Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Analyzing Subjective Well-Being Data with Misclassification*

Ekaterina Oparina & Sorawoot Srisuma

Analyzing Subjective Well-Being Data with Misclassification*

Ekaterina Oparina

London School of Economics

Sorawoot Srisuma‡

NUS and University of Surrey

*We are most grateful to Nattavudh Powdthavee for various discussions and suggestions especially at the early stage of this research. We thank the editor and two anonymous referees for comments and suggestions that helped improve the paper. We also thank Jo Blanden, Valentina Corradi, Szabolcs De ák, Mark Fabian, Zoe Fannon, Tomáš Jagelka, Anthoulla Phella, Luca Rondina, Yuya Sasaki, Matt Shum, Daniel Wilhelm, the seminar participants at Academia Sinica, IAAE Conference (2019), ESEM (2019) and the RES Junior Symposium (2019) for helpful comments and discussions.

‡*E-mail address*: s.srisuma@surrey.ac.uk

Abstract

We use novel nonparametric techniques to test for the presence of non-classical measurement error in reported life satisfaction (LS) and study the potential effects from ignoring it. Our dataset comes from Wave 3 of UK Understanding Society that is surveyed from 35,000 British households. Our test finds evidence of measurement error in reported LS for the entire dataset as well as for 26 out of 32 socioeconomic subgroups in the sample. We estimate the joint distribution of reported and latent LS nonparametrically in order to understand the mis-reporting behavior. We show this distribution can then be used to estimate parametric models of latent LS. We find measurement error bias is not severe enough to distort the main drivers of LS. But there is an important difference that is policy relevant. We find women tend to over-report their latent LS relative to men. This may help explain the gender puzzle that questions why women are reportedly happier than men despite being worse off on objective outcomes such as income and employment.

JEL CLASSIFICATION NUMBERS: C14, C51, I31

KEYWORDS: Measurement error, identification, subjective well-being, testing.

1 Introduction

Happiness or well-being economics first appeared in the economics literature in the early 1970s, see Van Praag (1971), Easterlin (1974). This fast growing, yet sometimes polarizing, subject studies causes and consequences of subjective well-being (SWB) and has provided many interesting insights into what makes people happy. Some of which have led to important policy lessons such as the idea that unemployment in the Western society is largely involuntary (Winkelmann and Winkelmann (1998)), that reducing the rates of joblessness should take priority over reducing the inflation rates (Di Tella et al. (2001)), that people partially adapt to serious disability over time (Oswald and Powdthavee (2008)), and that cigarette taxes actually improve the happiness of the likely smokers (Gruber and Mullainathan (2006)).

The central variable used in the well-being literature is life satisfaction (LS). LS is originally designed to capture the respondent's global well-being (Diener et al. (1985)). While LS has been shown to be correlated with a range of economic factors such as health and unemployment in expected ways, there is also ample evidence from the experimental literature showing that the reporting of LS is affected by irrelevant factors including mood, passing events, survey design, and pressures to provide socially desirable answers (see, e.g., Schwarz and Clore (1983), Feddersen et al. (2016), Diener et al. (2013)). We can therefore view *reported* LS as a possible mismeasurement of *latent* LS. Given the discrete nature of the SWB responses measurement error is also known as a misclassification.

Misclassification is a form of non-classical measurement error. A mismeasured LS can cause bias in empirical studies in arbitrary way. For this reason, one of the main conclusions from the well-known article by Bertrand and Mullainathan (2001), entitled: "Do people mean what they say? Implications for Subjective Survey Data", suggests researchers should not use LS as a dependent variable. Nevertheless, understanding the determinant of LS is one of the most fundamental tasks in the well-being literature. Since there is no obvious solution to the measurement error problem, LS is still routinely used as the dependent variable and the potential effects from measurement error have been unaccounted for.

In this paper we use novel econometric techniques to formally test for the presence of measurement error in reported LS and, if it exists, account for it and study its potential effects. We use survey data of 35,000 British households from the UK Understanding Society taken between January 2011 and June 2013. This (Wave 3) dataset is unique in that it contains what we believe are suitable variables that enable us to test for measurement errors and use the misclassification model of Hu (2008) to identify the joint distribution of the reported and latent LS nonparametrically. In particular, the LS distribution will be able to provide insights into the (mis-)reporting probabilities for people of different demographic and socioeconomic groups. We can also use this distribution to identify the determinants of latent LS in popular parametric models in the literature such as linear projection and ordered response models (e.g., logit and probit) without observing latent LS. Our results can have important policy implications, whether it is for the purpose of helping policy makers identify suitable groups of individuals for an intervention or for quantifying impacts of policies based on latent LS as opposed to reported LS.

The estimates from our ordinal regression are to be interpreted through the median rather than the mean. Bond and Lang (2019) have shown the mean ranking of ordinal data generally cannot be identified unless strong assumptions (such as homoskedasticity in probit/logit models) are a priori imposed. Subsequently, they use a (heteroskedastic) ordered probit to show some of the most well-known empirical results in the happiness literature can be arbitrarily reversed. We follow Chen, Oparina, Powdthavee and Srisuma (2019), who point out that the median ranking of ordinal data can be identified under weaker conditions and it should be used instead of the mean in well-being analysis. In this paper we focus on estimating a heteroskedastic ordered probit model without specifying the form of heteroskedasticity parametrically.

We begin our empirical study by testing for the presence of measurement error in reported LS. We adopt the nonparametric approach suggested recently by Wilhelm (2018) that, under suitable conditions, stochastic dependency between some auxiliary variables conditioning on the reported variable can be used to detect measurement error.[1] We use a Kolmolgorov-Smirnov type statistic and find evidence of measurement error in the reported LS for the entire dataset as well as for 26 out of 32 socioeconomic subgroups in the sample.

Next, we estimate a model of latent life satisfaction conditioning on key socioeconomic variables. We find the main drivers of LS in the model with latent LS are the same as those with reported LS, suggesting that the bias from measurement error may not be substantial enough to distort the effects of the main factors. For example, marriage and health have clear positive impact on LS while income and education have insignificant effects that could otherwise be positive due to substitution effects with health. However, there is one notable difference. We find that women systematically report themselves to be more satisfied with lives than they actually are relative to men. Measurement errors may thus help us solve the *gender puzzle* that women are happier than men in spite of the fact that they are often associated with less favorable objective measures in terms of health, income and employment level (see Dolan et al. (2008), Stevenson and Wolfers (2009)).

The validity of our empirical results relies on the conditions of the misclassification model of Hu (2008) being satisfied. Hu's identification procedure requires assumptions on the respondents' reporting behavior as well as some restrictions on the relation between reported LS and two auxiliary variables. The auxiliary variables need to be appropriately correlated with latent LS and satisfy a conditional independence assumption. These requirements pose contrasting qualities somewhat analogous to finding a good instrument.

The selection of appropriate auxiliary variables requires a transparent interpretation of the origin of misreporting. We build on Bertrand and Mullainathan (2001) and take the source of error to come from the lack of mental effort in answering survey questions. The information respondents can use to reduce mental effort can include mental effort can include mood, passing events, survey design or social desirability. These do not contribute to the LS in general. The two auxiliary variables we select for identification are: (i) a measure of mental well-being that is derived from the General Health Questionnaire (GHQ); (ii) a derived measure of neuroticism, which is one of the traits that underlies one's personality. The latter is currently collected only in Wave 3 of the UK Understanding Society survey. A useful feature of our dataset is that our auxiliary variables are constructed from a series of questions. Some of these questions are more narrow and objective than others, which allow us to perform robustness checks on the conditional independence assumption by constructing different versions of auxiliary variables.

Our work makes three main contributions. (1) To the best of our knowledge, we are the first to apply the nonparametric test of Wilhelm (2018) and misclassification model of Hu (2008) to analyze SWB data. These novel econometric methods do not rely on unjustified parametric assumptions, allow for non-classical errors, and do not require validation data (cf. Bound and Krueger (1991), Chen et al. (2005)). The latter two features in particular seem to be necessary for making any progress in accounting for measurement errors in self-reported subjective variables. (2) Using Wave 3 of the UK Understanding Society data, we show statistically that measurement error exists in reported LS and it may be used to solve the gender puzzle in the well-being literature. Measurement errors can therefore have practical implications if policy makers are to make decisions based on reported as opposed to latent LS. (3) We extend identify commonly used parametric linear and ordinal regression models where the dependent variable is measured with error. Our identification strategies lead to closed-form estimators for finite dimensional parameters of interest once a nonparametric estimator for the joint distribution of all variables in the model becomes available. Our parametric identification results are simple but also appear to be new; together with the proposed estimators, they add to the econometrics literature by complementing the results in Section 3 of Hu (2008) where he considered parametric regression models with misclassified covariates.

We organize the rest of the paper as follows. Section 2 gives a background on the current use of SWB data and issues with measurement error. Section 3 presents an econometric model of LS, gives conditions for identification of the parameters of interest and discuss practical inference. Section 4 describes the test we use to detect possible measurement errors in the reported LS. The empirical application is in Section 5. Section 6 concludes. The Online Appendix provides supplementary materials on further data descriptions and additional estimation and test results to support our main findings.

2 Background

Our background section consists of three parts. Section 2.1 provides a brief overview for a measure of well-being. Section 2.2 summarizes the main approaches to analyzing SWB data as well as some recent criticisms. Section 2.3 discusses measurement error in the reporting of life satisfaction and provide an interpretation of it for our application.

