Article:

Davillas, Apostolos, Jones, Andrew Michael orcid.org/0000-0003-4114-1785 and de Oliveira, Victor Hugo (2023) Is inconsistent reporting of self-assessed health persistent and systematic? Evidence from the UKHLS. Economics and Human Biology. 101219. ISSN 1570-677X

https://doi.org/10.1016/j.ehb.2022.101219

Is inconsistent reporting of self-assessed health persistent and systematic? Evidence from the UKHLS*

Apostolos Davillas
Department of Economics, University of Macedonia, Greece

Victor Hugo de Oliveira
Health Economics Group, University of East Anglia

Andrew M Jones
Department of Economics and Related Studies, University of York

Abstract

In this paper, we investigate whether individuals provide consistent responses to self-assessed health (SAH) questions in the UK Household Longitudinal Study (UKHLS), and the potential implications for empirical research in case of inconsistent reporting behaviour. We capitalise on an opportunity in the UKHLS, asking respondents the same SAH question twice: with a self-completion and an open interview mode, within the same household interview over four waves. We estimate multivariate models to explore which individual characteristics are systematically relevant for the likelihood and frequency of inconsistent reporting. About 11%-24% of those reported a particular SAH category in the self-completion reported inconsistently in the open interview. The probability of inconsistency is systematically associated with individual's demographics, education, income, employment status, cognitive and non-cognitive skills. The same characteristics also predict the frequency of inconsistent reporting across four UKHLS waves. Analysis of the implications of reporting inconsistencies shows no impact of SAH measurement on the association between income and health. A set of dimensions of people's physiological and biological health, captured using biomarkers, is associated equally with both SAH measures, suggesting that the interview mode does not play a role in the relationship between SAH and more objective health measures.

Keywords: measurement error; reporting bias; self-assessed health; UKHLS

JEL codes: C10, C33, C83, I10

Corresponding author: Apostolos Davillas, Department of Economics, University of Macedonia, 156 Egnatia Street, Thessaloniki, Greece; a.davillas@uom.edu.gr.

---

*Understanding Society is an initiative funded by the Economic and Social Research Council (ESRC) 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. The research data are distributed by the UK Data Service. We acknowledge financial support from the ESRC (ES/V003356/1). The funders, data creators and UK Data Service have no responsibility for the contents of this paper. We thank Paul Clarke and Philip Clarke for their useful comments on an earlier version of our study.

1. Introduction

The interplay between socioeconomic circumstances and health is of major importance for well-being and for human capital investment. To investigate these complex links, studies have often used survey data that combine information on social and economic circumstances with self-reported health measures. This is mostly due to the limited availability of datasets that combine more objective health measures with a wide range of socioeconomic data, along with the simplicity that self-reported health measures may offer. Among self-reported health measures, self-assessed health (SAH) is widely used in economics (e.g., Aoki and Santiago, 2018; Contoyannis and Jones, 2004; Contoyannis, Jones and Rice, 2004; Currie, Duque and Garfinkel, 2015; García-Gómez, Jones and Rice, 2010; Johnson, 2010; van Doorslaer et al., 2000), and in social research (e.g., Monden, 2010; Monk, 2015). Moreover, SAH measures are commonly used in epidemiological and medical research where an association with mortality has been demonstrated (e.g., Jylhä, 2009; Kaplan and Camacho, 1983; Kunst et al., 2004; Mossey and Shapiro, 1982; Hu et al., 2016).[1]

Studies have sought a better understanding of the extent to which SAH can be interpreted as a good proxy of underlying health between respondents of different socioeconomic backgrounds and in general (e.g., Au and Johnston, 2014; Bago d'Uva et al., 2008; Dowd and Zajacova, 2010; Etilé and Milcent, 2006; Lindeboom and van Doorslaer, 2004). A related issue to whether or not SAH is a good proxy of people's health is the reliability of SAH it terms of consistency to responses to SAH questions from the same individuals in a short time interval where genuine changes in people's health are not practically feasible. For example, a reliable proxy of people's health measure should be consistent in the case of repeated collection of the same measure when people's actual health doesn't change – i.e., in the context of our research, there should be consistency in responses to SAH within the same interview.

A short literature has tested the consistency of responses to SAH questions, by comparing repeated SAH questions for the same individuals collected over a short time (e.g., Black et al., 2017; Chen et al., 2021; Clarke and Ryan, 2006; Crossley and Kennedy, 2002). Differences in capacity to respond to survey questions or in

[1] Although the exact wording and response options of SAH vary across surveys, it is mainly based on individual ratings of current overall health, typically on a five-point ordinal scale (for example, from "excellent" to "poor").

reporting behaviour may depend on respondents' cognitive and non-cognitive skills, which may explain why some individuals tend to provide inconsistent response about their self-assessed health and others do not (Black et al., 2017). Moreover, individuals may respond inconsistently because they assess their health with some uncertainty or they have "learned" more about their health status because of the other questions that are asked between the first and the second SAH questions — for example, specific health or disability questions can influence subsequent responses about people's health status (Black et al., 2017; Clarke and Ryan, 2006; Crossley and Kennedy, 2002).[2] Strategic reporting may be another reason of reporting inconsistency — for example, the broader literature on measurement error in well-being indicators argues that individuals may exhibit a "justification bias" when they overstate their poor health condition in order to rationalize their economic inactivity (e.g., Black, Johnston and Suziedelyte, 2017; Bound, 1991; Kapteyn et al., 2007; Kerkhofs and Lindeboom, 1995). Mode of data collection may be relevant here; some of the existing literature has argued that self-completion, as opposed to open interview, mode may be more reliable in eliciting accurate responses to sensitive questions especially in the presence of other household members (e.g., Conti and Pudney, 2011).

The few studies that test the consistency of responses to SAH questions, and estimate misclassification by assessing repeated SAH questions for the same individuals over a short time frame, mainly use Australian data (Black et al., 2017; Chen et al., 2021; Clarke and Ryan, 2006; Crossley and Kennedy, 2002). These studies often compare responses from SAH questions with different wording and/or SAH measures that are asked within a wider time window (up to 30 days) rather than within the time window allowed at the same household interview (Black et al., 2017; Clarke and Ryan, 2006). For instance, a wider time window between SAH questions may introduce time-varying unobservable influences which may affect actual people's health and/or affect people's responses revising their responses regardless of whether or not their actual health may have changed. Consequently, comparisons of repeated SAH questions collected after a wide time interval may

[2] Individual learning between the two SAH questions may be understood as a priming effect. The psychology literature defines priming effects as implicit memory effects in which prior exposure to a question determines, to some extent, the response to a later question (see Voicu, 2015).

produce misleading results and uncertainty on whether or not this is driven by changes of peoples actual underlying health. In our study, respondents report their SAH using different collection modes in a short time window within the same UKHLS wave. Furthermore, often these studies are based on cross-sectional data and even when longitudinal data on duplicate responses to different SAH questions within each wave is available, they are often based on unequally spaced panels, which may limit analysis of persistent patterns in inconsistency SAH reporting behaviour (Black et al., 2017; Chen et al., 2021; Clarke and Ryan, 2006; Crossley and Kennedy, 2002).

Moreover, none of these studies have used more objectively measured health indicators, such as nurse-administered and blood-based biomarker data, and do not analyse whether individuals' biological health status is reflected differently across the two SAH measured administered using different survey modes (open interview and self-completion). Particularly, the scope of this approach is to validate individual responses to SAH questions from different interview modes by using measures of biomarkers as proxies of people's true underlying health regarding certain dimensions. Biomarkers are objective health measures that not only capture pathogenic processes but also reflect clinical condition and pre-symptomatic perception of individuals (Colburn et al., 2001). Some biomarkers (e.g., obesity) may capture conditions that are more visible than others (e.g., cholesterol), which may be associated with different response for the same SAH question from different modes of collection. For instance, social desirability, i.e., when respondents tend to take social norms into account in the case that interviews involve social interactions that is more likely to happen in the open-interview mode (Bowling, 2005), may be a driving force behind potential differences in the association between more objective measures of health and reporting behaviour to SAH questions by mode of administration if people face a more visible health condition.

Our paper contributes to this literature in a number of ways. We capitalise on the rare opportunity provided by UK Household Longitudinal Study (UKHLS) which asks respondents the same SAH question with identical wording twice (one with a self-completion mode and one with an open interview mode), mainly within the same household interview at UKHLS Waves 2, 3, 4 and 5.[3] Descriptive analysis of the data show considerable inconsistency in reporting of the SAH questions within each wave. We estimate multivariate models to explore the profile of those individuals who reported inconsistently to SAH questions between the open interview and the self-completion mode within each UKHLS wave.

Second, we implement analysis to explore the frequency of reporting inconsistencies. Specifically, we study inconsistent reporting within a time frame that does not justify changes in their actual health status. Although these results do not provide guidance on whether the self-completed or the open interview mode provides more reliable SAH measures, they do provide evidence on whether the observed patterns are systematic and persistent.[4] It should be noted that in this paper we do not aim to disentangle any potential role of true state dependence (if any), i.e., that inconsistent response to the two SAH questions in the past may have a structural impact on the probability of inconsistent SAH responses in the future, from spurious state dependence, which may be attributed to time invariant (or serially correlated) individual-level characteristics, in explaining the observed persistence. Besides, unlike Chen et al. (2021), our paper does not aim to assess which of the two SAH measures (the self-completion versus the open interview) is more accurate or the type and incidence of response errors associated with each of the two measures. The results suggest that reporting inconsistencies are a systematic behaviour that repeats over time for certain population groups. Non-random measurement error in SAH, that is associated with socioeconomic variables,

[3] The self-completion questionnaire, which also contains the relevant SAH question, was available as a paper questionnaire for Wave 2 (while administered using the computer assisted self-interviewing (CASI) survey technique at Waves 3, 4 and 5, in which the respondent uses interviewer's digital device to complete the SAH questionnaire without an interviewer administering it to the respondent). This may indicate that although both the self-completed and open-interview SAH questions are answered within an hour or so in Waves 3-5 (the duration of the household interview), there may be some delay in the completion of the self-reported paper questionnaire as opposed to open interview SAH at Wave 2; however, the instructions to the interviewers indicate that it is expected for the respondents to complete the self-completed Wave 2 interviews whilst the interviewer are in the household (and to give them back to the interviewer) rather than send it over later in time.

[4] Systematic inconsistent reporting indicates that socioeconomic characteristics are statistically significant in predicting reporting inconsistencies in SAH between the open interview and the self-completion mode. Persistence in reporting inconsistency is defined broadly as the continuity of this reporting inconsistency for more than one wave for the same individual.

may contaminate existing research using SAH measures as outcome or explanatory covariate.

