq 0$ in period $t-1$. Each dependent variable is estimated as a panel probit model, where μ_i is the individual specific random effect as follows:

$$\Delta s_{it}=\mathbf{1}\times\left[\mathbf{z}'_{it-1}\boldsymbol{\pi}+\lambda g_{it-1}+\mu_i+\varepsilon_{it-1}>0\right]. \tag{13}$$

The results are shown in Table 4 for males and Table 5 for females, where the first four columns focus on transitions in educational attainment, the next four consider labour market status and the final four columns focus upon transitions in financial behaviour, *i.e.*,

whether the individual becomes a saver.[13] Each table provides specifications employing: (A) the original $GHQ-12$; (B) the *adjusted* index (labelled as "Adj. 1"); (C) the *robust* method (labelled as "Adj. 2"); and (D) the *upper bound* measure (labelled as "Adj. 3"), as described above. For brevity, we only report the estimates of λ.

The results show that, in general, individuals who report a non-zero score derived from either the original $GHQ-12$ or one of the alternative adjusted measures, *i.e.* $g_{it-1}=1$, have a lower likelihood of increasing educational attainment (which is consistent with Cornaglia *et al.* (2015)), moving into employment as previously reported in the literature (for example, Boyce and Oswald, 2012) and, finally, in line with existing literature, becoming savers (for example, Guven, 2012; Frey and Stutzer, 2002). Furthermore, what is particularly noticeable is that both males and females, who report a non-zero score based upon the adjusted metrics, have an even lower probability of increasing educational attainment. For males, this is also evident for labour market transitions from unemployment into employment, *i.e.*, the negative effect of a non-zero score is more pronounced using the adjusted index relative to using the unadjusted index (there is no noticeable difference for females). A non-zero score is also associated with a lower probability of becoming a saver for both males and females across each alternative index, where again the alternative indices based upon the adjusted and robust metrics are typically larger in magnitude than the unadjusted index.

Moreover, what is also apparent is that the difference in the estimated parameters between the effects of a non-zero score based upon the original $GHQ-12$ and the alternative measures are generally statistically significant at the 5% level, as shown by the χ^2 statistics (the exception is female labour market transitions), with the magnitude of the coefficients typically being larger based upon the adjusted measures. This is perhaps not surprising given the inflation observed at the left hand extreme of the $GHQ-12$ distribution observed for both males and females.[14]

The results from these applications suggest that the over-reporting of the absence of

[13] When the results are based upon the adjusted metrics, *i.e.*, Columns 2 through to 4 for each outcome, given that the $GHQ-12$ measures are constructed from model estimates, the standard errors are bootstrapped using 200 replications.

[14] We have also investigated whether the persistence of reporting a non-zero score magnifies the likelihood of increasing educational attainment, transitioning from unemployment to employment, and becoming a saver. To do so, in equation (13) we condition on g_{it-1} and g_{it-2}, where $g_{it-2}=1$, if $GHQ-12\neq 0$ in period $t-2$. Interestingly, there is no evidence of long run effects on educational attainment or savings. However, for males and females, such persistence leads to a lower probability of transitioning into a state of employment.

mental health issues results in an under-estimate of its effect on transitions into improved economic states, such as employment and higher educational attainment. Such findings highlight the importance of allowing for potential mis-reporting in mental health measures from a policy perspective.

8 Conclusions

We have analysed the extent and implications of potential mis-reporting of mental health in the 12 sub-components of the $GHQ-12$, a very common and widely used measure of mental health. Using data from the British Household Panel Survey and Understanding Society over the period 1991 to 2018, we have employed inflated (latent-class type) models to account for a preponderance of zeros reported in the 12-item questionnaire. We then used posterior probabilities to adjust the $GHQ-12$ instrument. Importantly, the suggested approach is applicable to any health measure that is self-reported. The analysis shifts the distribution away from reporting no mental health issues. In our applications based upon using the adjusted measures, we find that over-reporting a score of zero for the $GHQ-12$ is generally associated with under-estimating the effect of mental health on a number of economic transitions relating to educational attainment, employment and financial vulnerability.