2.1 Subjective well-being

Well-being research is motivated by the ambition to understand the key drivers of individual's well-being. SWB is an umbrella term that includes a person's *cognitive well-being* such as LS (i.e., a judgment one makes about one's life overall), *affective well-being* (i.e., frequency and intensity of experienced emotions), and *eudaimonic well-being* (i.e., sense of purpose and worthwhileness). The economics of happiness literature traditionally uses LS, rather than measures of emotional states as a dependent variable.

Here we list some of the stylized facts of this literature as summarized in the World Happiness Report (Helliwell et al. (2012)):

- Richer people are on average happier than poorer people;
- LS is highly positively correlated with mental and physical health;
- Marriage has a positive correlation with LS;
- LS is U-shaped in age;
- Unemployment is significantly detrimental to LS;

- In most developed countries women report higher LS than men, despite being worse off in measurable socioeconomic outcomes;
- There is little correlation between a person's education level and his/her LS, but education is indirectly related to happiness through its effect on income: education increases income and income increases happiness.

Given the subjective nature of LS, the overwhelming majority of the findings are based on self-reported assessment: respondents are asked to report how satisfied they are with their life on a given scale. This approach favors personal evaluation of global well-being over the views of potential experts. Despite earlier concerns, self-reported measures of life satisfaction are proven to have a degree of validity. They converge in expected ways with each other and with non-self-reported measures, such as those based on other people's reports and the behavior of the respondent (Diener (2009), Layard (2010)). They are also predictive of future behaviors, such as job quit, divorce, and suicide (Diener et al. (2017)).

2.2 Estimating life satisfaction

The most common feature of empirical studies in the well-being literature is to use reported LS as a dependent variable and other characteristics, such as income, gender, health, employment statuses, etc., as covariates. There are two distinct approaches in how life satisfaction is modeled. One treats LS as a *cardinal* variable and the other as an *ordinal* one. The statistical techniques used for the former are based on least squares estimation or direct comparisons between sample averages. For the ordinal case, ordered logit or probit models are typically used. Both approaches are widely used in practice. See Ferrer-i-Carbonell and Frijters (2004) for an account for some (dis-)similarities of results between the two approaches.

The econometric analysis of SWB data has recently come under heavy criticisms. Whether least squares regression or ordered probit/logit estimation is used, similar to most other economic fields, a typical approach researchers take is to then draw conclusions based on statements about the relative *mean* happiness between groups of individuals (e.g., men and women, employed and unemployed, or across countries, etc.). Critiques point out that this research ignores the fact that SWB data are ordinal in nature. And the mean ranking of ordinal variables is only identified when it is stable across all increasing transformation. For examples, this means unless relevant stochastic dominance conditions hold, the raw average ranking and signs of least squares estimates may be reversed by monotonically transforming the ordinal scale, see Schröder and Yitzhaki (2017) and Bond and Lang (2019). Importantly, this issue goes deeper than "using OLS to estimate a discrete dependent variable", as Bond and Lang (2019) also show the mean ranking of latent happiness from ordinal models can also be arbitrarily reversed. They use a heteroskedastic ordered probit model to illustrate it for some of the most well-known results in the happiness literature. Explicitly allowing for heteroskedasticity is important because a homoskedastic model a priori effectively assumes the mean ranking to be identified. See Theorem 1 of Bond and Lang (2014).

There are ways to analyze happiness data that avoid these criticisms. For examples, direct comparisons of probabilities or probability odds of certain events between groups are not affected (e.g., Easterlin (1995)). But such a descriptive approach has limited scope for incorporating covariates. Chen, Oparina, Powdthavee and Srisuma (2019) suggest one solution is to use the *median* instead of the *mean* as a mode of comparison. The median rank is stable across all increasing transformations. Furthermore, they highlight the fact that the median and the mean in symmetric parametric models, like probit and logit, are the same. The median has therefore been frequently estimated but only interpreted as the mean.[2] This fact instantly nullifies the reversal of prior results in Bond and Lang (2019) by simply interpreting those estimates through the median. To this end, our paper emphasizes the use of an ordered response model with heteroskedasticity. We show in Section 4 it is simple to estimate a heteroskedastic probit model even without specifying the form of the heteroskedasticity parametrically.

2.3 Measurement error

We first ask how does one interpret measurement error in the context of LS? Unlike many economic variables that are known to be measured with error, for examples labor force status (Feng and Hu (2013)), number of bidders in an auction (An et al. (2017)) or voting participation (Bernstein et al. (2001)), a measure of LS is a less tangible concept. We will proceed by describing the factors that are known to influence the reporting of LS scores. Then we will give an interpretation of measurement error for LS.

The conventional view from the psychology literature is that evaluating LS can be a non-trivial mental task. This comes from the fact that respondents have to make a large number of comparisons across multiple dimensions without any guidance or criteria (see e.g., Diener et al. (2013) and references therein). Some economists share this view. In particular, Bertrand and Mullainathan (2001) point to this as the source of measurement error as they state: “An even more fundamental problem is that respondents may make little mental effort in answering the question, such as by not attempting to recall all the relevant information”.

The lack of mental effort means that a judgment which a respondent constructs while answering LS question can be influenced by various irrelevant factors (Strack et al. (1991)). The well-being and psychology literatures refer to this as a “shortcut” in the sense that responders use easily accessible information to make the well-being judgment less mentally challenging. The notion of a mental shortcut casts a wide net as it can be derived from mood[3], passing events[4], survey design [5] and subconsciously conforming to social norms [6].

We distinguish misreporting due to shortcuts described above from deliberate misreporting. Particularly, responders can be pressured into conforming with social norms during face-to-face interviews. For example, as Bertrand and Mullainathan (2001) put it, ‘Respondents want to avoid looking bad in front of the interviewer’.[7] For this reason, contemporary survey designs have self-completion components for sensitive questions (Schwarz et al. (1991), Tourangeau and Smith (1996), Presser and Stinson (1998)). The survey that our dataset comes from has a self-completion module for LS and the auxiliary variables (to be introduced in the next section). Thus, we do not consider measurement error from deliberate misreporting in this paper.

Most existing empirical applications involving LS take *reported* LS at face value and often use it as the dependent variable in a regression analysis. In this work we make explicit that reported LS, X, is a combination of *latent* LS, X^*, and measurement error, u: [8]

$$X = X^* + u.$$

We build on the statement made by Bertrand and Mullainathan (2001) (as quoted above) and interpret latent LS to be what responder would have reported if she gave full mental effort to reporting LS, irrespective of her interpretation and comprehension of the task. Measurement error is taken to be the cumulation of *irrelevant factors* that may affect the reporting of latent LS. Irrelevant factors include those that relate to mental shortcuts as well as benign unintentional mistakes[9].

The measurement error in our model is a discrete random variable because X and X^* are discrete. This type of error is also known as misclassification. Misclassification is non-classical by nature. For example, given the number of values the variable can take is finite, extreme values can only be mismeasured in one direction, so a zero-mean error (conditional on the true value) is impossible. Furthermore, the error term can be correlated with the covariates that are typically used in LS analysis. For example, Barrington-Leigh and Behzadnejad (2017) use two major health surveys in Canada to show that women and individuals with poor health condition are more affected by weather.

We are interested in understanding economic determinants of latent LS rather than reported LS since the latter is contaminated by irrelevant factors. The discussion above indicates that statistical analysis using reported LS as the dependent variable is expected to produce the results which are biased in unknown direction if measurement error is ignored. This is indeed one of the main conclusions stated in Bertrand and Mullainathan (2001); we refer the reader to their work for further discussions.

We propose a model of LS that explicitly deals with misclassification error in a general way in the next section. The approach we take follows from the misclassification model of Hu (2008) that assumes all the variables in the model are discrete. The discrete setup is suitable for analyzing LS as most variables that are used in this literature are discrete or can naturally be discretized. A more general treatment that allows for some continuous variables can be found in Hu and Schennach (2008). We refer the reader to the surveys by Schennach (2013) and Hu (2017) for examples of applications that rely on this type of identification results.

3 Model and identification strategy

In this section we describe the model of misclassification of Hu (2008) in the context of our application. In Section 3.1 we introduce the key variables in the model and discuss the assumptions for nonparametric identification. We consider parametric identification in Section 3.2. Section 3.3 discusses the numerical aspects of estimation and inference.

In what follows, we use $f_{A|B}(a|b)$ to denote $\Pr[A=a|B=b]$ for random variables (vectors) A and B taking values a and b respectively, and $f_A(a)$ to denote the $\Pr[A=a]$. We denote a generic matrix whose $ij-$th element is m_{ij} by a bold font $\mathbf{M}:=(m_{ij})$.

3.1 Nonparametric identification

The variables in our model are (X^*, X, Y, Z, W). The latent and observed LS are respectively X^* and X. Y and Z are auxiliary variables that play a similar role to instruments to be specified below. W denotes a vector of conditioning variables. $X^*, X,$ and Z are discrete random variables that have the same finite number of support points. Y is a binary variable. W is also discrete in our application but our identification results do not assume this.

We observe (X, Y, Z, W) but not X^*. So we want to identify $f_{X^*,X,Y,Z,W}$ from $f_{X,Y,Z,W}$. This is possible by Theorem 1 of Hu (2008) under some conditions. The following four assumptions are a version of his sufficient conditions.

Assumption 1 (CI). (X, Y, Z) *are mutually independent conditional on* (X^*, W).