Third, the richness of our data allows us to use a detailed set of demographic and socioeconomic characteristics, cognitive and noncognitive skills as well as proxies of the micro-social environment during the household interview, such as the presence of other adults or children. The nature of the questionnaire content within the two SAH questions does not allow us to explore whether any potential "learning" effects, or "justification biases" may explain the observed inconsistencies in SAH within each wave.[5] Instead, our analysis allows us to identify the profile of those who are more likely to report SAH inconsistently as well as those who repeat this inconsistent reporting SAH behaviour most frequently. Investigating the frequency of inconsistent responses help us to characterise the profile of those individuals who reported SAH without consistency more persistently. The latter suggests that reporting inconsistencies are a systematic behaviour that repeated over time for certain population groups and not simply a snapshot of a certain time (when focusing on modelling the probability of reporting SAH inconsistently at a particular wave). Our results show that the same individual profile that predicts the probability of inconsistent response to SAH questions also explain the high frequency of inconsistent responses. The potential implications of this measurement error are relevant when SAH is used as an outcome of interest.

Many datasets collect SAH measures, and researchers may pay limited attention to the collection mode (self-completion as opposed to open interview). SAH data have been routinely collected in many datasets, including the National Health and Nutrition Examination Survey (NHANES) in the US, the Survey of Health, Ageing and Retirement in Europe and the British Household Panel Survey (BHPS) in the UK. For example, SAH measures are collected using an open interview mode in the case of both the NHANES and the BHPS; however, most existing studies do not explicitly consider the potential implications of the SAH collection mode and the

[5] In the face-to-face interview, for instance, qualified individuals for a disability pension may exaggerate their poor health condition in order to justify their health-related work limitation (e.g., Black, Johnston and Suziedelyte, 2017). In the self-completion mode, responses to SAH question from those individuals tend to be more reliable because the more confidential nature of this mode of interview induces more truthful responses about sensitive issues (see List et al., 2004). These may therefore lead to inconsistent responses to SAH questions for these individuals across the different interview modes within the short time interval of the same interview.

survey design for the SAH collection on their analysis (e.g., Contoyannis, Jones and Rice, 2004, Fichera and Gathergood, 2016; Donni, Peragine and Pignataro, 2014). Often, even when multiple measures of SAH questions with the same or similar wording are available in a dataset (at least for some waves, such as in the case of HILDA in Australia, NHANES in the USA, and UKHLS in the UK), researchers do not use all available information and often consider responses to a particular SAH question despite concerns about the consistency of responses within repeated SAH questions (e.g., Au and Johnston, 2014, Davillas et al., 2019, Nesson and Robinson, 2019). If socioeconomic status (SES) plays an important role for reporting of SAH collected with different interview modes (self-completion versus the open interview)[6], this may be a concern for the robustness of the existing studies that use SAH as an outcome and where measurement error is part of the error structure of the SAH regression models.

We provide evidence on whether the SAH interview mode affects results for the income-health gradient — a popular topic in the socioeconomic determinants of health literature, where self-reported health measures are often used as health outcomes (e.g., Davillas et al., 2019; Foverskov and Holm, 2016; Fuchs, 2004; Frijters et al., 2005; Johnston et al., 2009; Larrimore, 2011; Ziebarth, 2010). Separate linear regression models on household income are estimated using the self-completion and the open-interview SAH measures as outcome. The analysis of income-health gradient is an attempt to shed light on the role of the mode of collection. Assuming that the parameter of interest is estimated with measurement error in both specifications, part of the potential difference between gradients should be attributed to the mode of SAH collection. The reason is that unobserved reporting behaviour depend on the interview mode (see List et al., 2004). Despite the implications of measurement error in both linear regression models, the absence of differences suggests that the interview mode plays little role in the income-health gradient.

[6] We should explicitly mention that in this study we cannot disentangle the mode interview effect from inconsistent behaviour of respondents when modelling inconsistent responses to SAH questions. The context of our work does not involve random assignment of survey participants to treatment groups receiving different versions of survey collection modes. However, it does not undermine our analysis as we are interested in investigating which individual's personal characteristics, including cognitive and non-cognitive abilities, and household context predict the differences in responses to SAH questions.

Finally, we use a detailed set of nurse-collected blood-based biomarker data. Unlike the self-reported health measures, biomarkers are more objective health measures and, beyond pathogenic cases, they also provide information on pre-disease stages that may be below clinical diagnosis thresholds. Certain dimensions of physiological and biological health may be reflected more strongly in responses to SAH self-completion questions as opposed to the open interview mode. Individuals without visible health conditions (e.g., obesity) may strategically report their health status in the open interview mode (see Black, Johnston and Suziedelyte, 2017), while the self-completion mode may lead them to provide more reliable or honest responses to sensitive survey questions such as their health status (see List et al., 2004). Moreover, the presence of interviewer in the open interview mode can trigger a general “put on a good show for the visitor” effect, which may lead respondents to overstate their true health status relative to the more private self-completion interview mode (see Conti and Pudney, 2011).

Testing whether the relationship between SAH and biomarkers differs depending on the interview mode contributes to the literature on better understanding SAH as a health outcome (e.g., Au and Johnston, 2014; Jylhä, 2009). A finding that different dimensions of health have differing patterns of association with the two SAH measures may explain why econometric results may differ between the open interview and self-completion SAH measures they are used as outcomes or explanatory variables. On the other hand, if there are no differences, other mechanisms on how individuals translate and report their actual health using SAH measures may be relevant (Jylhä, 2009).

The rest of the paper is organized as follows. Section 2 presents and describes the UKHLS dataset. Section 3 presents our econometric methods. Analysis of the observed inconsistent responses to SAH questions over time and the relevant longitudinal patterns are presented in Section 4. Multivariate analysis of the association between inconsistent reporting and socioeconomic factors as well as an analysis of the potential implications for measurement error in SAH for research on the income-health gradient are presented and discussed in Section 5. Section 5 also contains our analysis on whether physiological and biological health are reflected more strongly in the self-completion or the open interview SAH measures. Section 6 concludes and provides a summary of our findings.

2. The UKHLS dataset

The data come from Understanding Society, the UK Household Longitudinal Study (UKHLS). The UKHLS is a large, nationally representative panel survey, with a design that involves overlapping 2-year waves. Individuals have been interviewed annually since the initial wave in 2009–2010 (Wave 1). The BHPS sub-sample is absorbed into the UKHLS at Wave 2. UKHLS contains a detailed set of demographics, socioeconomic, health and well-being information for all household members on an annual basis.

A feature of this dataset is that information on SAH is collected twice for each respondent at each of the UKHLS Waves 2, 3, 4 and 5, with two modes of collection: open interview and self-completion. Responses on SAH are asked during the open, face-to-face interviews for each household member and, within the same wave, using a self-completion questionnaire. This was available as a paper questionnaire for Wave 2 and using the computer assisted self-interviewing (CASI) survey technique, in which the respondent uses a computer to complete the SAH questionnaire without an interviewer administering it to the respondent[7]. Specifically, the following SAH question is asked twice, within each wave and in the time frame of the UKHLS household interview: "In general, would you say your health is: Excellent, Very Good, Good, Fair or Poor?". Wording of the question and ordering of the health categories are identical between both measures within and between the UKHLS Waves 2-5. We have coded SAH so that higher values indicate a better health state.

To explore those factors that are associated with probability of reporting inconsistently between the open interview versus self-completion questions, within each wave, a set of longitudinally collected (unless otherwise stated) explanatory variables are used. These follow the related literature (Black et al., 2017; Chen et al., 2021; Crossley and Kennedy, 2002; Clarke and Ryan, 2006). We account for demographic and socioeconomic characteristics, measures of cognitive ability, personality traits, and indicators for the presence of other household members during the interview are included in our model specifications.

[7] To account for this difference in the survey design across waves, wave fixed effects are used in our regression models.

Specifically, gender and age group dummies for five-year intervals between 16 and 85 and a dummy for those over 85 years old are included in our analysis; this allows us to flexibly capture the role of age as well as gender on reporting inconsistency. Existing literature has shown that age and gender are systematic sources of reporting heterogeneity in SAH (Zajacova, Huzurbazar and Tood, 2017)). For instance, older individuals are more likely to assess their health better than younger individuals, equivalently healthy counterparts, while females assess their health lower than equivalently healthy men (e.g., Nesson and Robinson, 2019). Because older individuals expect to face health problems, younger individuals tend to interpret the information about their own health differently (see Jylhä, 2009). Nonetheless, it has been shown that males are more likely to inconsistently report their SAH when compared to women (see Black et al., 2017; Clark and Ryan, 2006; Crossley and Kennedy, 2002).

Two measures are used to capture the socioeconomic status of the respondents: highest educational attainment (degree, other higher qualification, A-level, GCSE, other lower qualification, no qualification) and household income (equivalized using the modified OECD scale and deflated). Recent evidence has shown that low-educated individuals are more inclined to provide inconsistent responses about their SAH (e.g., Chen et al. 2021), while high-educated individuals are more likely to show consistent reporting behaviour (e.g., Black et al., 2017). Better education is associated with better health literacy and greater control of people's own health status (Nutbeam, 2008), which may improve consistency to responses to SAH questions as people may be more certain about their health state. Existing literature has also documented a negative association between inconsistent SAH reporting behaviour and income (see Chen et al. 2021). Richer individuals are more likely to access preventive health care than poorer ones (e.g., Cookson et al., 2016), which may help individuals to reduce uncertainty about their true health status and, thus, it may improve consistent reporting to SAH questions

Employment status is captured by a four-category categorical variable (employed, unemployed, retired, and other job status). the existing literature is not conclusive about the relationship between labour market status and consistent reporting behaviour on SAH. While some evidence shows that unemployed individuals and those out of labour force are more likely to provide inconsistent responses (see Clarke and Ryan, 2006; Crossley and Kennedy, 2002), more recent studies do not find systematic associations (see Black et al., 2017; Chen et al., 2021).

Completing the vector of socioeconomic characteristics, a four-category variable is used to account for marital status (married, single, separated/divorced, and widowed). For example, married individuals may have some incentives to assess their health more positively in the open interview and in front of the partner rather than in the case of typically more honest self-completion mode. Existing literature from Australia provides mixed results about the association between marital status and inconsistent response to SAH questions (Black et al., 2017; Chen et al., 2021), highlighting some space for additional empirical evidence in the case of the UK.

Cognitive ability is considered as an important determinant of reporting behaviour (e.g., Black et al., 2017). Questionnaires place cognitive demands to respondents and, thus, it is likely that respondent's cognitive ability will affect the consistency of responses. Moreover, as cognitive ability is associated with a number of labour market outcomes (Lin et al., 2018), establishing the presence of systematic error in SAH that is relevant to cognitive ability may be of particular relevance to research exploring the effect of self-reported health measures on labour market outcomes. Recent evidence using Australian data shows that the probability of consistent SAH responses increases with the quantiles of the cognitive ability index that accounts for memory, cognitive function, and verbal skills (see Balck et al., 2017). Complementing the existing literature, a wide set of cognitive ability measures are employed in our study to explore the conditional association of each of these measures, capturing different aspects of people's cognitive ability, on the probability of inconsistent response to SAH questions. A large set of cognitive ability measures are collected at UKHLS wave 3; as no repeated data are available for waves 2, 4 and 5 we have to assume that cognitive ability remains fixed within the relatively short time interval between waves 2 and 5. The literature suggests that cognitive ability may be fairly stable with age (Lyons et al., 2017) and, thus, our results on the role of cognitive ability on reporting behaviour may not be contaminated by the absence of longitudinal data on cognitive ability.