Furthermore, the $GHQ-12$ index was developed to screen for general (non-psychotic) psychiatric morbidity (Goldberg and Williams, 1988), and the finding that mis-reporting bias is associated with individuals over-estimating their state of mental health is of policy concern. Interestingly, older individuals are more likely to mis-report sub-components of the $GHQ-12$,[15] which given an ageing population is worrying, especially when such metrics are employed as screening tools in primary health care meaning that ultimately long-term health costs may be under-estimated.[16]

Countries such as the UK are collecting information at a national level on subjective wellbeing. Since 2011, the UK Office for National Statistics has routinely collected measures of subjective wellbeing in the large scale Integrated Household Survey (IHS). This has become particularly pertinent following the Commission on the Measurement of Eco-

[15] See Tables A.3 to A.6 in the Appendix.

[16] Such costs are potentially not trivial, with a recent independent review for the UK government showing that the cost of poor mental health to the economy is between £74 and £99 billion per year. See https://www.gov.uk/government/publications/thriving-at-work-a-review-of-mental-health-and-employers (last accessed 14/01/2021).

nomic Performance and Social Progress, (Stiglitz *et al.*, 2009), and stems from concerns that traditional measures of living standards, for example, GDP per capita, do not adequately reflect economic and social progress. Hence, investigating mis-reporting of mental health and seeking alternative ways to take this into account is an important line of future enquiry, given the increasing prominence of wellbeing as an economic indicator.[17]

Supplementary material

Supplementary material is available on the OUP website. The online appendix provides additional results tables as well as Stata and Gauss code for replication purposes. The data used in this paper are available from the UK data archive https://www.data-archive.ac.uk/ (last accessed 14/01/2021).

Funding

This work was supported by the Australian Research Council [DP140100748 to S.B., M.H and P.S.].

Acknowledgements

We are grateful to the Data Archive, University of Essex, for supplying the British Household Panel Surveys (BHPS), waves 1 to 18, and Understanding Society (UKHLS), waves 1 to 9. The BHPS data were originally collected by the ESRC Research Centre on Micro-social Change at the University of Essex (now incorporated within the Institute for Social and Economic Research). The UKHLS is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. Neither the original collectors of the data nor the Archive bear any responsibility for the analyses or interpretations presented here. We would like to thank the Editor and two anonymous referees for excellent

[17] A caveat with this is that psychiatric distress (the focus of our analysis) and subjective wellbeing are different concepts measured on alternative scales. However, mis-reporting may also be present in measures of subjective wellbeing and hence this is a potential avenue for future research, where the approach developed herein is applicable to any health measure that is self-reported.

comments. We are also grateful to Raslan Alzuabi for excellent research assistance and would like to thank Daniel Gray, Arne Rise Hole, Jennifer Roberts and attendees at the 2019 Royal Economic Society conference for valuable comments. The normal disclaimer applies.

References

Abrevaya, J. and Hausman, J. A. (1999). Semiparametric estimation with mismeasured dependent variables: An application to duration models for unemployment spells. *Annales d'Economie et de Statistique* **55–56**, 243–275.

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J. and Hebert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American Journal of Epidemiology* **161**, 389–398.

Angrist, J. and Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton, USA: Princeton University Press, first edition.

Bharadwaj, P., Pai, M. M. and Suziedelyte, A. (2017). Mental health stigma. *Economic Letters* **159**, 57–60.

Bloom, D. E., Cafiero, E., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L. R., Fathima, S., Feigl, A. B., Gaziano, T., Hamandi, A., Mowafi, M., Pandya, A., Prettner, K., Rosenberg, L., Seligman, B., Stern, A. Z. and Weinstein, C. (2011). The global economic burden of noncommunicable diseases. Technical report, Geneva: World Economic Forum.

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

Boyce, C. J. and Oswald, A. J. (2012). Do people become healthier after being promoted? *Health Economics* **21**, 580–596.

Brown, S., Harris, M., Srivastava, P. P. and Zhang, X. (2018). Modelling illegal drug participation. *Journal of the Royal Statistical Society: Series A* **181**, 133–154.