Assumption 2 (RNK). *For all w, the matrices* $\mathbf{M}_{X|X^*,W=w} := \left(f_{X|X^*,W}(x_i \mid x_j^*, w) \right)$ and $\mathbf{M}_{X^*|Z,W=w} := \left(f_{X^*|Z,W}(x_i^* \mid z_j, w) \right)$ *are invertible.*

Assumption 3 (UNQ). *For all w,* $E[Y \mid X^* = x_i^*, W = w]$ *is different for different i.*

Assumption 4 (ORD). *For all w,* $f_{X|X^*,W}(x_I \mid x_i^*, w)$ *is strictly increasing in* $i = 1, \dots, I$.

We now introduce our choice of (*Y*, *Z*) and focus on explaining why Assumptions 1 to 4 can reasonably hold in our application. To fix ideas, let *X* and X^* be measured on a three-point scale: 1 – “dissatisfied”, 2 – “neither satisfied nor dissatisfied” or 3 – “satisfied”. *Y* is a derived measure of neuroticism that is indicative of a responder’s emotional stability. [10] *Y* takes value 1 if the responder’s level of neuroticism is above the median of the sample and 0 otherwise. *Z* is a measure of mental health that is derived from the General Health Questionnaire (GHQ). Unlike the covariates, which are collected through a face-to-face interviewer, each respondent provides (X, Y, Z) in the self-completed parts of the UK Understanding Society survey. We will defer other details on the dataset until Section 5.

Assumption 1 is the *conditional independence* assumption. While it is easy to find three independent variables in isolation, the challenge is to also have them satisfy Assumptions 2 to 4. We select *Z* and *Y* carefully so that they contain information on latent LS and some other information that we treat as errors. The errors are such, that based on the discussion in Section 2.3, they are independent from the irrelevant factors that contaminate X^* and between themselves once we control for X^* and *W*. Recall that the irrelevant factors have two components. One comes from mental shortcuts and the other comes from benign reporting errors. We assume the latter to be independent of everything and focus on the former.

We argue that Assumption 1 holds for the following reasons. First, the error in *Y* is orthogonal to the other components. This is because questions that are used to derive *Y* aim to learn about the respondent's personality traits rather than her assessments of LS or mental state. Particularly, neuroticism is based on an individual's stable characteristics that is unlikely to be affected by the irrelevant factors we associate with mental shortcuts for LS. On the other hand, *Z*, which is a measure of mental well-being, may be reasonably viewed as a second measure of X^* after *X*. However, there are important differences between the LS and GHQ questions. The LS question is a single question that involves a complex concept as the respondent is asked to put a score on her current life satisfaction overall. GHQ consists of a series of questions. Some are more objective or narrow in nature (e.g., amount of sleep or ability to concentrate) and they all concern recent changes in experiences. Questions that have better defined evaluation criteria require lower effort to answer and the respondents may not need to use a shortcut in the first place (Schwarz and Strack (1999)).[11] And one can argue any GHQ questions affected by mental shortcuts may do so in a different and independent way to the shortcut that affects evaluation of overall LS. This applies even for the GHQ question that appears closest to the LS question, which is "*Have you recently been feeling reasonably happy, all things considered*?" that has the following possible answers: 1 – "better than usual", 2 – "same as usual", 3 – "less than usual" and 4 – "much less than usual". The differences in questions and possible answers make apparent the different dimensions of well-being an individual is being asked to assess in the GHQ relative to the overall LS.[12] Therefore it does not seem implausible that conditional independence can hold. The empirical results in Section 5 use all of the questions in the GHQ to construct *Z*. We provide a sensitivity analysis by using different subsets of questions within the GHQ to construct *Z* in Appendix D.

Assumption 2 is a technical condition that ensures invertibility of certain matrices in the identification proof. In particular, when Assumption 1 holds, it can be shown Assumption 2 has an implication on observables as it is equivalent to the condition that $\mathbf{M}_{X,Z|W=w} := (f_{X,Z|W=w}(x_i, z_j \mid w))$ is an invertible matrix for all *w*.

Assumption 3 says that the probability that an individual whose level of neuroticism is above the median of the sample differs across sub-populations partitioned by X^*. Since neuroticism captures personal trait, which has been shown to be strongly related to the level of LS (see, e.g., Diener et al. (2009)), we expect this condition to hold.

Assumption 4 assumes two things. One, it assumes the respondent is more likely to report "satisfied" if their latent LS state is "satisfied" than at other states. Two, it assumes misreporting latent LS as "satisfied" is more likely to be gradual than extreme. These are intuitively plausible assumptions. Note that we are not ruling out truth-telling[13] where the respondent is more likely to report their latent LS state for other states as well. We want to explicitly impose conditions against extreme misreporting that can be interpreted as the respondent giving out random answers[14]. Technically, Assumption 4 is used to pin down a particular order in the matrix of eigenvectors following a diagonalization. We use the reported "satisfied" state as the anchor because it has an empirical significance, as it is the mode of the reported LS distribution (56% of the respondents in our sample report "satisfied", also see Figure 1 in Appendix A). A general violation of Assumption 4 indicates that reported LS may not have meaningful association with other aspects of well-being such as health or interpersonal relationship; similarly, it should have little meaningful predictive power on realized outcomes such as divorce and suicide. But this would go against the basic claims and findings from the multidisciplinary well-being literature that reported LS is correlated with various outcomes in intuitive ways as we have discussed at the end of Section 2.1.

Under Assumptions 1 to 4 we can identify $f_{X^*,X,Y,Z,W}$ from $f_{X,Y,Z,W}$. The former gives a complete characterization of the stochastic relation between all the variables in the model. We next show how it can be used to identify commonly used parametric models.

3.2 Parametric identification

Empirical studies are most often interested in the coefficients of linear and probit/logit models of LS given a vector of covariates. We consider two parametric models that are most often used in practice and show how to identify the parameters of interest.

3.2.1 Linear projection model

Here X^* is treated as a cardinal variable. Suppose X^* is observed. Let $\tilde{W} = \left(1, W^\top\right)^\top$. We are interested in β_C, which comes from the following linear projection model:

$$X^* = \tilde{W}^\top \beta_C + \varepsilon, \text{ where } E\left[\tilde{W}\varepsilon\right] = 0. \quad (1)$$

Then we can identify β_C as a least squares solution under familiar conditions. We state this as a proposition without proof.

Proposition 1. *Suppose Assumptions 1 to 4 hold and* $\left(X^*, \tilde{W}\right)$ *satisfies (1). If* $E\left[\tilde{W}\tilde{W}^\top\right] < \infty$ *and has full rank, then*

$$\beta_C = \left(E\left[\tilde{W}\tilde{W}^\top\right]\right)^{-1} E\left[\tilde{W}X^*\right]. \ (2)$$

Under Assumptions 1 to 4 we can identify $E\left[\tilde{W}\tilde{W}^\top\right]$ and $E\left[\tilde{W}X^*\right]$ from $f_{X^*,X,Y,Z,W}$.

3.2.2 Ordered probit model

Now let X^* be an ordinal variable generated from an ordered response model. Suppose X^* is observed. We are interested in β_O, which comes from the following ordered probit model:

$$X^* = i \times \mathbf{1}\left[\mu_{i-1} < \tilde{W}^\top \beta_O + \sigma(W)\varepsilon \le \mu_i\right] \text{for } i = 1, \ldots, I, \qquad (3)$$

where $\left(\mu_i\right)_{i=1}^{I-1}$ is an increasing sequence of reals with $\mu_0 = -\infty$ and $\mu_I = +\infty$, $\sigma(W)$ denotes a skedastic function that is positive almost surely, and ε has a standard normal distribution.

We can interpret $\tilde{W}^{\top}\beta_O + \sigma(W)\varepsilon := U^*$ in the traditional way. I.e U^* is an underlying continuous well-being variable that gets transformed into discrete level of LS. By symmetry of the normal distribution $\tilde{W}^{\top}\beta_O$ is the conditional median (and mean) of U^*.

In what follows we denote the CDF of ε by Φ. It is well-known that an ordered probit is not identified and some normalizations have to be made. In this paper we set $(\mu_1, \mu_2) = (0,1)$.[15] Next, we show in Lemma 1 that σ is identified without further assumptions. The proof of this result uses the identification strategy from Chen and Khan (2003).

Lemma 1. *Suppose Assumptions 1 to 4 hold. Then σ is identified and*

$$\sigma(W) = \frac{1}{\Phi^{-1}\left(\Pr\left[X^* \le 2 \mid W\right]\right) - \Phi^{-1}\left(\Pr\left[X^* \le 1 \mid W\right]\right)}. \qquad (4)$$

Proof. From (3), we have:

$$\Pr\left[X^* = i \mid W\right] = \Phi\left(\frac{\mu_i - \tilde{W}^{\top}\beta_O}{\sigma(W)}\right) - \Phi\left(\frac{\mu_{i-1} - \tilde{W}^{\top}\beta_O}{\sigma(W)}\right), \quad i = 1, \ldots, I. \quad (5)$$

It then follows that

$$\Pr\left[X^* \le 1 \mid W\right] = \Phi\left(\frac{-\tilde{W}^{\top}\beta_O}{\sigma(W)}\right), \qquad (6)$$

$$\Pr\left[X^* \le 2 \mid W\right] = \Phi\left(\frac{1-\tilde{W}^{\top}\beta_O}{\sigma(W)}\right). \qquad (7)$$

So that $\frac{1}{\sigma(W)} = \Phi^{-1}\left(\Pr\left[X^* \le 2 \mid W\right]\right) - \Phi^{-1}\left(\Pr\left[X^* \le 1 \mid W\right]\right)$. Under Assumptions 1 to 4 $f_{X^*|W}$ is identified. Therefore σ is identified. ∎

An interesting feature of the heteroskedastic ordered response model above is that we only need information on $\Pr\left[X^* = i \mid W\right]$ for $i = 1, 2$ to identify σ even if I is larger than 3. In fact, the same can be said for the identification of β_O.