We control for the following cognitive ability measures: episodic memory, i.e., the number of words the respondent can recall from a carefully recorded list of ten words (two variables for immediate, and delayed word recall); working memory, measured by counting the correct answer to a series of five (simple) numerical subtraction questions; semantic or category fluency, measured by counting the number of correct and incorrect responses to naming as many animals as the respondent can in 60 minutes; practical numerical knowledge, measured by counting the number of correct answers to five questions.

Personality traits are also collected at UKHLS Wave 3 using a 15-item questionnaire version of the Big-Five Inventory (John et al.,1991). Responses to a set of three questions (from the total of 15 questions) pertaining to each trait are then used to calculate each of the five-personality trait scores: agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience. Higher scores indicate that the particular trait is more relevant to the respondent. The "Big-Five" personality traits have been used extensively in the economics literature and are viewed as a stable input in regression models; they are characterized by a limited variability over time, with any potential intra-individual personality change being mostly unrelated to adverse life events (Cobb-Clark and Schurer, 2012). Using Australian data, it has been shown that conscientious respondents are less likely to provide inconsistent response to SAH questions (see Black et al., 2017); the hypothesis behind this association is that individual effort and consideration about answering the survey questions may varies according to certain personality traits (see Bertrand and Mullainathan, 2001). We complement the existing literature by providing evidence of the conditional association of all "Big-Five" personality traits,i.e., agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience, on reporting behaviour to SAH question in the context of the UK population.

Finally, to account for the role of the micro-social environment during the open interview (which may affect reporting behaviour at the open interview but to lesser extent for self-completion), we have included two dummy variables for the presence of other adults and children (aged 10-15) during the household interviews. Existing research has shown that social desirability bias, which is especially relevant to the presence of other household members during the open interview, may affect reporting behaviour in well-being outcomes (Conti and Pudney, 2011). Regional dummies are included to account for regional variations in health.

Longitudinal information from these explanatory variables (where available) are used for our models exploring the factors associated with the likelihood of inconsistent responses at each wave. However, in our subsequent analysis of the frequency of inconsistent reporting (i.e., the total number of inconsistent responses across the four waves), a snapshot of the explanatory variables is employed (mainly from Wave 2; and Wave 3 for the personality and cognitive ability measures).

Nurse administered and blood-based biomarker data

Nurse-collected and blood-based biomarkers are collected by trained nurses as part of their visits following the UKHLS Wave 2 and Wave 3 main waves. Wave 2 nurse visits collect data from the original UKHLS sample, while Wave 3 collects data from the BHPS sample that was absorbed into the UKHLS. We use a pooled sample of Waves 2 and 3 in our analysis to explore whether these objectively measured biomarkers have different patterns of association with our two SAH measures. Following the literature, we use a range of biomarkers (e.g., Davillas and Jones, 2020; Davillas and Pudney, 2020). We use the waist-to-height ratio (WHR) to measure adiposity. Resting heart rate and blood pressure are measured followed standard measurement protocols. Systolic blood pressure, the maximum pressure in an artery when the heart is pumping blood, diastolic blood pressure, the lowest pressure when the heart is resting, and the pulse rate are used as continuous variables; higher values indicate higher cardiovascular risks. We use a set of blood-based biomarkers relevant to inflammation, steroid hormones, fat in the blood, blood sugar and liver functioning. C-reactive protein (CRP) is our biomarker for systemic inflammation, which rises as part of the immune response to infection. [8] The dihydroepiandrosterone suphate (DHEAS) is the most common steroid hormone in the body — a primary mechanism through which psychosocial stressors may affect people's health. Low levels of DHEAS are associated with cardiovascular and all-cause mortality risks (Ohlsson et al., 2010). The "good" cholesterol, high-density lipoprotein cholesterol (HDL), is used as our fat in the blood biomarker; lower HDL levels are associated with increased cardiovascular risks. HbA1c is a biomarker that measures blood sugar, regarded as a diagnostic test for diabetes. As a liver function test, we use albumin, the main liver protein; lower albumin levels suggest impaired liver function (Davillas and Pudney, 2020). In addition to specific markers, an index of multi-system risk that measures the wear and tear on the body, approximating

[8] We follow the conventional practice and exclude those with CRP over 10 mg/L, as those values may reflect current transient infections and not chronic processes (Davillas and Pudney, 2020; Pearson et al., 2003).

the allostatic load, is also employed. Following exiting literature (Davillas and Pudney, 2020), HDL, Albumin and DHEAS are transformed to negative values to reflect ill health, and then each of the measures described above is converted into *z*-scores and summed. To facilitate comparisons all our biomarkers and allostatic load are transformed to reflect derivations from their standard deviation.

3. Methods

Panel probit models for the likelihood of reporting inconsistency

We investigate the determinants of within-wave inconsistent SAH responses by modelling whether the responses to SAH question differ between the self-completion (H_{sc}) and the open interview (H_{OI}) , administered within the same wave for all participants. Our dependent variable, y_{it}, is defined as a dichotomous variable that, for each respondent i in UKHLS wave t (2, 3, 4 and 5), takes the value of 1 if they reported differently in the H_{sc} versus the H_{OI} , and 0 otherwise.

We estimate pooled probit models. The likelihood for this model corresponds to assuming independence in the error terms across time, however the maximum likelihood estimator of the model has the quasi-maximum likelihood property and is robust to arbitrary serial correlation (e.g., Wooldridge, 2002). The probability of reporting inconsistently is given by:

$$\Pr(y_{it} = 1) = \Phi(x_{it}\beta) \tag{1}$$

where, the underlying latent variable model is given by $y_{it}^{*} = x_{it}\beta + e_{it}$, with $y_{it} = 1[y_{it}^{*} > 0]$. The vector x_{it} is the set of covariates used in our analysis, β are the respective coefficients to be estimated, and e_{it} is the independently and normally distributed error term. The term $\Phi(\cdot)$ is the cumulative distribution function of the normal standard distribution.

Our second specification is the random effects probit model which assumes an error components specification. This allows the error term to be decomposed into permanent and transitory components, but it is not robust to misspecification of the serial correlation. A random effect term (c_i) is included as part of the error structure, i.e., $u_{it} = c_i + e_{it}$. A random effects (RE) probit model can be the estimated as:

$$\Pr(y_{it} = 1|x_{it}, c_i) = \Phi(x_{it}\gamma_c + c_i). \quad (2)$$

where the underlying latent variable model is given by $y_{it}^* = x_{it}\gamma_c + u_{it}$, with $y_{it} = 1[y_{it}^* > 0]$. The terms c_i and e_{it} are assumed to be normally distributed and independent of x_{it} and of each other. Conventional maximum likelihood estimation (MLE) methods, alongside the Gaussian quadrature procedure, can be then used for the consistent estimation of the explanatory variable coefficients γ_c and the variance of unobserved heterogeneity σ_c^2 (Butler and Moffitt, 1982; Greene, 2003; Wooldridge, 2002).[9] The Average marginal effects are estimated for the pooled and random effects models, with standard errors estimated using the delta method (Wooldridge, 2005).[10]

Fractional response model for the frequency of the inconsistent reporting

We analyse the profile of those who reporting inconsistently more frequently than others over the four waves available. The frequency of inconsistent SAH responses is defined by the number of waves in which the (H_{SC}) differs from (H_{OI}). Our dependent variable is defined as the fraction of waves with inconsistent responses to SAH questions relative to total number of waves available in our dataset (4 waves); values of the resulting dependent variable lie in the interval [0,1]. This outcome variable is modelled using a cross sectional fractional response model using explanatory variables from baseline (Wave 2 or Wave 3 for the cognitive and non-cognitive measures).

The fractional response model is fitted by a quasi-maximum likelihood estimator, which does not require the true distribution of the entire model to obtain consistent parameter estimates for the conditional mean (Papke and Wooldridge, 1996). The conditional mean of the outcome variable can be expressed as:

$$E(y_i|x_i) = G(x_i\beta) \quad (3)$$

where, $0 \leq y_i \leq 1$, and $G(\cdot)$ is a known function satisfying $0 < G(z) < 1$ for all $z \in \mathbb{R}$. The function $G(\cdot)$ is specified as the normal standard cumulative distribution

[9] The marginal effect for the random effect probit model takes into account the variance of unobserved heterogeneity, i.e., $\gamma_c = \gamma/(1 + \sigma_c^2)^{1/2}$ (Wooldridge, 2002).

[10] We have calculated the marginal effect for each explanatory variable keeping all other variables in their mean values.

function. The quasi-maximum likelihood estimator can be implemented by using a Bernoulli log-likelihood function (Papke and Wooldridge, 1996). Marginal effects are also estimated to facilitate quantitative interpretation of our results[11]

Measurement of SAH and the income-health gradient

Among the covariates used to explore inconsistent reporting, we find that household income has a strong association. Given the large literature on the health-income gradient, this implies that differences in the interview mode may matter for studies of the income-health gradient that rely on SAH measures. Using all the available UKHLS waves (Waves 2-5) we estimate pooled OLS regression models for each mode of SAH on household income, adjusted for individual's age, gender and wave fixed effects. SAH is coded so that higher values indicate a better heath state. These models, although parsimonious, allow us to explore potential differences in the association between income and health that may be attributed to SAH measurement.