Brown, S., Taylor, K. and Wheatley-Price, S. (2005). Debt and distress: Evaluating the psychological cost of credit. *Journal of Economic Psychology* **26**, 642–663.

Clark, A. (2003). Unemployment as a social norm: Psychological evidence from panel data. *Journal of Labor Economics* **21**, 323–352.

Corbin, J. and Morse, J. (2003). The unstructured interactive interview: Issues of reciprocity and risks when dealing with sensitive topics. *Qualitative Inquiry* **9**, 335–354.

Cornaglia, F., Crivellaro, E. and McNally, S. (2015). Mental health and education decisions. *Labour Economics* **33**, 1–12.

Dustmann, C. and Fasani, F. (2015). The effect of local area crime on mental health. *The Economic Journal* **126**, 978–1017.

Dustmann, C. and Van Soest, A. (2001). Language fluency and earnings: Estimation with misclassified language indicators. *Review of Economics and Statistics* **83**, 663–674.

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

Frey, B. and Stutzer, A. (2002). What can economists learn from happiness research? *Journal of Economic Literature* **40**, 402–435.

Gardner, J. and Oswald, A. (2004). How is mortality affected by money, marriage, and stress? *Journal of Health Economics* **23**, 1181–1207.

Goldberg, D. and Williams, P. (1988). *A user's guide to the General Health Questionnaire*. NFER-Nelson.

Goldberg, D. P. and Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine* **9**, 139–145.

Greene, W. (2012). *Econometric Analysis 7e*. New Jersey, USA: Prentice Hall, seventh edition.

Guven, C. (2012). Reversing the question: Does happiness affect consumption and savings behaviour? *Journal of Economic Psychology* **33**, 701–717.

Hankins, M. (2008). The reliability of the twelve-item general health questionnaire (ghq-12) under realistic assumptions. *BMC public health* **8**, 1–7.

Hartmann, P. (1994). Interviewing when the spouse is present. *International Journal of Public Opinion Research* **6**, 298–306.

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

Hebert, J. R., Ebbeling, C. B., Matthews, C. E., Hurley, T. G., Yunsheng, M., Druker, S. and Clemow, L. (2002). Systematic errors in middle-aged women's estimates of energy intake: Comparing three self-report measures to total energy expenditure from doubly labeled water. *Annals of Epidemiology* **12**, 577–586.

Hinshaw, S. P. (2009). *The mark of shame: Stigma of mental illness and an agenda for change*. Oxford University Press.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics* **144**, 27–61.

Hu, Y., Stewart-Brown, S., Twigg, L. and Weich, S. (2007). Can the 12-item General Health Questionnaire be used to measure positive mental health? *Psychological Medicine* **37**, 1005–1013.

Jackle, A., Lynn, P., Sinibaldi, J. and Tipping, S. (2013). The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys. *Survey Research Methods* **7**, 1–15.

Jenkins, R., Bhugra, D., Bebbington, P., Brugha, T., Farrell, M., Coid, J., Fryers, T., Weich, S., Singleton, N. and Meltzer, H. (2008). Debt income and mental disorder in the general population. *Psychological Medicine* **38**, 1485–1493.

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

Kleinjans, K. J. and Van Soest, A. (2014). Rounding, focal point answers and nonresponse to subjective probability questions. *Journal of Applied Econometrics* **29**, 567–585.

Kristensen, N. and Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics* **15**, 96–117.

Lewbel, A. (2000). Identification of the binary choice model with misclassification. *Econometric Theory* **16**, 603–609.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology* **22**, 5–55.

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

Mahajan, A. (2006). Identification and estimation of regression models with misclassification. *Econometrica* **74**, 631–665.

Manski, C. F. and Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics* **28**, 219–231.

Mathers, C., Fat, D. and Boerma, J. (2008). *The Global Burden of Disease: 2004 Update*. World Health Organization.

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires*. Oxford university press.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

Metcalfe, R., Powdthavee, N. and Dolan, P. (2011). Destruction and distress: Using a quasi-experiment to show the effects of the September 11 attacks on mental well-being in the United Kingdom. *The Economic Journal* **121**, 81–103.