Suppose that $I \geq 3$, then the additional information from $\Pr\left[X^* = i \mid W\right]$ for $i \geq 3$ can be used for identifying $\mu_O := \left(\mu_3, \ldots, \mu_{I-1}\right)$.

Proposition 2. *Suppose Assumptions 1 to 4 hold and $\left(X^*, W\right)$ satisfies (3). If $E\left[\tilde{W}\tilde{W}^\top\right] < \infty$ and has full rank, then*

$$\beta_O = \left(E\left[\tilde{W}\tilde{W}^\top\right]\right)^{-1} E\left[\tilde{W}\tilde{X}^*(W)\right], \qquad (8)$$

where $\tilde{X}^*(W) := -\sigma(W)\Phi^{-1}\left(\Pr\left[X^* = 1 \mid W\right]\right)$ *and*

$$\mu_i = \tilde{W}^\top \beta_O + \sigma(W)\Phi^{-1}\left(\Pr\left[X^* \leq i \mid W\right]\right) \text{ for } i = 3, \ldots, I-1. \qquad (9)$$

Proof. Re-arrange (6) to obtain,

$$-\sigma(W)\Phi^{-1}\left(\Pr\left[X^* = 1 \mid W\right]\right) = \tilde{W}^\top \beta_O.$$

Pre-multiply both sides of the display above by $\tilde{W}$. Take expectation and solve it to identify β_O.

We can identify μ_O by solving $\Pr\left[X^* \leq i \mid W\right] = \Phi\left(\dfrac{\mu_i - \tilde{W}^\top \beta_O}{\sigma(W)}\right)$ for all $i \geq 3$, where the latter expression is implied by (5). ∎

By inspecting the proof of Proposition 2, note that we can equivalently use (7) to identify β_O. In particular, the normalization restrictions impose the condition that $\sigma(W)\Phi^{-1}\left(\Pr\left[X^* \leq 1 \mid W\right]\right) = \sigma(W)\Phi^{-1}\left(\Pr\left[X^* \leq 2 \mid W\right]\right) - 1$.

Our discussion above assumes normality of ε in (3) for concreteness. Other parametric models, such as the logit, can be identified analogously by replacing Φ with another CDF of a continuous variable that has full support on $\mathbb{R}$.

3.3 Practical estimation and inference

The nonparametric identification result of Hu (2008) is known to be constructive. Our identification strategies for the parametric models in Section

3.2 are also constructive. We now discuss how to estimate the parameters of interest from data.

3.3 Nonparametric estimation

Suppose we have a random sample $\{(X_n, Y_n, Z_n, W_n)\}_{n=1}^N$ drawn from a population (X, Y, Z, W) that satisfies Assumptions 1 to 4. We can then consistently estimate $f_{X,Y,Z,W}$ under standard regularity conditions. One way to estimate $f_{X^*,X,Y,Z,W}$ is to trace through the proof of Theorem 1 in Hu (2008) and replace $f_{X,Y,Z,W}$ by its estimator. However, this approach may not work well in practice due to sampling error. In particular, the proof of Hu (2008) involves diagonalizing population matrices where their eigenvalues and eigenvectors are probabilities that are expected to satisfy properties of what probabilities are as well as the additional requirements imposed by Assumptions 3 and 4. Diagonalizing estimated matrices in practice may lead to outcomes that violate some of these conditions.[16]

We estimate $f_{X^*,X,Y,Z,W}$ by maximum likelihood where suitable constraints can be explicitly imposed. Many applications using related identification results of Hu (2008) take the same approach. For examples, see Hu (2017). In particular, under Assumption 1, we have

$$f_{W,X,Y,Z}(w,x,y,z) = \sum_{x^* \in \mathcal{X}^*} f_{X^*,X,Y,Z|W}(x^*,x,y,z \mid w) f_W(w)$$
$$= \sum_{x^* \in \mathcal{X}^*} f_{X|X^*,W}(x \mid x^*,w) f_{Y|X^*,W}(y \mid x^*,w) f_{Z|X^*,W}(z \mid x^*,w) f_{X^*|W}(x^* \mid w) f_W(w).$$

We construct a likelihood function based on the joint probability above where the parameters of interest are $\left(f_{X|X^*,W}, f_{Y|X^*,W}, f_{Z|X^*,W}, f_{X^*|W}, f_W\right)$. In our application the conditioning variables are all discrete. This is the norm in well-being applications and we will focus on this case. Then the maximum likelihood estimator of f_W corresponds to the empirical distribution of $\{W_n\}_{n=1}^N$, which can be obtained independently of the other parameters. Maximum likelihood estimation of the other parameters can be performed conditionally on W.

Let $\mathcal{S}_W, \mathcal{S}_X, \mathcal{S}_Y$ and $\mathcal{S}_Z$ denote the cardinalities of the support of *W*, *X*, *Y* and *Z*. Then for each *w* in the support of *W*, there are $\mathcal{S}_{XYZ} := \mathcal{S}_X \mathcal{S}_Y \mathcal{S}_Z$ possible realizations of (X, Y, Z). [17] We can enumerate these distinct events by $\{x_j, y_j, z_j\}_{j=1}^{\mathcal{S}_{XYZ}}$ coupled with $\{m_j\}_{j=1}^{\mathcal{S}_{XYZ}}$ where m_j counts how many times realization *j* occurs in the sub-sample when $W_n = w$. We then estimate the parameters of interest by maximizing the following conditional log-likelihood function

$$M_N(\mathbf{p}; w) = \sum_{j=1}^{\mathcal{S}_{XYZ}} m_j \ln \sum_{x^* \in \mathcal{X}^*} p_{X|X^*,W}\left(x_j \mid x^*, w\right) p_{Y|X^*,W}\left(y_j \mid x^*, w\right) p_{Z|X^*,W}\left(z_j \mid x^*, w\right) p_{X^*|W}\left(x^* \mid w\right), \tag{10}$$

where $\mathbf{p} = \left(p_{X|X^*,W}, p_{Y|X^*,W}, p_{Z|X^*,W}, p_{X^*|W}\right)$ lies in the parameter space $\mathcal{P}$ that satisfies the constraints that components of **p** constitute to valid probability distributions and the inequality relations in Assumption 4. We do this for all *w* in the support of *W*. Once the nonparametric estimators of $\left(f_{X|X^*,W}, f_{Y|X^*,W}, f_{Z|X^*,W}, f_{X^*|W}, f_W\right)$ are available, we can proceed to the parametric estimation stage.

Constrained maximum likelihood estimation is not a computationally simple task. There are $\mathcal{S}_X\left(\mathcal{S}_X + \mathcal{S}_Y + \mathcal{S}_Z - 3\right) + \mathcal{S}_X - 1$ free parameters to optimize over in (10) for each possible value that *W* takes. I.e., we have to solve this type of optimization problem $\mathcal{S}_W$ times. The numerical challenge increases with the support size of the variables in the model. Furthermore, the objective function is not concave so there can be many local maxima. In practice, we suggest numerical searches should be performed at different starting points in order to help locate the global maximum. See Lu et al. (2014) for a further discussion on the numerical aspects of maximum likelihood estimation in misclassification models.

3.3 Parametric Estimation

Once an estimator for $f_{X^*|W}$ is available, population quantities involving X^* such as $E\left[X^* \mid W\right]$ and $\Pr\left[X^* \le i \mid W\right]$ can be estimated even if we do not observe latent LS. For example for the linear probability model, from (2), we can write $\beta_C = \left(E\left[\tilde{W}\tilde{W}^{\top}\right]\right)^{-1} E\left[\tilde{W}E\left[X^* \mid \tilde{W}\right]\right]$. We can then estimate β_C by replacing the (unconditional) expectation by the sample counterparts.

For the ordered probit model, we can estimate σ by replacing $\Pr\left[X^* \leq i \mid W\right]$ in (4) by its estimator. Then we can construct estimators for β_O and μ_O by replacing the population moments in (8) and (9) respectively by their sample counterparts. Alternatively, a perhaps more convenient numerical approach is to estimate the parameters of interest with the build-in functions of statistical software providing it with the skedastic function based on (4).

3.3 Inference

We propose to perform inference by bootstrapping. A bootstrap sample can be generated by random resampling from the observed data with replacement. The estimators and tests of nonparametric probabilities and parameters in Propositions 1 and 2 have regular asymptotic properties that can be bootstraped as long as the true parameters lie in the interior of the parameter space (Andrews (1999, 2000)). In practice, estimates of probabilities being close to 0 or 1, or any other a priori (if used) constraints (Assumptions 3 and 4) that appear to be numerically binding should raise concerns that the assumption of an interior solution is not being satisfied.

4 Test for presence of measurement error

We want to test the hypothesis of no measurement error in LS:

$$H_0^A : \Pr\left[X = X^*\right] = 1. \qquad (11)$$

Suppose we have (X, Y, Z) that satisfies Assumptions 1 - 4 unconditionally[18]. Then we can identify $f_{X^*,X}$ from $f_{X^*,X,Y,Z}$. One way to test (11) directly is to look for evidence that $f_{X^*,X}\left(x^*, x\right) > 0$ for some $x^* \neq x$. But performing such test is difficult because the null would imply that $f_{X^*,X}\left(x^*, x\right) = 0$ for all $x^* \neq x$; parameters at the boundary will require a non-standard testing procedure. For example, see Andrews (2001). We instead follow the approach of

Wilhelm (2018), who shows it is possible to construct a simple test for the presence of measurement errors under weaker conditions and without the need to first identify the entire model.