The association between biomarkers and SAH

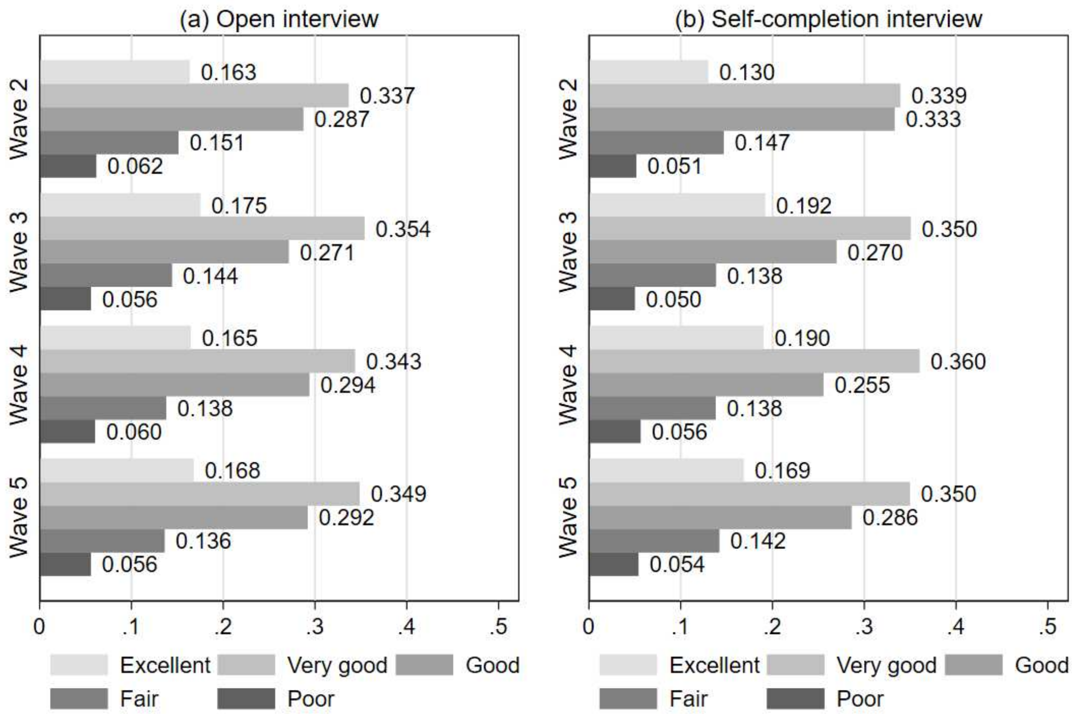
Given the cross-sectional format of our biomarker data, linear regression models of SAH on each of our biomarkers (and for allostatic load) are estimated for the pooled Waves 2 and 3 sample; each of these models also accounts for age, gender, regional dummies and wave fixed effects. As the biomarkers are measured health indicators, comparisons of the magnitude of their associations with our two SAH measures (open interview versus the self-completion mode) may give information on whether particular dimensions of physiological and biological health are reflected more strongly in one or the other SAH measure.

[11] Marginal effects can be interpreted as percentage points change in the frequency of within-wave inconsistent SAH reporting across all four waves due to a unit variation in the explanatory variable of interest.

4. Descriptive analysis of inconsistent responses

Figure 1 presents the distribution of responses to SAH questions for the open and the self-completion questions separately for all waves (Waves 2-5) where both of these measures are collected for each individual. These show that the *overall* marginal distributions of SAH are very similar for the two modes of data collection, especially at Wave 5. However, the two modes do not give identical distributions, and, in the earlier waves, more respondents report better health in the open interview mode. Overall, these preliminary results give us limited information about the inter-individual differences in SAH reporting between the open interview and the self-completion and only indicate the presence of moderate differences in the *overall* distribution of SAH categories.

Figure 1. Histogram of SAH responses (self-completion and open interview) by waves: unbalanced sample (obs. 161,242)



The preliminary analysis in Figure 1 is based on an unbalanced sample. However, for our analysis we need to follow the same individuals across the four waves and ensure the presence of valid responses for both SAH questions for each individual within and across waves as well as non-missing information on all explanatory variables used in our analysis — thus, a balanced sample is used for the remaining analysis. Figure 2 shows that the frequencies of the different SAH categories (for both the self-completion and the open interview question) are almost identical between the unbalanced and balanced samples; this suggests that the implications of these exclusion restrictions should be very limited for our analysis.

Figure 2. Histogram for SAH responses (self-completion and open interview): unbalanced sample (obs. 161,242), balanced sample (obs. 97,456) and balanced sample with no missing data on all variables used in the analysis (obs. 90,600).

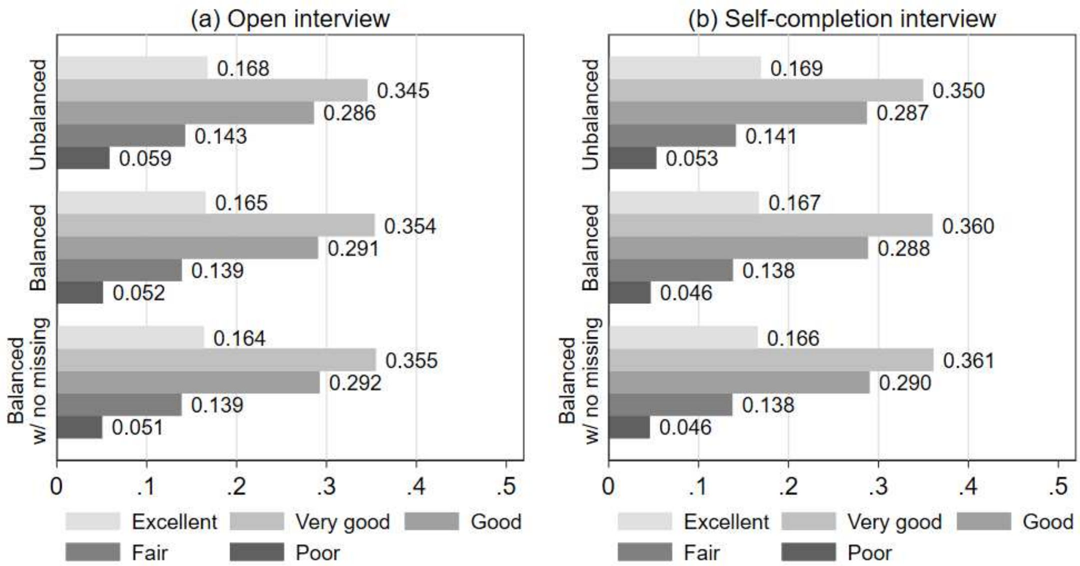
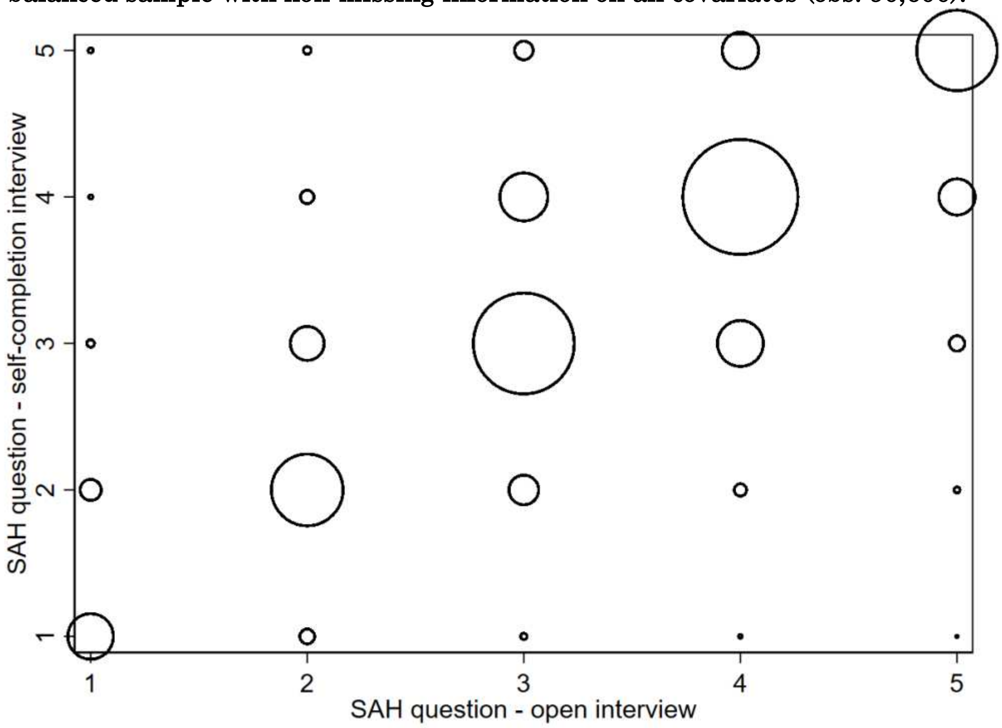


Figure 3 shows a bubble plot of responses to self-completion as opposed to open interview, using the complete cases balanced sample. This illustrates the extent and pattern of *inter-individual differences* in reporting between the two SAH measures. Although we observe that there is a concordance in responses to both SAH responses for most of the respondents (along the main diagonal), there is a sizeable proportion of respondents (as show by the size of the bubbles above and below the main diagonal) that reported their SAH inconsistently within the same interview. Table 1 presents the corresponding proportions of reporting a particular SAH status in the open interview mode conditional on responses to the self-completion SAH questions. For example, for those who reported excellent health in the self-completion interview, about 80% reported excellent health in the open interview

SAH question, while the remaining respondents are distributed across all other categories (with 15% of those reporting the nearest possible category — very good health). Overall, our results show that 11%-24% of those who reported a given SAH category in self-completion mode, reported inconsistently in the open interview, with the majority of the inconsistent responses concentrated in the SAH categories that are adjacent to their self-completion responses.

Figure 3. Bubble plot — SAH self-completion versus open interview: balanced sample with non-missing information on all covariates (obs. 90,600).



Note: Each bubble is weighted by the number of respondents. Higher self-assessed health (SAH) values indicate a better health state: "1" stands for Poor, "2" Fair, "3" Good, "4" Very good and "5" Excellent.

Table 1. Distribution of SAH responses in self-completion versus the open interview questionnaire.

Self-completion interview	Open interview					
	Excellent	Very good	Good	Fair	Poor	Total
Excellent	12,015 (80.12%)	2,283 (15.22%)	578 (3.85%)	85 (0.57%)	35 (0.23%)	**14,996 (100%)**
Very good	2,365 (7.22%)	25,738 (78.63%)	4,321 (13.20%)	293 (0.90%)	17 (0.05%)	**32,734 (100%)**
Good	396 (1.51%)	3,849 (14.65%)	19,926 (75.83%)	2,026 (7.71%)	80 (0.30%)	**26,277 (100%)**
Fair	46 (0.37%)	261 (2.09%)	1,610 (12.92%)	9,802 (78.64%)	745 (5.98%)	**12,464 (100%)**
Poor	7 (0.17%)	14 (0.34%)	51 (1.24%)	354 (8.57%)	3,703 (89.68%)	**4,129 (100%)**
Total	**14,829**	**32,145**	**26,486**	**12,560**	**4,580**	**90,600**

Capitalising on the longitudinal nature of our data, we explore the unconditional dynamics of the observed within-wave inconsistencies. Following Black et al. (2017) and Chen et al. (2021), Table A1 (Appendix) provides shares of consistent and inconsistent responses to SAH questions from open and self-completion modes using the balanced sample of waves 2-5. Table A1 shows that from those reporting SAH inconsistently (SAH-OI > SAH-SC or SAH-OI < SAH-SC) between the open interview and the self-completion SAH measure, there is no consistent pattern of reporting better health in the SAH open interview as opposed to self-completion (and vice versa) across waves; the fraction of our sample reporting SAH consistently (SAH-OI = SAH-SC) over time increases from 76% to 81% when moving from wave 2 to wave 5 for the same individuals (balanced sample). However, we refrain from modelling these patterns explicitly in our paper and we focus part of our analysis on a dichotomous variable on consistent reporting in SAH questions or not for reasons we describe below. We believe that these comparisons (SAH-OI > SAH-SC and SAH-OI < SAH-SC) are more affected by the ceiling (or floor) effects.[12]