Meyer, B. and Mittag, N. (2017). Misclassification in binary choice models. *Journal of Econometrics* **200**, 295–311.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics* **144**, 81–117.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica* **46**, 69–85.

Niccoletti, C. and Peracchi, F. (2005). Survey response and survey characteristics: Micro level evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A* **168**, 763–781.

Ratcliffe, A. (2015). Wealth effects, local area attributes, and economic prospects: On the relationship between house prices and mental wellbeing. *Review of Income and Wealth* **61**, 75–92.

Ratcliffe, A. and Taylor, K. (2015). Who cares about stock market booms and busts? Evidence from data on mental wellbeing. *Oxford Economic Papers* **67**, 816–845.

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

Roberts, J., Hodgson, R. and Dolan, P. (2011). It's driving her mad: Gender differences in the effects of commuting on psychological health. *Journal of Health Economics* **30**, 1064–1076.

Schrapler, J. P. (2004). Respondent behaviour in panel studies: A case study for income-nonresponse by means of the German-Socio Economic Panel (SOEP). *Sociological Methods and Research* **33**, 118–156.

Scottish-Government (2013). Scottish health survey 2012 - volume 1 main report. Technical report, see http://www.scotland.gov.uk/Publications/2013/09/3684/5 (last accessed 14/01/2021).

Shevlin, M. and Adamson, G. (2005). Alternative factor models and factorial invariance of the ghq-12: a large sample analysis using confirmatory factor analysis. *Psychological Assessment* **17**, 231.

Stein, M. B., Cox, B. J., Afifi, T. O., Belik, S.-L. and Sareen, J. (2006). Does co-morbid depressive illness magnify the impact of chronic physical illness? A population-based perspective. *Psychological Medicine* **36**, 587–596.

Stern, S. (1989). Measuring the effect of disability on labor force participation. *Journal of Human Resources* **24**, 361–395.

Stiglitz, J., Sen, A. and Fitoussi, J.-P. (2009). Report by the commission on the measurement of economic performance and social progress. Technical report, https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf (last accessed 14/01/2021).

Thomas, C., Benzeval, M. and Stansfeld, S. A. (2005). Employment transitions and mental health: An analysis from the British Household Panel Survey. *Journal of Epidemiology and Community Health* **59**, 243–249.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

University of Essex (2019). Institute for Social and Economic Research, NatCen Social Research, Kantar Public. Understanding Society: Waves 1-9, 2009-2018 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 12th Edition. UK Data Service. SN: 6614, http://doi.org/10.5255/UKDA-SN-6614-13.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A. and Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society: Series A* **174**, 575–595.

Van Soest, A. and Hurd, M. (2008). A test for anchoring and yea-saying in experimental consumption data. *Journal of the American Statistical Association* **103**, 126–136.

Vassallo, R., Durrant, G. B. and Smith, P. W. F. (2015). Interviewer effects on non-response propensity in longitudinal surveys: A multilevel modelling approach. *Journal of the Royal Statistical Society: Series A* **178**, 83–99.

WHO (2001). The World Health Report 2001: Mental Health: New Understanding, New Hope. Https://www.who.int/whr/2001/en/ (last accessed 14/01/2021).

Wildman, J. (2003). Income related inequalities in mental health in Great Britain: Analysing the causes of health inequality over time. *Journal of Health Economics* **22**, 295–312.

Zipp, J. and Toth, J. (2002). She said, he said, they said. the impact of spousal presence in survey research. *Public Opinion* **2**, 177–208.