Theorem 1 in Wilhelm (2018) states that: if $Y \perp Z \mid X^*$, then (11) implies $Y \perp Z \mid X$. (Here we use "$\perp$" to denote independence.) We can then construct a test to detect potential measurement errors based on a conditional independence hypothesis:

$$H_0^B : Y \perp Z \mid X. \qquad (12)$$

We state this as a proposition.

Proposition 3. *Suppose* $Y \perp Z \mid X^*$. *Then violation of* H_0^B *implies violation of* H_0^A.

Note that the conditional independence assumption in Proposition 3 is weaker than what is assumed in Assumption 1. Testing H_0^B is just a test of conditional independence on observed variables. There are many options available for consistent tests that are easy to construct. In this paper we use a Kolmogorov-Smirnov type statistic that is based on the sample counterpart of the following, equivalent, way to write (12):

$$H_0^B : \max_{(x,y,z)\in S_{XYZ}} \left| f_{Y,Z|X}(y,z \mid x) - f_{Y|X}(y \mid x)\, f_{Z|X}(z \mid x) \right| = 0.$$

In our application we use the frequency estimator for $\left(f_{Y,Z|X}, f_{Y|X}, f_{Z|X} \right)$, which corresponds to the maximum likelihood estimator since (X,Y,Z) are discrete. Denoting the frequency estimator by $\left(\hat{f}_{Y,Z|X}, \hat{f}_{Y|X}, \hat{f}_{Z|X} \right)$, we have the following test statistic:

$$TS = \max_{(x,y,z)\in S_{XYZ}} \left| \hat{f}_{Y,Z|X}(y,z \mid x) - \hat{f}_{Y|X}(y \mid x)\, \hat{f}_{Z|X}(z \mid x) \right|. \quad (13)$$

We perform inference by bootstrapping. We construct bootstrap critical values for *TS* from the percentiles of $\left\{ TS^b \right\}_{b=1}^{B}$, where

$$TS^b = \max_{(x,y,z)\in S_{XYZ}} \left| \hat{f}^b_{Y,Z|X}(y,z|x) - \hat{f}^b_{Y|X}(y|x)\,\hat{f}^b_{Z|X}(z|x) - \left(\hat{f}_{Y,Z|X}(y,z|x) - \hat{f}_{Y|X}(y|x)\,\hat{f}_{Z|X}(z|x) \right) \right|, \quad (14)$$

and $\hat{f}^b_{A|B}$ denotes the frequency estimator of $f_{A|B}$ based on the bootstrap sample. These bootstrap critical values are consistent as long as $f_{X,Y,Z}$ takes values in the interior of $(0,1)$ as discussed at the end of Section 3.[19]

It is worth emphasizing that Proposition 3 only provides a sufficient condition to detect measurement errors. On the other hand, H_0^B generally does not imply H_0^A unless additional conditions hold on the joint distribution of $f_{X^*,X,Y,Z}$. We refer the reader to Wilhelm (2018) for further details as to when the two hypotheses are equivalent.

5 Application

We begin this section by describing the dataset and explaining how it is used in our applications. We report the results of the test for the presence of measurement error in Section 5.2. We study the effect measurement error has on general models of LS in Section 5.3.

5.1 Data

We use Wave 3 of the representative household longitudinal data from UK Understanding Society. The survey covers members of over 35,000 households in the United Kingdom. These data were collected between January 2011 and June 2013. We choose Wave 3 because, unlike the other waves, it includes questions on personality traits, which is important for us as we use neuroticism as one of the auxiliary variables for identification.

Understanding Society measures LS on a scale from 1 – ‘completely dissatisfied’ to 7 – ‘completely satisfied’. For our application, we aggregate responses to the LS question into 3 larger groups, where 1st group is those dissatisfied with life overall (‘completely dissatisfied’ and ‘mostly dissatisfied’), 3rd group is those satisfied (‘mostly satisfied’ and ‘completely satisfied’) and the 2nd group is those in between (‘somewhat dissatisfied’, ‘neither satisfied or dissatisfied’ and ‘somewhat satisfied’). For the GHQ measure, which runs from 0 - ‘the least distressed’ to 36 – ‘the most distressed’, we construct Z to share the same cardinality as X by aggregating all responses below 33rd percentile in group 1, those between 33rd and 66th percentile in group 2, all the rest in group 3. The neuroticism score is originally calculated as an average of 3 questions on a scale from 1 to 7. Indicator Y takes the value of 1 if the level of neuroticism of the individual is above the sample median and zero otherwise.

We aggregate the data on LS to ensure stable solutions for our constrained maximum likelihood with both the observed and bootstrap samples. This reduces the number of parameters to be estimated from 97 to 17 for each possible realization of the conditioning variables. We try to avoid local maxima and dependence on the starting search value by maximizing each of our likelihood function 10 times using a different starting point. Almost all of our estimates converge to the same solution. On the other hand, if we use a 7-point scale for life satisfaction we often find numerical optimization starting at a different point leads to distinct local maxima; in this case we do not have the confidence that global solutions can be reached in feasible time.

We only use data for the respondents who reported satisfaction with life overall. This gives us 40,359 observations from the 49,739 available in the survey (over 81%). Of those, 56% are women, 44% are men. All the participants are of age 16 or above. 34% of the respondents have a long-standing illness or disability, 51.8% are married, 23.1% have obtained a university degree, 5.3% are unemployed.

Our W consists of: university degree (*degree*), gender (*fem*), long-standing illness or disability (*illness*), income above the sample median (*inc*) and marital status (*married*). Each of the covariates is a binary variable. That gives $\mathcal{S}_W = 2^5 = 32$. We are unable to condition on additional variables because some socioeconomic groups would have too few observations for nonparametric estimation. We compute our estimators as described in Section 3.3; in particular the skedastic function is nonparametric (see (4)).

We remark that we would prefer to use more conditioning variables in theory because it lessens the burden on the conditional independence assumption (see Assumption 1). However, related to the comment above, in finite sample there may be a problem of too few observations for some subsamples.

5.2 Measurement error in reported LS

We test for the presence of measurement error in reported LS unconditionally and conditionally on the covariates. The unconditional test assumes that Y and Z are independent conditionally on X^*. Each of our conditional test assumes a weaker independence assumption specific to a particular socioeconomic group. Our arguments for conditional independence given in Section 3.1 are applicable here as well.

The unconditional test assumes H_0^B under the null, uses (13) as the test statistic, and (14) to construct the critical values. Table 1 compares the value of the test statistic against the bootstrap critical values at the different significance levels. We find very strong evidence against the no measurement error hypothesis as H_0^B is rejected at 1% significance level.

We next look for the presence of measurement error across different socioeconomic groups. The conditional test partitions the data into $\mathcal{S}_W$ subgroups. In this case, for each $w \in \mathcal{S}_W$ we consider the following hypothesis:

$$H_0^C(w): \max_{(x,y,z)\in\mathcal{S}_{XYZ}} \left| f_{Y,Z|X,W}(y,z \mid x,w) - f_{Y|X,W}(y \mid x,w)\, f_{Z|X,W}(z \mid x,w) \right| = 0.$$

We alter (13) and (14) to accommodate the conditioning on W accordingly with the frequency estimator. They are then used respectively to construct test statistics and bootstrap critical values. The description of different socioeconomic groups and conditional test results can be found in Appendices B and C, respectively.

We find very strong evidence that measurement error exists for many subgroups of the population. In particular, we reject $H_0^C(w)$ at 1% significance level for 22 out 32 of socioeconomic groups. Out of the 10 groups we do not reject the null at 1%, we reject 4 of them at 5%. It is worth noting that the number of observations in these groups are very small relative to the rest, especially for the groups that we do not reject the null. The lack of (stronger) evidence to detect measurement error in some of those groups may be due to small sample size.[20]

5.3 Estimation results

Our main results will focus on the distribution of the reported and latent LS and their ordered probit estimates. In particular, parameter estimates from probit models are to be interpreted as a component of the conditional median of the underlying continuous happiness variable. Before we present them, we consider the effects of reducing the support of reported LS from a 7-point scale to a 3-point scale as well as from leaving out some other covariates. In addition to the variables we have already introduced we will also use: logarithm of gross personal income (*l_inc*), unemployment dummy (*unempl*) and age (*age*) and age squared (*age2*).[21]

Table 2 reports the estimates for the linear projection model and for the ordered probit model for reported LS with full support (7-point scale) and reduced support (3-point scale). Here we use personal income instead of the dummy indicator that a respondent's income is above the median or not. In this case the heteroskedastic ordered probit is fully parametric. It is estimated using the oglm STATA command, where the skedastic function is specified by an exponential function with a linear index (Williams (2010)), in order to abstract away from the need to select tuning parameters from nonparametric estimation (e.g., with kernel smoothing, see Chen and Khan (2003)).