Table 2 presents summary statistics for all the observed sequences of reporting inconsistency/consistency. These sequences cover four waves, resulting in 16 (= 2^4) distinct sequences. We assign a value of 1 to respondents with inconsistent SAH responses between the self-completion and open interview measures at each specific wave, and 0 otherwise. Table 2 shows that only 42.6% of our sample are classified as "always consistent" (0000). Turning to the remaining sequences, the results show that more than half of our sample (57.4%) is formed by respondents who have provided inconsistent responses about their SAH at least once. Specifically, grouping the remaining sequences with respect to frequency or consistent/inconsistent responses, those who are "mostly consistent" (0001, 0010, 0100 and 1000), i.e., within-wave inconsistent SAH reporting only once among the four waves, represent 35.5% of our total sample. Sequences which could be grouped as "moderately inconsistent" (1001, 1010, 1100, 0011, 0101, 0110), accounting for less than 1/3 of the total sample (16.4%). The proportion of the "mostly inconsistent"

[12] A potential issue related to responses to SAH questions is the ceiling (or floor) effects, where individuals reporting the highest (or lowest) level of SRH cannot report subsequent improved (or worsened) health (see Gunasekara et al., 2012; Lumsdaine and Exterkate, 2013). As a way to reduce sensitivity to ceiling or floor effects, we measure inconsistency by indicating those respondents that provide diverging answers in open and self-completing modes of interview.

sequences, i.e., those who are reporting SAH within-wave inconsistently in three out of four waves ("0111", "1011", "1101", "1110") accounts for the 4.8% of our sample. Only 0.75% are "always inconsistent" (1111).[13]

Table 2: Distribution of inconsistent/consistent responses to SAH: balanced sample

Sequences	Frequency	Percent	Cumulative
0000	38,568	42.57	42.57
0001	6,472	7.14	49.71
0010	7,276	8.03	57.74
0100	8,712	9.62	67.36
1000	9,660	10.66	78.02
1001	2,340	2.58	80.6
1010	2,488	2.75	83.35
1100	2,964	3.27	86.62
0011	2,460	2.72	89.34
0101	2,104	2.32	91.66
0110	2,512	2.77	94.43
0111	1,124	1.24	95.67
1011	1,012	1.12	96.79
1101	1,036	1.14	97.93
1110	1,196	1.32	99.25
1111	676	0.75	100
Total	**90,600**	**100**	-

Notes: "0" stands for consistent SAH responses between the self-completion and open interview SAH measures for each particular wave; "1" for inconsistent SAH responses.

Table 3 provides some preliminary evidence on whether the inconsistent reporting patterns over time are systematically associated with individual characteristics. For example, Table 3 shows that males are more prevalent in the "mostly inconsistent" and "always inconsistent" groups. Mean values for the age groups show heterogeneous patterns across the different sub-groups. Turning to education, lower educational categories are more prevalent for the "moderately

[13] One may argue that the fact that only 0.75% of our sample members belongs to the always inconsistent ("1111") category indicates that persistence is low for inconsistency responses. However, this is not true. Given that the mean probability of reporting inconsistently is 0.214, under independence, this means that we would expect that the proportion of our sample "always inconsistent" ("1111") should be 0.2144=0.002 (or 0.21%). By contrast, in our sample, we observe 676 individuals (i.e., 0.75% of our sample) classified in the "always inconsistent" category.

inconsistent", "mostly inconsistent" and "always inconsistent" groups as opposed to the "always consistent" or "mostly consistent" groups (as well as for the total sample). The average log of income for the total sample is about 7.2 and decreases as the number of waves with inconsistent SAH responses increases; the difference in household income between the "always consistent" and "always inconsistent" categories is approximately 25%. These results suggest that socioeconomic status (proxied by education and household income) is relevant for reporting behaviour in SAH. Regarding job status, our unconditional summary statistics show that the mean unemployment prevalence increases from 3.1% in the case of the "always consistent" sub-sample to 8.1% for the "always inconsistent" sub-sample. Turning to marital status, the unconditional summary statistics show that while the proportion of married respondents is decreasing with the increasing number of waves that individuals reported SAH inconsistently within each wave, the proportion of single respondents is increasing.

Table 3 also shows that the more consistent sub-samples have higher cognitive ability scores for the set of cognitive ability measures used in our analysis. Turning to the Big 5 personality traits, the mean values for agreeableness, conscientiousness and extraversion are higher for the always inconsistent sub-group as opposed to the whole sample and the always consistent sub-group. Concerning the micro-social environment, the unconditional mean for the presence of children (ages 10-15) during the household interview progressively increased as moving from the "always consistent" to "always inconsistent" sub-samples.

Table 3. Sample means of explanatory variables: balanced samples

Variables	Total	# of waves: inconsistent SAH measures 0 waves (always consistent)	1 wave (mostly consistent)	2 waves (moderately inconsistent)	3 waves (mostly inconsistent)	4 waves (always inconsistent)
Male†	0.427	0.402	0.434	0.458	0.488	0.497
Aged 16-25†	0.081	0.074	0.084	0.091	0.089	0.084
Aged 26-35†	0.144	0.154	0.143	0.125	0.130	0.127
Aged 36-45†	0.202	0.208	0.199	0.196	0.192	0.212
Aged 46-55†	0.199	0.197	0.191	0.208	0.228	0.260
Aged 56-65†	0.177	0.184	0.178	0.172	0.145	0.112
Aged 66-75†	0.135	0.131	0.137	0.141	0.142	0.130
Aged 76-85†	0.054	0.046	0.061	0.056	0.066	0.058
Aged 86 and older†	0.008	0.007	0.008	0.012	0.010	0.016
Degree†	0.254	0.300	0.239	0.192	0.167	0.216
Other higher qualification†	0.129	0.134	0.131	0.120	0.114	0.099
A-level†	0.203	0.202	0.203	0.208	0.201	0.186
GCSE†	0.209	0.189	0.213	0.238	0.240	0.275
Other low qualification†	0.097	0.085	0.101	0.115	0.119	0.101
No qualification†	0.108	0.091	0.112	0.126	0.159	0.123
Log of HH income	7.157	7.200	7.139	7.114	7.070	6.980
Employed†	0.589	0.594	0.589	0.581	0.582	0.559
Unemployed†	0.038	0.031	0.039	0.049	0.056	0.081
Retired†	0.248	0.244	0.252	0.251	0.245	0.241
Other job status†	0.124	0.130	0.121	0.118	0.118	0.118
Married†	0.687	0.702	0.679	0.672	0.671	0.629
Single†	0.162	0.153	0.166	0.174	0.168	0.209
Separated/divorced†	0.091	0.091	0.092	0.092	0.095	0.092
Widowed†	0.060	0.054	0.064	0.062	0.066	0.071
Immediate word recall	6.447	6.632	6.397	6.228	6.020	5.870
Delayed word recall	5.425	5.626	5.380	5.160	4.974	4.828
Number of correct subtractions	4.507	4.592	4.482	4.413	4.282	4.325
Verbal fluency: correct words	22.567	23.230	22.299	21.737	21.601	22.000
Verbal fluency: incorrect words	0.329	0.300	0.334	0.371	0.414	0.325
Numeric ability: correct answers	3.727	3.851	3.695	3.561	3.478	3.515
Agreeableness	5.638	5.637	5.649	5.633	5.571	5.781
Conscientiousness	5.505	5.520	5.498	5.489	5.474	5.568
Extraversion	4.583	4.562	4.604	4.588	4.594	4.604
Neuroticism	3.546	3.591	3.540	3.489	3.401	3.462
Openness to experience	4.545	4.558	4.546	4.534	4.466	4.533
Multiple adult interviews/HH†	0.720	0.728	0.712	0.715	0.720	0.753
Children (aged 10-15) interviewed/HH†	0.149	0.145	0.147	0.158	0.178	0.183
Sample size	**90,600**	**38,568**	**32,120**	**14,868**	**4,368**	**676**

Notes: † Dichotomous variables.

5. Multivariate Models

5.1 The likelihood of an inconsistent response

Table A2 (Appendix) presents the estimated coefficients from the pooled and random effect (RE) probit models for the probability of reporting SAH inconsistently in the self-completion and open interview within each wave. Overall, the coefficients from both models are very similar in terms of the direction of the associations and statistical significance[14]. The RE probit model imposes an error component structure on the data, decomposing the error term and modelling the time-invariant unobserved heterogeneity. Under these assumptions, the intra-class correlation coefficient (rho) shows that about 12% of the unexplained variation in inconsistent SAH reporting behaviour is attributable to the individual effect — this is relatively low in magnitude, albeit highly statistically significant. This indicates that after controlling for our set of observed characteristics, there are still unexplained differences in the probability of reporting SAH inconsistently with only a low proportion of this variation (12%) attributed to the time invariant error component in a random effects model.

Table 4 presents the marginal effects, providing an indication of the magnitude of the association between our explanatory variables and the probability of reporting inconsistency. In general, the marginal effects are similar for the pooled and RE probit models; this may reflect the fairly small role of individual effects in explaining the variability of inconsistent SAH responses, which is used to scale the corresponding marginal effects for the RE model. On average men are 0.026 more likely to report inconsistent SAH than females. Those belonging to the 26-35, 36-45 and 56-65 age groups have a lower probability on average (between 0.017 and 0.021) of reporting SAH inconsistently compared to our youngest age group (16-25, reference category). We also find systematic education and income gradients in the probability of reporting SAH inconsistently. For example, the probability of inconsistent responses for those with no qualification is 0.035 higher compared to those with a university degree (reference group); the corresponding probability for those with O-level or GCSE is 0.041 higher compared to the reference group. This indicates the presence of a non-monotonic but positive association between the probability of reporting inconsistent SAH and the lower level of

[14] It should be noted here that the scaling of coefficients in the pooled and RE specifications are different and they should not be compared directly in Table A2 (Appendix); this highlights the need to calculate the corresponding marginal effects which are on the same scale and can be compared.

educational attainment as opposed the reference category (degree). Moreover, there is systematic and positive association with income, indicating that higher income is associated with a lower probability of reporting SAH inconsistently. These results are broadly consistent with existing evidence that measures of socioeconomic status are important determinants of reporting behaviour in SAH and other health outcomes (e.g., Bago d'Uva, O'Donnell and van Doorslaer, 2008; Black et al., 2017; Crossley and Kennedy, 2002; Clarke and Ryan, 2006; Etilé and Milcent, 2006; Johnston et al., 2009). Regarding respondents' job status, those who are retired and with other job statuses experience on average a higher probability of reporting inconsistently, compared to those employed, of about 0.027.