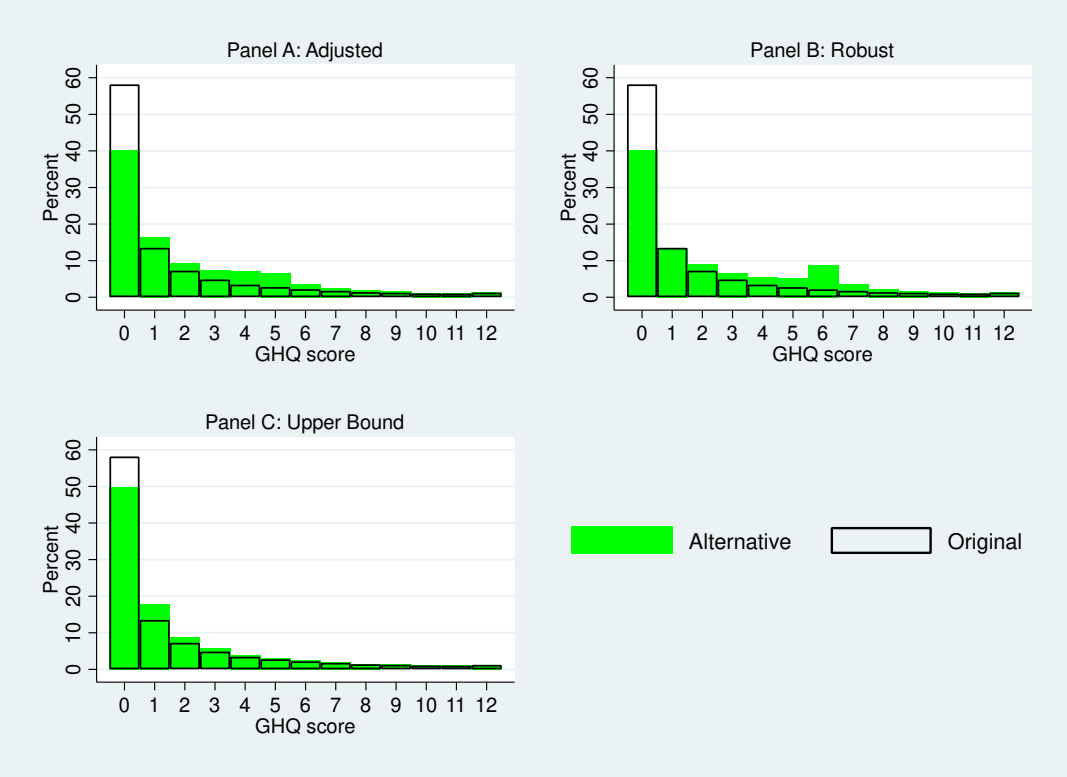

Figure 1: Males - Alternative $GHQ-12$ index adjusted using Posterior Probabilities