Reducing the support of LS has negligible or no difference in how covariates affect LS apart from income, where the positive income effect on LS measured on a 3-point scale is much more pronounced. This pattern holds in all models. In particular, we note the similarities between the least squares and the probit estimates for all covariates (cf. Ferrer-i-Carbonell and Frijters (2004)) as well as the similarities between results from homoskedastic and heteroskedastic models (cf. Chen et al. (2008)[22]). These results are largely consistent with the literature. Married people are more satisfied with their lives than their non-married counterparts. Long-standing illnesses or disability and unemployment significantly reduce LS. Women report to be more satisfied than men. The effect from age supports the U-shaped pattern based on a quadratic specification. Money does buy some happiness. Although there are some conflicted findings on the income effect, the literature in general seems to find the support for the idea that income influences LS positively with diminishing returns (e.g. see Clark et al. (2008)). Education is known to influence LS indirectly through the increase in income and health. A positive effect from having more education is common result for the studies that cannot fully control for health[23], including those for the UK (see, e.g., Dolan et al., 2008).

Table 3 contains analogous statistics to Table 2 but is based on the reduced set of covariates that we later use to estimate latent LS. Most of the results between the two tables are qualitatively very similar. One notable difference, again, is on the income effect. Table 3 reports that the income effect remains positive in all cases for the reduced support LS. But the model with the full support LS yields negative estimates with the probit specification. The negative income effect is, however, weak as it is insignificant and significant at 10% in the homoskedastic and heteroskedastic cases respectively. Our discussion on the income effect from the previous paragraph applies. Importantly, Table 2 and 3 suggest that using a median income dummy and omitting unemployment and age have little impact, as well as reducing the support of LS.

We now provide the estimates from reported and latent LS. Table 4 reports the estimates from the linear projection model and the ordered probit models for reported and latent LS. Here the skedastic function for the heteroskedastic ordered probit model is estimated nonparametrically and the normalization is as discussed in Section 3.2.2.

The results show that the latent LS estimates are qualitatively very similar across all three models. There is a difference in the signs of the estimates of income, but the income effect is weak and insignificant. Comparing the results from the models with reported and latent LS, we find the two prominent predictors of LS agree on their effects: health and interpersonal relationships. People who suffer from long-standing illness or disability are less satisfied with their life, while married people are more satisfied than their single counterparts. While the effect of health becomes more pronounces when we control for the presence of measurement error, it appears to have substituted the effect on education and income, making them insignificant with very high p-values. The most striking difference we find is the gender effect: the female dummy has a positive coefficient for the reported LS, but negative for the latent one. Appendix D shows this finding is robust to different combinations of GHQ questions used to construct Z.

Our results provide a potential explanation of the *gender puzzle* based on systematic differences in the misreporting behavior between men and women. While LS has been widely accepted to be correlated with health and interpersonal relationship in obvious ways (see, e.g., Helliwell et al. (2012)), the correlation between well-being and gender observed in practice is less intuitive. Many surveys find that females report themselves to be more satisfied with their life than men, e.g., see Dolan et al. (2008). These findings are in contradiction with being worse off in many measurable social and economic outcomes, which are known to be the sources of well-being (pay gap and unemployment gap, to name a few). More recently, Meisenberg and Woodley (2015) use a dataset of 90 countries represented in the World Values Survey to find that gender equality, gainful employment and prolonged schooling decrease female well-being; while women are happier in the countries that maintain traditional gender roles.

In order to better understand the difference in reporting behavior of men and women, we consider 4 particular socioeconomic groups of respondents. Group 0 contains single men with income below the median, with no degree and no long-standing illness or disability. Group H contains the same respondents who suffer from long standing health issues. The other two groups are the female respondents with the same characteristics. Distributions for the other groups are presented in Appendix E.

Comparing the upper (0 and H) and the lower (F and FH) blocks of Table 5 explains the different signs of the gender dummy coefficients in the two models. The distribution of reported LS, $\mathbf{M}_X$, is similar for men (upper block) and women (lower block) with women slightly more likely to report the high state, hence positive coefficient for the gender dummy. The comparison of latent distributions, $\mathbf{M}_{X^*}$, shows the opposite: women are more likely to be in the low state and less likely to be in the high one. However, we do not observe lower levels of LS among women in the data, because they misreport in a systematically different way compared to men. Comparing the matrices of misreporting probabilities, $\mathbf{M}_{X|X^*}$, shows that though all the respondents are prone to report higher states that they latently are, women do it more emphatically. I.e., women are more likely than men to report the highest state, regardless of their latent state.

Our econometric analysis can identify differences in reporting behavior but does not provide a behavioral answer to rationalize them. At the moment we can only offer potential explanations. One particular conjecture is based on the distinct gender patterns of conforming to social roles and social stereotypes. Kahneman et al. (1999) and references therein suggest that according to traditional gender roles women are usually seen as more cheerful and enthusiastic. As a result, women might report higher states conforming with the existing norm.

6 Conclusion

There is an enormous interest in using subjective well-being data in economics and related disciplines. Existing research almost always ignores measurement error despite the fact that the literature acknowledges its likely presence. In particular, the error is non-classical and its potential effects on subsequent analysis is completely unknown. In this paper we use novel nonparametric techniques to formally test for the presence of measurement error and empirically investigate its effects. We also extend the existing nonparametric identification results to identify parametric models that are commonly used in the literature, namely the linear projection and probit models, as part of our analysis.

Our tests are based on the idea proposed in Wilhelm (2018) and we use a misclassification model of Hu (2008). The application of these nonparametric methods in itself is not entirely trivial. Primarily there is an empirical challenge in finding appropriate data that satisfies assumptions somewhat analogous to finding a good instrument. We use Wave 3 of the UK Understanding Society survey because it is the only wave that contains questions on neuroticism that we believe is crucial for identification.

We find evidence of measurement error in LS for the whole sample as well as 26 out of 32 socioeconomic subgroups of the data. We use covariates that define these subgroups to estimate parametric models of LS. We find most important drivers of LS affect latent LS and reported LS in the same way. But there is a notable difference in the gender effect. The happiness literature often finds women reporting higher levels of well-being than men despite being worse off in measurable objective outcomes (e.g., income and employment). This is known as the *gender puzzle*. The puzzle can be rationalized by our model because women are more likely to report themselves to be happier than they actually are compared to men.

The puzzling relations between female well-being and socioeconomic measures were also found in the panel data. Stevenson and Wolfers (2009) show that reported well-being of women in the United States declined in the last 35 years despite the improvement of women's positions in many objective outcomes. The authors label this result as *The Paradox of Declining Female Happiness*. In order to further investigate whether the gender puzzle can be explained by measurement error, we need to have panel data to extend our analysis.

One methodological recommendation of our research is for future surveys to consider collecting data that increase the scope to apply modern econometric techniques to solve old problems like measurement error. For example, the UK Understanding Society data is in fact longitudinal. But the lack of information on neuroticism from all Waves other than Wave 3 prevents us from controlling from individual specific effects that would be very helpful in well-being studies.

References

An, Y., Baye, M. R., Hu, Y., Morgan, J., and Shum, M. (2017). Identification and Estimation of Online Price Competition With an Unknown Number of Firms. *Journal of Applied Econometrics*, 32(1):80–102.

Andrews, D. W. K. (1999). Estimation When a Parameter is on a Boundary. *Econometrica*, 67(6):1341–1383.

Andrews, D. W. K. (2000). Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space. *Econometrica*, 68(2):399–405.

Andrews, D. W. K. (2001). Testing When a Parameter Is on the Boundary of the Maintained Hypothesis. *Econometrica*, 69(3):683–734.

Barrington-Leigh, C. and Behzadnejad, F. (2017). The impact of daily weather conditions on life satisfaction: Evidence from cross-sectional and panel data. *Journal of Economic Psychology*, 59:145–163.

Bernstein, R., Chadha, A., and Montjoy, R. (2001). Overreporting Voting: Why It Happens and Why It Matters. *The Public Opinion Quarterly*, 65(1):22–44.

Bertrand, M. and Mullainathan, S. (2001). Do People Mean What They Say? Implications for Subjective Survey Data. *American Economic Review*, 91(2):67–72.

Bond, T. and Lang, K. (2014). The Sad Truth About Happiness Scales. Technical Report w19950, National Bureau of Economic Research, Cambridge, MA.

Bond, T. N. and Lang, K. (2019). The Sad Truth about Happiness Scales. *Journal of Political Economy*, 127(4):1629–1640.

Bound, J. and Krueger, A. B. (1991). The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? *Journal of Labor Economics*, 9(1):1–24.

Chen, L. Y., Oparina, E., Powdthavee, N., and Srisuma, S. (2019). Have Econometric Analyses of Happiness Data Been Futile? A Simple Truth About Happiness Scales. *Working paper*.

Chen, S. and Khan, S. (2003). Rates of convergence for estimating regression coefficients in heteroskedastic discrete response models. *Journal of Econometrics*, 117(2):245–278.

Chen, X., Hong, H., and Tamer, E. (2005). Measurement Error Models with Auxiliary Data. *The Review of Economic Studies*, 72(2):343–366.

Clark, A. E., Frijters, P., and Shields, M. A. (2008). Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles. *Journal of Economic Literature*, 46(1):95–144.

Clark, H. H. and Schober, M. F. (1992). Asking questions and influencing answers. In *Questions about questions: Inquiries into the cognitive bases of surveys*, pages 15–48. Russell Sage Foundation, New York, NY, US.

Di Tella, R., MacCulloch, R. J., and Oswald, A. J. (2001). Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness. *The American Economic Review*, 91(1):335–341.

Diener, E. (2009). *Assessing Well-Being: The Collected Works of Ed Diener*. Springer Science & Business Media.

Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, 49(1):71–75.