Our measures of cognitive ability are systematically associated with the probability of reporting inconsistently; this is broadly in line with relevant research using data from Australia (Black et al., 2017). Our detailed set of cognitive ability measures gives us the opportunity to further explore which aspects of cognitive ability are more strongly related to consistent reporting behaviours in SAH. To facilitate comparisons across the different cognitive ability variables, the estimates of the relevant marginal effects are scaled in terms of standard deviations of the variables. Overall, word recall (particularly immediate word recall), number of correct subtractions and the number of correct answers in some simple numeric ability tests are more strongly associated with the probability or reporting inconsistently. Given that word recall is related to episodic memory, i.e., memory associated with a specific event or episode, and numeracy is a measure of practical numerical knowledge, we argue that the role of these cognitive skills is much more pronounced on explaining inconsistent behaviours in reporting SAH than the associations for verbal fluency. For example, our marginal effects show that one standard deviation increase in numeric ability is associated with a lower probability of reporting inconsistently of about 0.019; the corresponding marginal effects for verbal fluency (correct words) indicates that one standard deviation increase in the measure is associated with 0.007 reduction in the probability of reporting inconsistently.

We find that the probability of reporting inconsistently is associated with conscientiousness, neuroticism and openness to experience. Higher conscientiousness scores, usually characterizing those individuals with a higher level of self-discipline, are associated with a systematically lower probability of reporting SAH inconsistently. Openness to experience is positively associated with reporting inconsistency. There are arguments that individuals more open to experiences may experience an increased probability of conflicting appraisals (Barford and Smillie, 2016); the later may explain our findings that openness to experience is associated with individuals' tendency to report SAH inconsistently in a short time period. Higher neuroticism scores are associated with a lower probability of reporting inconsistently. Some studies have argued that neuroticism is associated with medical conditions, negatively perceived health status and frequency of visits to the GP (Jerram and Coleman, 1999; Nouri et al., 2019); those with higher scores are then more likely to have a more concrete perception of their SAH and, thus, more likely to report their SAH status more consistently.

Table 4. Pooled probit model and random effects (RE) probit model for inconsistent responses to SAH questions: marginal effects

	Pooled probit		RE probit model	
Covariates	Marginal Effects	SE	Marginal effects	SE
Male	0.0255***	0.0034	0.0259***	0.0034
Aged 26-35	-0.0210***	0.0067	-0.0208***	0.0067
Aged 36-45	-0.0173**	0.0069	-0.0173**	0.0069
Aged 46-55	-0.0019	0.0071	-0.0022	0.0071
Aged 56-65	-0.0178**	0.0074	-0.0173**	0.0074
Aged 66-75	-0.0056	0.0093	-0.0061	0.0092
Aged 76-85	-0.0055	0.0110	-0.0059	0.0110
Aged 86 and older	0.0006	0.0189	-0.0021	0.0187
Other higher qualification	0.0195***	0.0055	0.0198***	0.0055
A-level	0.0244***	0.0049	0.0248***	0.0049
GCSE	0.0405***	0.0051	0.0412***	0.0051
Other low qualification	0.0369***	0.0065	0.0379***	0.0065
No qualification	0.0346***	0.0070	0.0351***	0.0069
Log of the HH income	-0.0135***	0.0035	-0.0122***	0.0035
Unemployed	0.0026	0.0077	-0.0001	0.0076
Retired	-0.0275***	0.0057	-0.0274***	0.0057
Other job status	-0.0271***	0.0047	-0.0252***	0.0047
Single	-0.0009	0.0052	-0.0010	0.0052
Separated/divorced	-0.0045	0.0058	-0.0045	0.0058
Widowed	-0.0001	0.0075	0.0009	0.0075
Immediate word recall†	-0.0134***	0.0025	-0.0134***	0.0025
Delayed word recall†	-0.0055**	0.0024	-0.0054**	0.0024
Number of correct subtractions†	-0.0088***	0.0018	-0.0088***	0.0018
Verbal fluency: correct words†	-0.0066***	0.0019	-0.0069***	0.0019
Verbal fluency: incorrect words†	0.0041***	0.0014	0.0041***	0.0015
Numeric ability: correct answers†	-0.0186***	0.0020	-0.0188***	0.0020
Agreeableness†	-0.0028*	0.0017	-0.0028*	0.0017
Conscientiousness†	-0.0039**	0.0017	-0.0039**	0.0017
Extraversion†	0.0018	0.0016	0.0019	0.0016
Neuroticism†	-0.0075***	0.0016	-0.0075***	0.0016
Openness to experience†	0.0059***	0.0017	0.0060***	0.0017
Multiple adult interviews/HH	-0.0035	0.0041	-0.0037	0.0041
Children (aged 10-15) interviewed/HH	0.0093**	0.0045	0.0079*	0.0045
Wave 3	-0.0107***	0.0036	-0.0108***	0.0036
Wave 4	-0.0275***	0.0036	-0.0278***	0.0036
Wave 5	-0.0441***	0.0035	-0.0444***	0.0035
Sample size	90,600		90,600	

Note: Marginal effects are estimated at sample means. Reference categories for the categorical variables included in our model: age groups (age 16-25), gender (female), education (Degree), marital status (married/cohabiting), job status (employed), and regional dummies (North East).

† Expressed in terms of deviations from their standard deviation.

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

Presence of other adults or children during the open interview seems to have a limited association with reporting inconsistencies, this suggests a limited role for the so-called micro-social environment during the open interview (Conti and Pudney, 2011) on affecting reporting behaviour in the context of SAH. Finally, our results show that the probability of reporting inconsistent SAH is monotonically decreasing across waves compared to Wave 2 (reference). This may reflect utilization of the computer-based self-completion SAH questionnaires (to replace the paper questionnaire at Wave 2, as discussed earlier) and the potential role of a "learning-by-doing" or "priming" process among respondents in answering SAH questions twice within each wave. As the panel become more mature in time, respondents improved understanding of the questionnaire content and/or became more confident with interviewers or with the study as a whole. Respondents may also learn how to strategically answer survey questionnaire with the objective of reducing the interview length (see Fisher, 2019). This may lead respondents to more consistent within-wave SAH responses as the panel ages.

5.2 The frequency of inconsistent responses

Our analysis so far has explored those factors that are associated with the probability of reporting inconsistently in the two SAH questions (open interview versus self-completion) within each wave. In this sub-section we characterize the profile, based on individuals' characteristics at baseline (mainly Wave 2 or Wave 3), of those reporting inconsistently more frequently than others across our four UKHLS waves. Our dependent variable is, thus, defined as the fraction of waves with inconsistent responses to SAH questions relative to total number of waves. Table 5 displays the marginal effects from the corresponding fractional response model; the model coefficients are presented in Table A3 (Appendix).

Overall, the results from the fractional response model show that those baseline characteristics that predict the probability of reporting inconsistently to SAH questions are also systematically associated with a higher frequency of reporting SAH inconsistently, i.e., the presence of "persistent" behaviour in reporting inconsistency to SAH questions across the four waves. Demographics (age and gender), education, income, employment status, cognitive and non-cognitive skills are the baseline characteristics that are most associated with a higher frequency of reporting inconsistently across the four UKHLS waves.

Table 5: Fractional response model for fraction of waves (as opposed to the total number of waves available) with inconsistent responses to SAH questions: marginal effects.

	Coeff.	SE
Male	0.025***	0.003
Aged 26-35	-0.018***	0.007
Aged 36-45	-0.009	0.007
Aged 46-55	-0.005	0.007
Aged 56-65	-0.019**	0.008
Aged 66-75	-0.011	0.010
Aged 76-85	-0.007	0.012
Aged 86-104	-0.006	0.023
Other higher qualification	0.017***	0.005
A-level	0.021***	0.005
GCSE	0.038***	0.005
Other low qualification	0.035***	0.006
No qualification	0.037***	0.006
Log of the HH income	-0.018***	0.004
Unemployed	0.007	0.008
Retired	-0.023***	0.006
Other job status	-0.029***	0.005
Single	0.001	0.005
Separated/divorced	-0.002	0.006
Widowed	-0.003	0.008
Immediate word recall	-0.014***	0.003
Delayed word recall	-0.005**	0.002
Number of correct subtractions	-0.009***	0.002
Verbal fluency: correct words	-0.007***	0.002
Verbal fluency: incorrect words	0.005***	0.001
Numeric ability: correct answers	-0.017***	0.002
Agreeableness	-0.002	0.002
Conscientiousness	-0.003**	0.002
Extraversion	0.002	0.002
Neuroticism	-0.007***	0.002
Openness to experience	0.006***	0.002
Multiple adult interviews/HH	-0.001	0.004
Children (aged 10-15) interviewed/HH	0.015***	0.004
Log-likelihood	-11,629.0	
Sample size	22,650	

Notes: The outcome variable is defined as the fraction of waves with inconsistent responses to SAH questions relative to total number of waves (varying from 0 to 1). Reference categories for the categorical variables included in our model: age groups (age 16-25), gender (female), education (Degree), marital status (married/cohabiting), job status (employed), and regional dummies (North East). Standard errors (SE) are clustered at individual level. Balanced sample is used for these estimations.
* p<0.10, ** p<0.05, *** p<0.01

The marginal effects from the fractional response model provide an indication of the magnitude of the association between the baseline characteristics used in our analysis and the frequency of reporting SAH inconsistently across the available four waves (Table

5). Men have an increased frequency of within-wave inconsistent SAH reporting across all four UKHLS waves. Age has a negative but non-monotonic association with inconsistently reporting SAH more frequently. For example, respondents belonging to the 26-35 or 56-65 age group at baseline, as opposed to our reference category (16-25 age group), show reduced frequency of reporting inconsistent responses to SAH questions across waves.

Turning to education, we observed a strong gradient, with lower education at baseline being positively associated with a higher frequency of being inconsistent SAH responses. For example, no educational qualifications, as compared to having a university degree (reference group), is associated with an increased frequency of inconsistent SAH reporting within each of the four UKHLS waves of about 3.7 percentage points. Concerning household income, there is a negative gradient, with richer respondents at baseline show reduced frequency of within-wave inconsistently reporting SAH. Compared to those employed, retired respondents and those of other job statuses show smaller frequency of inconsistently reporting SAH. The presence of children during the open interview is associated with increased frequency of inconsistent SAH reporting within each of the four UKHLS waves.

The cognitive ability measures have a systematic gradient with the frequency of inconsistent reporting to SAH; higher levels of cognitive ability are associated with reduced frequency of inconsistent SAH reporting within each of the four UKHLS waves. The magnitude of the marginal effects, which are expressed in terms of deviations for their standard deviations, show that episodic memory (captured by the word recall measures) and numerical knowledge are those cognitive skills that have a higher association with the frequency of reporting inconsistency. The role of verbal fluency is lower in magnitude, but highly statistically significant.