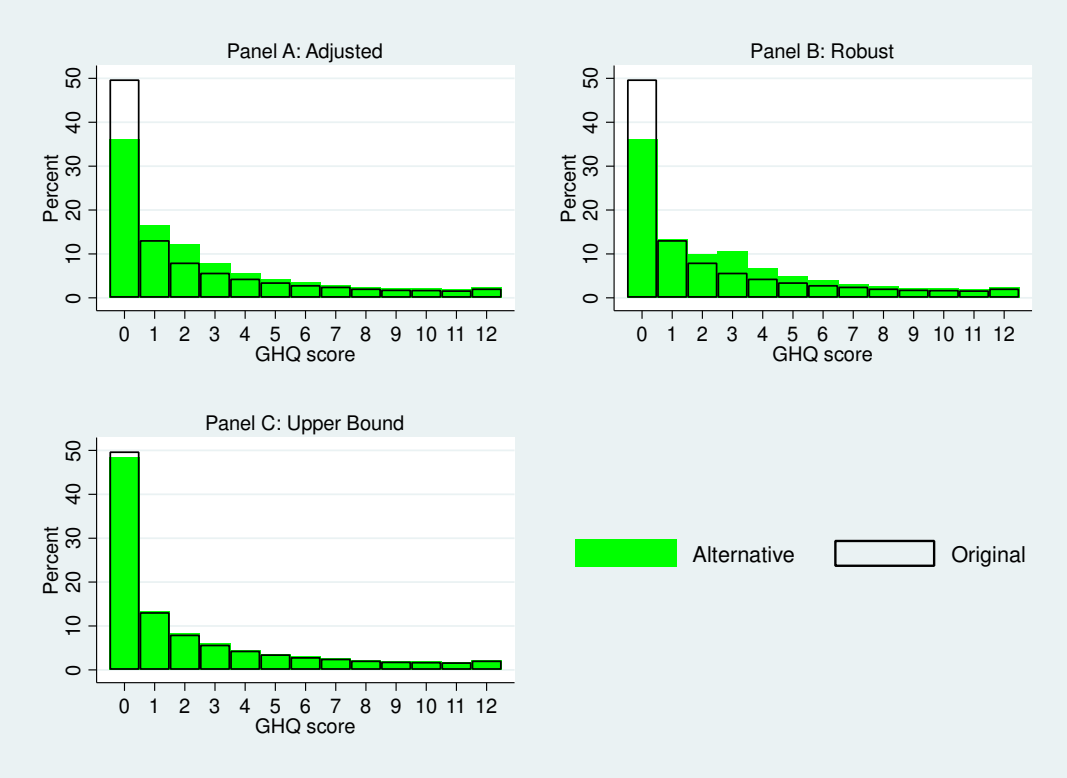

Figure 2: Females - Alternative $GHQ-12$ index adjusted using Posterior Probabilities

Table 1: Summary Statistics: $GHQ-12$ and Binary Sub-components

	MALES				FEMALES			
	Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum
Overall $GHQ-12$ index	1.549	2.709	0	12	2.152	3.196	0	12
Sub-components of $GHQ-12$								
GHQ1 – concentration	0.154	0.361	0	1	0.216	0.412	0	1
GHQ2 – sleep loss	0.146	0.353	0	1	0.219	0.413	0	1
GHQ3 – usefulness	0.118	0.323	0	1	0.143	0.350	0	1
GHQ4 – capability	0.076	0.264	0	1	0.116	0.320	0	1
GHQ5 – strain	0.230	0.421	0	1	0.292	0.455	0	1
GHQ6 – overcoming difficulties	0.115	0.319	0	1	0.164	0.370	0	1
GHQ7 – enjoy activities	0.168	0.374	0	1	0.206	0.404	0	1
GHQ8 – face up to problems	0.086	0.280	0	1	0.135	0.342	0	1
GHQ9 – unhappy or depressed	0.174	0.379	0	1	0.240	0.427	0	1
GHQ10 – losing confidence	0.109	0.311	0	1	0.173	0.378	0	1
GHQ11 – worthless person	0.061	0.239	0	1	0.094	0.292	0	1
GHQ12 – feeling reasonably happy	0.114	0.318	0	1	0.154	0.361	0	1
Individuals (N)	14,531				16,382			
Observations ($N \times T$)	122,247				148,056			

Table 2: Males - Predicted Probabilities and Reporting Bias for Individual $GHQ-12$ Components

	Proportion of reported psychological distress	Predicted rate of psychological distress	Reporting bias	Predicted marginal probability of misreporting zeros	Posterior Probabilities	
					0-Score	mis-reporting
		$\mathbf{Pr}(\widetilde{y}=1\|\mathbf{x})$	%	$\mathbf{Pr}(r=0\|\mathbf{x})$	$\mathbf{Pr}(\widetilde{y}=0\|\mathbf{x},\mathbf{y}=\mathbf{0})$	$\mathbf{Pr}(\widetilde{y}=1,r=0\|\mathbf{x},\mathbf{y}=\mathbf{0})$
	(1)	(2)	(3)	(4)	(5)	(6)
GHQ1	0.154	0.329	-114%	0.229	0.778	0.222
		(0.014)***		(0.017)***	(0.017)***	(0.017)***
GHQ2	0.146	0.247	-70%	0.148	0.856	0.144
		(0.013)***		(0.016)***	(0.016)***	(0.016)***
GHQ3	0.118	0.229	-93%	0.402	0.850	0.150
		(0.024)***		(0.025)***	(0.027)***	(0.027)***
GHQ4	0.076	0.101	-33%	0.086	0.953	0.047
		(0.005)***		(0.018)***	(0.006)***	(0.006)***
GHQ5	0.230	0.422	-83%	0.212	0.715	0.285
		(0.010)***		(0.010)***	(0.013)***	(0.013)***
GHQ6	0.115	0.331	-189%	0.367	0.742	0.258
		(0.022)***		(0.034)***	(0.026)***	(0.026)***
GHQ7	0.168	0.426	-154%	0.355	0.684	0.316
		(0.020)***		(0.022)***	(0.026)***	(0.026)***
GHQ8	0.086	0.061	28%	0.094	0.996	0.004
		(0.003)***		(0.051)***	(0.002)***	(0.002)***
GHQ9	0.174	0.257	-48%	0.138	0.886	0.114
		(0.009)***		(0.017)***	(0.012)***	(0.012)***
GHQ10	0.109	0.275	-154%	0.309	0.796	0.204
		(0.017)***		(0.031)***	(0.020)***	(0.020)***
GHQ11	0.061	0.047	23%	0.422	0.992	0.008
		(0.002)***		(0.033)***	(0.001)***	(0.001)***
GHQ12	0.114	0.197	-72%	0.181	0.881	0.119
		(0.010)***		(0.017)***	(0.012)***	(0.012)***