Diener, E., Inglehart, R., and Tay, L. (2013). Theory and Validity of Life Satisfaction Scales. *Social Indicators Research*, 112(3):497–527.

Diener, E., Oishi, S., and Lucas, R. (2009). Subjective Well-Being: The Science of Happiness and Life Satisfaction. *The Oxford Handbook of Positive Psychology*.

Diener, E., Oishi, S., and Tay, L. (2018). Advances in subjective well-being research. *Nature Human Behaviour*.

Diener, E., Sandvik, E., Pavot, W., and Gallagher, D. (1991). Response Artifacts in the Measurement of Subjective Well-Being. *Social Indicators Research*, 24(1):35–56.

Diener, E., Tay, L., Heintzelman, S. J., Kushlev, K., Wirtz, D., and Lutes, L. D. (2017). Findings All Psychologists Should Know From the New Science on Subjective Well-Being. *Canadian Psychology/Psychologie canadienne*, 58(2):87–104.

Dolan, P., Peasgood, T., and White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, 29(1):94–122.

Easterlin, R. A. (1974). Does economic growth improve the human lot?: some empirical evidence. *Nations and households in economic growth: essays in honor of Moses Abramovitz*.

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization*, 27(1):35–47.

Feddersen, J., Metcalfe, R., and Wooden, M. (2016). Subjective wellbeing: why weather matters. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(1):203–228.

Feng, S. and Hu, Y. (2013). Misclassification Errors and the Underestimation of the US Unemployment Rate. *American Economic Review*, 103(2):1054–1070.

Ferrer-i-Carbonell, A. and Frijters, P. (2004). How Important is Methodology for the estimates of the determinants of Happiness? *The Economic Journal*, 114(497):641–659.

Giné, E. and Zinn, J. (1990). Bootstrapping General Empirical Measures. *The Annals of Probability*, 18(2):851–869.

Gruber, J. and Mullainathan, S. (2006). Do Cigarette Taxes Make Smokers Happier? In Ng, Y.-K. and Ho, L. S., editors, *Happiness and Public Policy: Theory, Case Studies and Implications*, pages 109–146. Palgrave Macmillan UK, London.

Helliwell, J., Layard, R., and Sachs, J. (2012). World happiness report. LSE Research Online Documents on Economics, London School of Economics and Political Science, LSE Library.

Holtgraves, T. (2004). Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding. *Personality and Social Psychology Bulletin*, 30(2):161–172.

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

Hu, Y. (2017). The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of Econometrics*, 200(2):154–168.

Hu, Y. and Schennach, S. M. (2008). Instrumental Variable Treatment of Nonclassical Measurement Error Models. *Econometrica*, 76(1):195–216.

Kahneman, D., Diener, E., and Schwarz, N., editors (1999). *Well-being: The foundations of hedonic psychology.* Well-being: The foundations of hedonic psychology. Russell Sage Foundation, New York, NY, US.

Kaminska, O. and Foulsham, T. (2016). Eye-tracking Social Desirability Bias. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 130(1):73–89.

Layard, R. (2010). Measuring Subjective Well-Being. *Science*, 327(5965):534–535.

Lee, M. J. (1992). Median regression for ordered discrete response. *Journal of Econometrics*, 51(1):59–77.

Lu, R., Luo, Y., and Xiao, R. (2014). An MPEC estimator for misclassification models. *Economics Letters*, 125(2):195–199.

Mahajan, A. (2006). Identification and Estimation of Regression Models with Misclassification. *Econometrica*, 74(3):631–665.

Manski, C. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333.

McCrae, R. R. and Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1):81–90.

Meisenberg, G. and Woodley, M. A. (2015). Gender Differences in Subjective Well-Being and Their Relationships with Gender Equality. *Journal of Happiness Studies*, 16(6):1539–1555.

Oswald, A. J. and Powdthavee, N. (2008). Does happiness adapt? A longitudinal study of disability with implications for economists and judges. *Journal of Public Economics*, 92(5):1061–1077.

Presser, S. and Stinson, L. (1998). Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance. *American Sociological Review*, 63(1):137–145.

Schennach, S. M. (2013). *Measurement Error in Nonlinear Models – A Review*, volume 3 of *Econometric Society Monographs*, pages 296–337. Cambridge University Press.

Schröder, C. and Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92:337–358.

Schwarz, N. and Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3):513–523.

Schwarz, N. and Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In *Well-being: The foundations of hedonic psychology*, pages 61–84. Russell Sage Foundation, New York, NY, US.

Schwarz, N., Strack, F., Hippler, H.-J., and Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5(3):193–212.

Schwarz, N., Strack, F., Kommer, D., and Wagner, D. (1987). Soccer, rooms, and the quality of your life: Mood effects on judgments of satisfaction with life in general and with specific domains. *European Journal of Social Psychology*, 17(1):69–79.

Stevenson, B. and Wolfers, J. (2009). The Paradox of Declining Female Happiness. *American Economic Journal: Economic Policy*, 1(2):190–225.

Strack, F., Argyle, M., and Schwarz, N. (1991). *Subjective Well-being: An Interdisciplinary Perspective*. International series in experimental social psychology. Pergamon Press.

Tourangeau, R. and Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, 60(2):275–304.

Van Praag, B. (1971). The welfare function of income in Belgium: An empirical investigation. *European Economic Review*, 2(3):337–369.

Wilhelm, D. (2018). Testing for the Presence of Measurement Error. *CeMMAP Working Paper CWP45/18*.

Williams, R. (2010). Fitting heterogeneous choice models with oglm. *The Stata Journal*, 10(4):540–567.

Winkelmann, L. and Winkelmann, R. (1998). Why are the Unemployed So Unhappy? Evidence from Panel Data. *Economica*, 65(257):1–15.

Table 1 Unconditional test for the presence of measurement error.

TS	Critical values		
	90%	95%	99%
0.106***	0.007	0.008	0.011

* p <0.10, ** p <0.05, *** p <0.01

Table 2 Linear projection and ordered probit models estimates for the model with reported LS (full and reduced support) and full list of covariates.

	Linear model		Homoskedastic ordered probit		Heteroskedastic ordered probit	
	Full support	Reduced support	Full support	Reduced support	Full support	Reduced support
degree	0.212***	0.104***	0.201***	0.163***	0.175***	0.143***
	(0.0182)	(0.0080)	(0.0216)	(0.0124)	(0.0212)	(0.0165)
fem	0.0316**	0.0167**	0.054***	0.027***	0.068***	0.049***
	(0.0154)	(0.0067)	(0.0180)	(0.0103)	(0.0181)	(0.0120)
health	-0.488***	-0.195***	-0.588***	-0.300***	-0.585***	-0.360***
	(0.0166)	(0.0073)	(0.0196)	(0.0110)	(0.0291)	(0.0256)
mrd	0.290***	0.124***	0.355***	0.191***	0.375***	0.248***
	(0.0166)	(0.0073)	(0.0194)	(0.0110)	(0.0233)	(0.0195)
l_inc	0.0132*	0.00899***	0.00126	0.0144***	0.0150*	0.0203***
	(0.00683)	(0.00299)	(0.0081)	(0.0046)	(0.0089)	(0.0049)
unempl	-0.550***	-0.216***	-0.546***	-0.290***	-0.516***	-0.267***
	(0.0373)	(0.0163)	(0.0431)	(0.0237)	(0.0496)	(0.0273)
age	-0.0518***	-0.0194***	-0.0662***	-0.0300***	-0.0762***	-0.0435***
	(0.00251)	(0.00110)	(0.00297)	(0.00170)	(0.00403)	(0.00282)
age2	0.000599***	0.000227***	0.000770***	0.000356***	0.000878***	0.000520***
	(0.0000247)	(0.0000108)	(0.0000292)	(0.0000168)	(0.0000430)	(0.0000328)

* p <0.1, ** p <0.05, *** p <0.01

Table 3 Linear projection and ordered probit models estimates for the model with reported LS (full and reduced support) and reduced list of covariates.

	Linear model		Homoskedastic ordered probit		Heteroskedastic ordered probit	
	Full support	Reduced support	Full support	Reduced support	Full support	Reduced support
degree	0.132***	0.0710***	0.103***	0.111***	0.0847***	0.0803***
	(0.0183)	(0.00798)	(0.0213)	(0.0124)	(0.0198)	(0.0146)
fem	0.0286*	0.0170**	0.0393**	0.0284***	0.0360**	0.0495***
	(0.0153)	(0.00664)	(0.0177)	(0.0101)	(0.0175)	(0.0118)
illness	-0.409***	-0.161***	-0.483***	-0.245***	-0.466***	-0.301***
	(0.0158)	(0.0069)	(0.0183)	(0.0104)	(0.0185)	(0.0114)
income	0.0367**	0.0293***	-0.0172	0.0435***	-0.0328*	0.0379***
	(0.0157)	(0.00683)	(0.0182)	(0.0104)	(0.0182)	(0.0121)
mrd	0.244***	0.108***	0.289***	0.170***	0.291***	0.232***
	(0.0150)	(0.00653)	(0.0175)	(0.0100)	(0.0175)	(0.0128)

* p <0.1, ** p <0.05, ** * p <0.01

Table 4 Linear projection and ordered probit models estimates for the models with reported and latent LS and reduced list of covariates.