Regarding the Big-5 personality traits, conscientiousness and neuroticism are most strongly associated with reporting patterns in SAH. Openness to experience is associated with a higher frequency of inconsistent SAH reporting within each of the four UKHLS waves. Overall, these results highlight that the personality traits that are relevant to a higher level of self-discipline, exploring perceptual information in inflexible and divergent ways and neuroticism are relevant for persistent patterns in these behaviours across the four UKHLS waves.

5.3 Implications for the income-health gradient

Our results so far show that the observed inconsistencies in reporting in SAH within each wave and the frequency of inconsistency across waves are systematic and associated with individual characteristics. Among these individual characteristics, household income may be of particular interest given the existing research on the income-health gradient based on SAH measures. We find that household income is negatively associated with the probability of reporting SAH inconsistently at each wave as well as with the probability of this being persistent across waves.

To explore whether difference in interview mode (self-completion versus open interview) affects the association between income and SAH, we estimate linear regression models for SAH regressed on household income after adjusting for age, gender and wave fixed effects (Table 6, Panel A). Recall, that we have coded SAH so that higher values indicate a better health state. Our estimates for the full sample suggest evidence of a positive and highly significant association between income and health with remarkably similar income coefficients in the case of self-completion and open interview SAH measures.

We then implement further analysis, excluding from our estimation sample certain sub-samples with specific patterns in reporting SAH.[15] As expected, identical income coefficients are evident for the case of the two models when we restrict our analysis to those who consistently report SAH within each of the four waves used here (Table 6, Panel B: Always consistent) and there is evidence of positive and systematic associations between income and health. Analysis across all the remaining sub-samples show practically identical results, with very limited differences in the estimated associations between income and health when using our two SAH measures.

[15] We need to highlight here that our scope is not to explore the implications of measurement error in SAH on the income-health gradient. Our focus is much narrower, and it simply limited to explore whether the two SAH measures (open interview versus the self-completion) result into different income coefficients. Although we draw some general argument for the measurement error literature (especially to interpret the reduction in the income coefficients as moving from Panel B to Panel E in Table 6), our aim here is not to assess the implications of any measurement error in SAH more generally but the implications of employing SAH measures that are based on different interview modes inconspicuously as it is done in existing studies. For example, even when multiple SAH measures are available in a dataset within the same wave, the authors, unconsciously, consider responses of one SAH measure for their analysis despite concerns regarding the consistency of these SAH measures (e.g., Contoyannis et al., 2004, Fichera and Gathergood, 2016; Li Donni et al., 2014).

Table 6. Income gradients in SAH measured using self-completion and open interview mode: pooled OLS.

	SAH – Open Interview (H_{OI})	SAH – Self-completion (H_{SC})	Difference in coefficients
	Coef. (SE)	Coef. (SE)	p-value
Panel A: Full sample			
Log of the HH income	0.419*** (0.011)	0.419*** (0.011)	0.948
Sample size	90,600		
Panel B: Always consistent (0 inconsistencies across four waves)			
Log of the HH income	0.518*** (0.018)	0.518*** (0.018)	–
Sample size	38,568		
Panel C: Full sample excluding those always consistent (1-4 inconsistencies across four waves)			
Log of the HH income	0.335*** (0.013)	0.338*** (0.013)	0.575
Sample size	52,032		
Panel D: Inconsistent responses in 1 or 2 out of four waves in total (mostly consistent and moderately inconsistent)			
Log of the HH income	0.347*** (0.014)	0.347*** (0.014)	0.926
Sample size	46,988		
Panel E: Inconsistent responses in 3 or 4 out of four waves in total (mostly inconsistent / always inconsistent)			
Log of the HH income	0.214*** (0.040)	0.269*** (0.036)	0.123
Sample size	5,044		

Notes: Pooled OLS estimation accounts for wave fixed effects, gender (female as reference category) and age dummies (age 16-25 as reference category). Standard errors (SE) are clustered at individual level. The balanced sample is used for estimation. Higher SAH values indicate a better health state.
* p<0.10, ** p<0.05, *** p<0.01

Our results from the models on reporting inconsistency in SAH shows that income is a statistically significant correlate, which may suggest the presence of a systematic correlation between inconsistency reporting to SAH questions and income, our explanatory covariate of interest; in this case, if drawing from the measurement error literature, the relevant OLS estimates may be biased and inconsistent. A closer look to the results from Table 6 show that as moving from Panel B to Panel E, there is a stronger absolute correlation between reporting inconsistencies and income as we increasingly focus on sub-samples that report SAH inconsistently. The observed decline in the magnitude of the income coefficients between Panel B and Panel E suggest that there is an under-estimation of the income-health gradient to what would have been observed if focusing only to those who reported SAH with consistency.

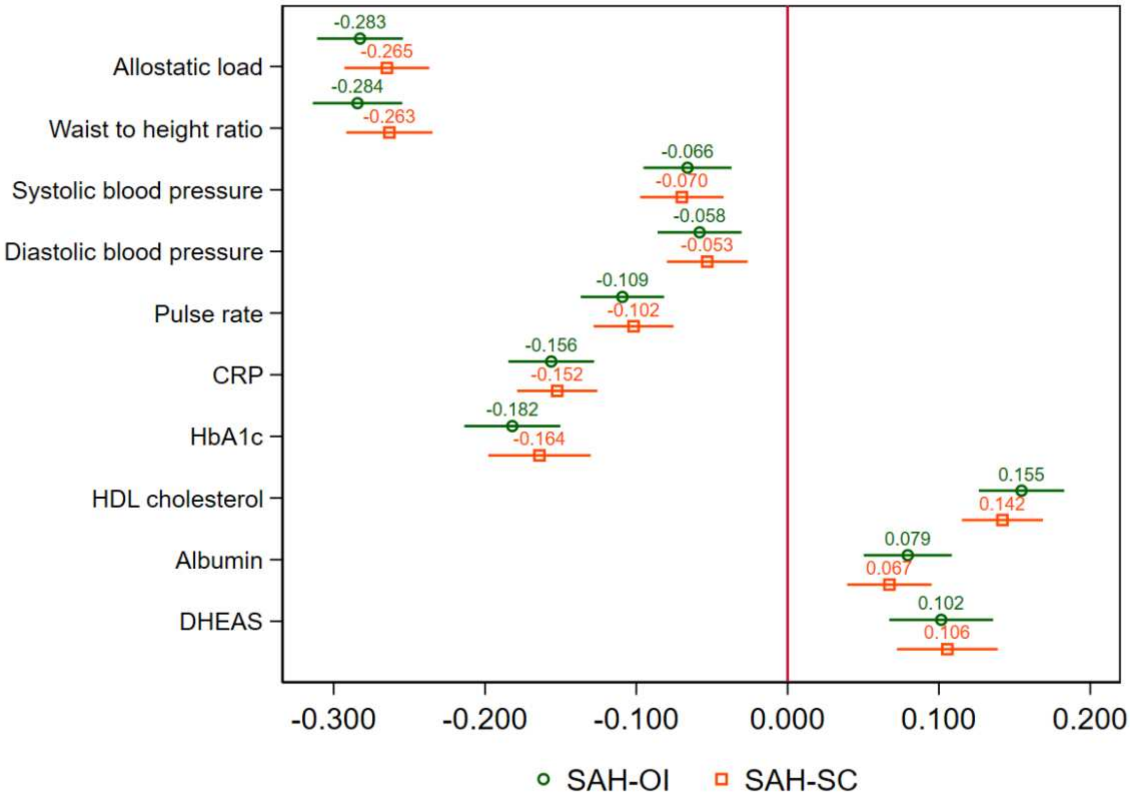
5.4 The association between biomarkers and SAH

Figure 4 presents the coefficient estimates for each of the biomarkers and for allostatic load, measured in units of standard deviations, where each of those is included separately in a linear regression model for both SAH measures. We follow Crossley (2002) for simplicity and adopt linear regression approach for this analysis.[16] This simple model helps us to test the hypothesis that if the mode of data collection does not matter for individual reporting behavior about their health status, we should not observe differences in the association between objective and subjective health measures. Moreover, we are also concerned with the potential multicollinearity among biomarkers which may lead to wide confidence intervals of the estimates. This is the reason why we run specific regression for each biomarker as the explanatory variable of interest. To summarize the selected biomarkers into a composite measure, we use allostatic load index (see Davillas and Pudney, 2020). All coefficients have the expected sign: those that reflect higher health risks are negatively associated with SAH. Positive associations are observed for HDL cholesterol, Albumin and DHEAS as higher values of these biomarkers as associated with lower health risks.

Overall, the biomarker coefficient estimates do not vary systematically between the two SAH measures, indicating that there are limited differences in how these objectively collected health dimensions are associated with the two SAH measures; this is also the case for the coefficient estimates of our composite health measure — allostatic load. Allostatic load (standardized by its standard deviation), our composite biological measure, is strongly associated with SAH. Turning to the underlying biomarkers (standardized by their standard deviation), adiposity, followed by the diabetes biomarker (HbA1c), inflammation (CRP), "fat in the blood" biomarker (HDL cholesterol), resting pulse rate and stress related steroids (DHEAS) are most strongly associated with both SAH measures.

[16] Crossley (2002) uses linear regression approach to analyze the attenuation bias associated with SAH measures when explaining individual employment status.

Figure 4. Coefficient estimates from biomarkers in self-completion and open interview SAH regression models.



Notes: Each of the biomarkers (and allostatic load) are included separately in linear regression models of SAH outcomes measured using the self-completion and the open interview mode. Higher SAH values indicate a better health state. All regressions account for age dummies, gender, regional dummies and wave fixed effects. Sample size: 5,907 observations.

6. Conclusion

Beyond concerns to better understand the extent to which SAH can be interpreted as a good proxy of individual's underlying health, there is a small literature on whether individuals answer SAH questions consistently. We add to this and capitalise on the rare opportunity provided by the UKHLS asking respondents the same SAH question with identical wording twice (one with a self-completion and one with an open interview mode), within the same household interview, at UKHLS Waves 2, 3, 4 and 5. Descriptive analysis shows substantial inconsistency in reporting behaviour. Within each wave, about 11%-24% of those who reported a given SAH category in self-completion mode reported inconsistently in the open interview, with the majority of the inconsistent responses concentrated in the SAH categories adjacent to their self-completion responses. Descriptive analysis of the sequences of reporting reveals that only 43% of sample members are "always consistent".

We use multivariate models to explore the profile of those who are more likely to report SAH inconsistently. Sex, age, educational attainment, household income, employment status, cognitive and non-cognitive skills are associated with the probability of reporting SAH inconsistency within a wave. Despite existing evidence that the micro-social environment during the household interviews (proxied by the presence of other adults or children during the interview) potentially affects responses to the self-completed wellbeing measures (e.g., Black et al., 2017; Conti and Pudney, 2011), we find less pronounced evidence that they play a systematic role in reporting inconsistencies in SAH. For instance, the presence of another adult during the open interview does not systematically associated with reporting inconsistency to SAH questions, while the association between the presence of children during the interview in the household and inconsistency reporting in the relevant open interview as opposed to self-completion SAH question is not robust when controlling for unobserved heterogeneity.