Note: GHQ1 – concentration; GHQ2 – sleep loss; GHQ3 – usefulness; GHQ4 – capability; GHQ5 – strain; GHQ6 – overcoming difficulties; GHQ7 – enjoy activities; GHQ8 – face up to problems; GHQ9 – unhappy or depressed; GHQ10 – losing confidence; GHQ11 – worthless person; GHQ12 – feeling reasonably happy. Standard errors are given in parentheses.* significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 3: Females - Predicted Probabilities and Reporting Bias for Individual $GHQ-12$ Components

	Proportion of reported psychological distress	Predicted rate of psychological distress	Reporting bias	Predicted marginal probability of misreporting zeros	Posterior Probabilities	
					0-Score	mis-reporting
		$\mathbf{Pr}(\widetilde{y}=1\|\mathbf{x})$	%	$\mathbf{Pr}(r=0\|\mathbf{x})$	$\mathbf{Pr}(\widetilde{y}=0\|\mathbf{x},\mathbf{y}=\mathbf{0})$	$\mathbf{Pr}(\widetilde{y}=1,r=0\|\mathbf{x},\mathbf{y}=\mathbf{0})$
	(1)	(2)	(3)	(4)	(5)	(6)
GHQ1	0.216	0.335	-55%	0.143	0.839	0.161
		(0.011)***		(0.013)***	(0.015)***	(0.015)***
GHQ2	0.219	0.316	-44%	0.110	0.899	0.101
		(0.009)***		(0.016)***	(0.012)***	(0.012)***
GHQ3	0.143	0.403	-181%	0.368	0.698	0.302
		(0.017)***		(0.023)***	(0.022)***	(0.022)***
GHQ4	0.116	0.139	-21%	0.062	0.961	0.039
		(0.005)***		(0.015)***	(0.006)***	(0.006)***
GHQ5	0.292	0.510	-75%	0.172	0.754	0.246
		(0.011)***		(0.014)***	(0.018)***	(0.018)***
GHQ6	0.164	0.153	7%	0.003	0.997	0.003
		(0.003)***		(0.001)***	(0.001)***	(0.001)***
GHQ7	0.206	0.447	-117%	0.301	0.693	0.307
		(0.017)***		(0.019)***	(0.023)***	(0.023)***
GHQ8	0.135	0.200	-48%	0.071	0.931	0.069
		(0.006)***		(0.012)***	(0.009)***	(0.009)***
GHQ9	0.240	0.423	-77%	0.160	0.802	0.198
		(0.011)***		(0.015)***	(0.017)***	(0.017)***
GHQ10	0.173	0.203	-17%	0.028	0.970	0.030
		(0.004)***		(0.005)***	(0.005)***	(0.005)***
GHQ11	0.094	0.099	-5%	0.068	0.967	0.033
		(0.004)***		(0.015)***	(0.005)***	(0.005)***
GHQ12	0.154	0.194	-25%	0.066	0.964	0.036
		(0.005)***		(0.013)***	(0.006)***	(0.006)***

Note: GHQ1 – concentration; GHQ2 – sleep loss; GHQ3 – usefulness; GHQ4 – capability; GHQ5 – strain; GHQ6 – overcoming difficulties; GHQ7 – enjoy activities; GHQ8 – face up to problems; GHQ9 – unhappy or depressed; GHQ10 – losing confidence; GHQ11 – worthless person; GHQ12 – feeling reasonably happy. Standard errors are given in parentheses.* significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 4: Males - Application of the Adjusted $GHQ-12$ to Modelling Transitions in Economic Outcomes