	Linear model		Homoskedastic ordered probit		Heteroskedastic ordered probit	
	Reported LS	Latent LS	Reported LS	Latent LS	Reported LS	Latent LS
degree	0.0710***	-0.0503	0.112***	-0.0754	0.0811***	-0.126
	(0.00813)	(0.0491)	(0.0130)	(0.0750)	(0.0148)	(0.0988)
fem	0.0170***	-0.135***	0.0284***	-0.198***	0.0505***	-0.146*
	(0.00636)	(0.0480)	(0.0096)	(0.0732)	(0.0110)	(0.0826)
illness	-0.161***	-0.279***	-0.245***	-0.408***	-0.305***	-0.353***
	(0.00703)	(0.0423)	(0.0103)	(0.0645)	(0.0122)	(0.0698)
income	0.0293***	-0.00297	0.0436***	0.00171	0.0368***	0.0530
	(0.00661)	(0.0423)	(0.0100)	(0.0638)	(0.0119)	(0.0662)
mrd	0.108***	0.116***	0.170***	0.169***	0.240***	0.131**
	(0.00563)	(0.0386)	(0.00861)	(0.0585)	(0.0105)	(0.0638)

* p <0.1, ** p <0.05, *** p <0.01

Table 5 Distribution of reported and latent LS.

0	H
$\mathbf{M}_X = \begin{bmatrix} 0.0887 & 0.3606 & 0.5507 \\ (0.0054) & (0.0083) & (0.0088) \end{bmatrix}$	$\mathbf{M}_X = \begin{bmatrix} 0.1657 & 0.4709 & 0.3634 \\ (0.0100) & (0.0151) & (0.0142) \end{bmatrix}$
$\mathbf{M}_{X\mid X^*} = \begin{bmatrix} 0.2395 & 0.0617 & 0.0720 \\ (0.0518) & (0.0109) & (0.0082) \\ 0.6915 & 0.4945 & 0.1460 \\ (0.0379) & (0.0382) & (0.0216) \\ 0.0691 & 0.4437 & 0.7819 \\ (0.0614) & (0.0366) & (0.0223) \end{bmatrix}$	$\mathbf{M}_{X\mid X^*} = \begin{bmatrix} 0.4562 & 0.0554 & 0.0562 \\ (0.0656) & (0.0325) & (0.0192) \\ 0.5438 & 0.5899 & 0.2541 \\ (0.0641) & (0.0609) & (0.0435) \\ 0.0001 & 0.3547 & 0.6898 \\ (0.0177) & (0.0712) & (0.0454) \end{bmatrix}$
$\mathbf{M}_{X^*} = \begin{bmatrix} 0.1254 & 0.4195 & 0.4551 \\ (0.0347) & (0.0516) & (0.0459) \end{bmatrix}$	$\mathbf{M}_{X^*} = \begin{bmatrix} 0.2745 & 0.4090 & 0.3165 \\ (0.0493) & (0.0530) & (0.0519) \end{bmatrix}$
F	FH
$\mathbf{M}_X = \begin{bmatrix} 0.0940 & 0.3226 & 0.5834 \\ (0.0050) & (0.0076) & (0.0076) \end{bmatrix}$	$\mathbf{M}_X = \begin{bmatrix} 0.1472 & 0.4402 & 0.4125 \\ (0.0078) & (0.0109) & (0.0105) \end{bmatrix}$
$\mathbf{M}_{X\mid X^*} = \begin{bmatrix} 0.1525 & 0.0649 & 0.0751 \\ (0.0270) & (0.0092) & (0.0096) \\ 0.5947 & 0.3794 & 0.0902 \\ (0.0402) & (0.0350) & (0.0179) \\ 0.2528 & 0.5557 & 0.8346 \\ (0.0596) & (0.0360) & (0.0179) \end{bmatrix}$	$\mathbf{M}_{X\mid X^*} = \begin{bmatrix} 0.2662 & 0.0882 & 0.0448 \\ (0.0284) & (0.0187) & (0.0163) \\ 0.6216 & 0.4497 & 0.0924 \\ (0.0258) & (0.0434) & (0.0342) \\ 0.1121 & 0.4621 & 0.8628 \\ (0.0382) & (0.0523) & (0.0384) \end{bmatrix}$
$\mathbf{M}_{X^*} = \begin{bmatrix} 0.2847 & 0.3068 & 0.4085 \\ (0.0460) & (0.0588) & (0.0403) \end{bmatrix}$	$\mathbf{M}_{X^*} = \begin{bmatrix} 0.3831 & 0.4061 & 0.2108 \\ (0.0562) & (0.0440) & (0.0421) \end{bmatrix}$

Notes

[1] The insight that conditional independence can indicate the presence of measurement error was first explored in a regression context by Mahajan (2006), who considers a binary regressor that may be measured with error.

[2] Estimating the median without any parametric distributional assumption is also possible (Manski (1985), Lee (1992)). Chen et al. (2019) suggest the semiparametric median can be estimate using modern constrained mixed integer optimization technique; they apply it to study the Easterlin paradox.

[3] Multiple experiments have shown that LS scores can be influenced by mood changing events like finding a dime in a copy machine, spending time in a pleasant environment or watching a football team win (Schwarz et al. (1987)).

[4] In a large-scale survey setting, the responses can be influenced by weather (Schwarz and Clore (1983)); there are well-known diurnal and day-of-the-week variations in SWB (see Diener et al. (2018)).

[5] Respondents have been found to provide answers consistent with the previous ones, so the ordering and phrasing of questions matters (Clark and Schober (1992)).

[6] Providing socially desirable answers may be cognitively easier than performing all the necessary comparisons and forming a judgment. See Holtgraves (2004) and Kaminska and Foulsham (2016).

[7] The socially desirable responding specific to SWB is ‘happy image management’ that would result in reporting higher or lower well-being than experienced to appear happier/less happy (Diener et al. (1991)).

[8] In this paper we use the term *latent* to mean a measurement without error. In Section 3.2.2 we model X^* using an ordered response model, which traditional interprets X^* to be derived from an underlying continuous

happiness variable that plays an analogous role to utility in McFadden's random utility maximization model.

[9] One can imagine there can be unintentional reporting mistakes even when respondents are not taking mental shortcuts. We do not digress in this direction because we will assume such error satisfies the conditional independence assumption when it comes to identification.

[10] A neurotic individual can be defined by such terms as worrying, insecure, self-conscious, and temperamental (McCrae and Costa (1987)).

[11] E.g., experiments by Strack et al. (1991) show that though satisfaction with life overall is influenced by the mood manipulations, they didn't find the significant effect on the evaluation of life domains.

[12] All questions and possible answers from the survey that are used to construct (X, Y, Z) can be found in Appendix A.

[13] The truth-telling assumption corresponds to Assumption 2.7 in Hu (2008). It would also suffice for identification as an alternative to our Assumption 4.

[14] Truth-telling assumption does not impose any conditions on the off-diagonal elements of $\mathbf{M}_{X|X^*,W=w}$ other than they have to be less than the main diagonal within each column.

[15] Alternatively normalizations can be made on β_O. For example, the intercept can be set to 0 and one of the slope parameters can be set to 1.

[16] E.g., we obtained complex eigenvalues as well as eigenvectors that have both positive and negative elements from diagonalization matrices with our dataset.

[17] We assume the joint support of (W, X, Y, Z) is the same for all realizations of W for notational simplicity.

[18] For the ease of exposition and to keep the notation as close as possible to related papers on testing we omit the covariates in this section.

[19] Let $\mathcal{F}(x,y,z) := f_{Y,Z|X}(y,z|x) - f_{Y|X}(y|x) f_{Z|X}(z|x)$ for $(x,y,z) \in \mathcal{S}_{XYZ}$. It is clear that $\mathcal{F}(x,y,z)$ is a continuous function of $f_{X,Y,Z}$. Under random sampling, the asymptotic distribution of $\sqrt{N}\left(\hat{f}_{X,Y,Z} - f_{X,Y,Z}\right)$ can be consistently estimated by $\sqrt{N}\left(\hat{f}^b_{X,Y,Z} - \hat{f}_{X,Y,Z}\right)$ since empirical measures can be bootstrapped (e.g., see Giné and Zinn (1990)). The asymptotic percentiles of *TS* can then be consistently estimated using $\left\{TS^b\right\}_{b=1}^{B}$ by an application of the Continuous Mapping Theorem.

[20] The results in Appendix C are in fact conservative relative to equivalent tests that are based on,

$$H_0^D(w): \max_{(x,y,z)\in\mathcal{S}_{XYZ}} \left| f_{X|W}(x|w) f_{X,Y,Z|W}(x,y,z|w) - f_{X,Y|W}(x,y|w) f_{X,Z|W}(x,z|w) \right| = 0,$$

where we would reject the no measurement error hypothesis for all but 2 cases at 5% or lower significance level. The results of these tests are available upon request.

The above hypothesis can be useful in small sample or when $f_{X|W}(x|w)$ is close to 0 as we can avoid divisions by zeros since the event $\hat{f}_{X|W}(x|w) = 0$ and $\hat{f}^*_{X|W}(x|w) = 0$ both have positive probabilities. (We did not experience such event in our empirical study.)

[21] We need to reduce the support of LS for numerical stability of the maximum likelihood procedure and limit the number and support of covariates in order to ensure there is a sufficient number of observations with each socioeconomic group. For example, from Appendix C, we have 10 socioeconomic groups with under 500 observations (with DH the lowest at 117). If we split the sample further with employment status (only 5.3% are unemployed) and age bands, there will be groups with too few observations to estimate 17 parameters.

[22] This empirical indifference is in stark contrast to the theoretical implication illustrated in Bond and Lang (2019).

[23] The dataset does not allow us to fully control for the state of health and we only account for the presence of long-standing illness or disability.