	Educational attainment *(increase in highest qualification obtained)*				Labour market status *(unemployed to employee)*				Savings *(non-saver to saver)*			
	λ (1)	λ (2)	λ (3)	λ (4)	λ (1)	λ (2)	λ (3)	λ (4)	λ (1)	λ (2)	λ (3)	λ (4)
Definition of $GHQ-12$:												
A: Original	0.003				-0.131***				-0.061***			
	(0.016)				(0.020)				(0.011)			
B: Adj. 1		-0.024**				-0.185***				-0.071***		
		(0.013)				(0.024)				(0.012)		
C: Adj. 2			-0.024**				-0.185***				-0.071***	
			(0.017)				(0.024)				(0.012)	
D: Adj. 3				0.034				-0.179***				-0.069***
				(0.022)				(0.022)				(0.012)
Obs. ($N \times T$)	107,716				87,495				107,716			
χ^2 equality	4.68				4.97				3.74			
$\lambda(1) = \lambda(2)$	*p=0.028*				*p=0.026*				*p=0.038*			
χ^2 equality	4.68				4.97				3.74			
$\lambda(1) = \lambda(3)$	*p=0.028*				*p=0.026*				*p=0.038*			
χ^2 equality	1.83				3.28				0.04			
$\lambda(1) = \lambda(4)$	*p=0.176*				*p=0.04*				*p=0.834*			

Note: results in each column are based upon random effects probit estimates conditioning on a quadratic in age, marital status, total income, housing tenure, year of interview and region of residence. Additional controls in the educational attainment models are labour market status. Additional controls in the labour market status models are highest educational attainment. The savings model includes both labour market status and highest educational attainment. Coefficients are reported with associated standard errors given in parentheses. The label "Adj. 1" refers to the *adjusted* method, "Adj. 2" refers to the *robust* method and "Adj. 3" refers to the *upper bound*, as described in section 6. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 5: Females - Application of the Adjusted $GHQ-12$ to Modelling Transitions in Economic Outcomes

	Educational attainment *(increase in highest qualification obtained)*				Labour market status *(unemployed to employee)*				Savings *(non-saver to saver)*			
	λ (1)	λ (2)	λ (3)	λ (4)	λ (1)	λ (2)	λ (3)	λ (4)	λ (1)	λ (2)	λ (3)	λ (4)
Definition of GHQ − 12 :												
A: Original	-0.022**				-0.087***				-0.031***			
	(0.010)				(0.016)				(0.009)			
B: Adj. 1		-0.042**				-0.084***				-0.061***		
		(0.017)				(0.018)				(0.010)		
C: Adj. 2			-0.035**				-0.084***				-0.056***	
			(0.017)				(0.018)				(0.010)	
D: Adj. 3				-0.023**				-0.090***				-0.034***
				(0.010)				(0.017)				(0.009)
Obs. ($N \times T$)	131,674				104,989				131,674			
χ^2 equality	3.01				0.04				7.04			
$\lambda(1) = \lambda(2)$	*p=0.04*				*p=0.843*				*p=0.008*			
χ^2 equality	3.01				0.04				7.04			
$\lambda(1) = \lambda(3)$	*p=0.04*				*p=0.843*				*p=0.008*			
χ^2 equality	0.18				0.02				0.26			
$\lambda(1) = \lambda(4)$	*p=0.672*				*p=0.875*				*p=0.612*			

Note: results in each column are based upon random effects probit estimates conditioning on a quadratic in age, marital status, total income, housing tenure, year of interview and region of residence. Additional controls in the educational attainment models are labour market status. Additional controls in the labour market status models are highest educational attainment. The savings model includes both labour market status and highest educational attainment. Coefficients are reported with associated standard errors given in parentheses. The label "Adj. 1" refers to the *adjusted* method, "Adj. 2" refers to the *robust* method and "Adj. 3" refers to the *upper bound*, as described in section 6. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.