We capitalise on the longitudinal data available to explore the profile of those who report inconsistently more frequently over the four waves of data. We find that those baseline characteristics that predict the probability of reporting inconsistently are also systematically associated with a higher frequency of reporting inconsistently.

Our evidence suggests that the observed reporting inconsistency in SAH is not purely random measurement error; individuals with certain characteristics are on average less likely to report SAH consistently across the two measures using different modes of administration (self-completed versus open interview). Socioeconomic status, among other individual characteristics, plays an important role in reporting inconsistency. This may suggest that existing studies that use SAH as an outcome, modelled as a function of socioeconomic status, may be contaminated with the corresponding results potentially encompass significant biases.[17] Analysis of the association between income and health shows no evidence that employing the self-completion SAH measure as opposed to the open interview SAH (and vice versa) may affect the results. Specifically, we show that the interview mode, which is associated with inconsistent reporting behaviour, does not play a role in the estimation of the income-health gradient either based on the self-completion or the open interview SAH measures. Finally, of particular interest, we do not find systematic differences in the association between our SAH measures, administered using the open interview and the self-completion mode, and a large set of objectively measured nurse-collected and blood-based biomarkers. This may suggest that reporting inconsistencies are driven by mechanisms other than people's physiological and biological health (or at least the dimensions of physical health captured by our set of biomarkers). Overall, our results show that inconsistent reporting behaviour about individual SAH is associated with socioeconomic and demographic characteristics, cognitive skills, and personality traits. However, in term of our income-health gradient illustrative example, we found no evidence that using the open interview SAH questionnaire as opposed to the self-completion may affect our results.

Researchers working with SAH measures should bear in mind the sensitivity of SAH measurement as individuals may report inconsistently to the same SAH question within a very short time interval and without any changes in their underlying health. Despite our reassuring evidence that the income-health gradient remained robust to SAH measures collected with the two interview modes, further research is needed to explore potential implications of inconsistency reporting on the reliability of SAH as a health measure and its implications for the exiting research.

[17] If inconsistent reporting behaviour is associated with socioeconomic status, measurement error in SAH will be correlated with explanatory variables in regression model leading to biased estimates (see Clarke and Ryan, 2006).

References

Aoki, Y., Santiago, L. (2018). Speak better, do better? Education and health of migrants in the UK. *Labour Economics,* 52, 1-17.

Au, N., Johnston, D.W. (2014). Self-assessed health: what does it mean and what does it hide? *Social Science & Medicine*, 121, 21-28.

Bago d'Uva, T., O'Donnell, O., van Doorslaer, E. (2008). Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology*, 37, 1375-1383.

Barford, K. A., Smillie, L. D. (2016). Openness and other Big Five traits in relation to dispositional mixed emotions. *Personality and Individual Differences*, 102, 118-122.

Bertrand, M., Mullainathan, S., 2001. Do people mean what they say? Implications for subjective survey data. *American Economic Review*, 91 (2), 67e72.

Black, B., Johnston, D. W., Suziedelyte, A. (2017). Justification bias in self-reported disability: New evidence from panel data. *Journal of Health Economics*, 54, 124-134.

Black, N., Johnston, D.W., Shields, M.A., Suziedelyte, A. (2017). Who provides inconsistent reports of their health status? The importance of age, cognitive ability and socioeconomic status. *Social Science & Medicine*, 191, 9-18.

Bound, J. (1991). Self-Reported Versus Objective Measures of Health in Retirement Models. *Journal of Human Resources*, 26, 106-138.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*. 27 (3), 281–291.

Butler, J. S., Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, 50, 761-764.

Chen, L., Clarke, P.M., Petrie, D.J., Staub, K.E. (2021). The effects of self-assessed health: Dealing with and understanding misclassification bias. *Journal of Health Economics*, 102463.

Clarke, P.M., Ryan, C. (2006). Self-reported health: reliability and consequences for health inequality measurement. *Health Economics*, 15 (6), 645–652.

Colburn, W., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A., Woodcock, J., et al., (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Biomarkers Definitions Working Group. *Clinical Pharmacology & Therapeutics*, 69, 89–95.

Conti, G., Pudney, S. (2011). Survey design and the analysis of satisfaction. *Review of Economics and Statistics*, 93, 1087-1093.

Cobb-Clark, D.A., Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, 115(1), 11-15.

Contoyannis, P. Jones, A. M. (2004). Socio-economic status, health and lifestyle. *Journal of Health Economics*, 23, 965-995.

Contoyannis, P., Jones, A.M., Rice, N. (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics,* 19, 473-503.

Cookson, R., Propper, C., Asaria, M., Raine, R. (2016). Socio-Economic Inequalities in Health Care in England. *Fiscal Studies*, 37, 371-403.

Crossley, T.F., Kennedy, S. (2002). The reliability of self-assessed health status. *Journal of Health Economics,* 21, 643-658.

Currie, J., Duque, V., Garfinkel, I. (2015). The Great Recession and mothers' health. *The Economic Journal*, 125, F311-F346.

Davillas, A., Jones, A. M., Benzeval, M. (2019). The income-health gradient: evidence from self-reported health and biomarkers in understanding society. in Panel Data Econometrics, Academic Press.

Davillas, A., Jones, A.M. (2020). Ex ante inequality of opportunity in health, decomposition and distributional analysis of biomarkers. *Journal of Health Economics*, 69, 102251.

Davillas, A., Pudney, S. (2020). Biomarkers as precursors of disability. *Economics & Human Biology*, 36, 100814.

Donni, P. L., Peragine, V., Pignataro, G. (2014). Ex-ante and Ex-post measurement of equality of opportunity in health: A normative decomposition. *Health Economics*, 23(2), 182-198.

Dowd, J.B., Zajacova, A. (2010). Does self-rated health mean the same thing across socioeconomic groups? Evidence from biomarker data. *Annals of Epidemiology*, 20, 743-749.

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

Fichera, E., Gathergood, J. (2016). Do wealth shocks affect health? New evidence from the housing boom. *Health Economics*, 25, 57-69.

Fisher, P. (2019). Does repeated measurement improve income data quality? *Oxford Bulletin of Economics and Statistics*, 81, 989-1011.

Foverskov, E., Holm, A. (2016). Socioeconomic inequality in health in the British household panel: tests of the social causation, health selection and the indirect selection hypothesis using dynamic fixed effects panel models. *Social Science & Medicine*, *150*, 172-183.

Frijters, P., Haisken-DeNew, J. P., Shields, M.A. (2005). The causal effect of income on health: Evidence from German reunification. *Journal of Health Economics*, *24*, 997-1017.

Fuchs, V. R. (2004). Reflections on the socio-economic correlates of health. *Journal of Health Economics*, *23*(4), 653-661.

García-Gómez, P., Jones, A. M., Rice, N. (2010). Health effects on labour market exits and entries. *Labour Economics,* 17, 62-76.

Greene, W. H. Econometric Analysis. 5th ed., Pearson Education Prentice-Hall, Upper Saddle River, NJ (2003).

Gunasekara, F. I., Carter, K., Blakely, T. (2012). Comparing self-rated health and self-assessed change in health in a longitudinal survey: Which is more valid? *Social Science & Medicine*, 74(7), 1117-1124.

Hu, Y., van Lenthe, F. J., Borsboom, G. J., Looman, C. W. N., Bopp, M., Burström, B., Dzúrová, D., Ekholm, O. Klumbiene, J., Lahelma, E., Leinsalu, M., Regidor, E., Santana, P., de Gelder, R., Mackenbach, J. P. (2016). Trends in socioeconomic inequalities in self-assessed health in 17 European countries between 1990 and 2010. *Journal of Epidemiology and Community Health*,70, 644-652.

Jerram, K. L., Coleman, P. G. (1999). The big five personality traits and reporting of health problems and health behaviour in old age. *British Journal of Health Psychology*, 4(2), 181-192.

John, O., Donahue, E., Kentle, R. (1991). The big five inventory-Versions 4a and 54. Berkeley, CA: University of California. Berkeley, Institute of Personality and Social Research.

Johnson, R. C. (2010). The health returns of education policies from preschool to high school and beyond. *American Economic Review,* 100, 188-194.

Johnston, D. W., Propper, C., Shields, M.A. (2009). Comparing subjective and objective measures of health: Evidence from hypertension for the income/health gradient. *Journal of Health Economics*, 28(3), 540-552.

Jylhä, M. (2009). What is self-rated health and why does it predict mortality? Towards a unified conceptual model. *Social Science & Medicine*, 69, 307-316.

Kaplan, G.A., Camacho, T. (1983). Perceived health and mortality: a nine-year follow-up of the human population laboratory cohort. *American Journal of Epidemiology*, 117, 292-304.

Kapteyn, A., Smith, J.P., van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, 97, 461-473.

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

Kunst, A.E., Bos, V., Lahelma, E., Bartley, M., Lissau, I., Regidor, E., Mielck, A., Cardano, M., Dalstra, J.A., Geurts, J.J., Helmert, U., Lennartsson, C., Ramm, J., Spadea, T., Stronegger, W.J., Mackenbach, J.P. (2004). Trends in socioeconomic inequalities in self-assessed health in 10 European countries. *International Journal of Epidemiology*, 34, 295-305.

Larrimore, J. (2011). Does a higher income have positive health effects? Using the earned income tax credit to explore the income-health gradient. *The Milbank Quarterly*, *89*, 694-727.

Li Donni, P., Peragine, V., Pignataro, G. (2014). Ex-ante and ex-post measurement of equality of opportunity in health: a normative decomposition. *Health Economics*, 23, 182-198.

Lin, D., Lutter, R., Ruhm, C. J. (2018). Cognitive performance and labour market outcomes. *Labour Economics*, 51, 121-135.

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

List, J. A., Berrens, R. P., Bohara, A. K., Kerkvliet, J. (2004). Examining the role of social isolation on stated preferences. *American Economic Review*, 94, 741–752.

Lumsdaine, R. L., Exterkate, A. (2013). How survey design affects self-assessed health responses in the Survey of Health, Ageing, and Retirement in Europe (SHARE). European Economic Review, 63(C), 299-307.

Lyons, M. J., Panizzon, M. S., Liu, W., McKenzie, R., Bluestone, N. J., Grant, M. D., Franz, C. E., Vuoksimaa, E. P., Toomey, R., Jacobson, K. C., Reynolds, C. A., Kremen, W. S., Xian, H. (2017). A longitudinal twin study of general cognitive ability over four decades. *Developmental Psychology*, 53(6), 1170.

Monden, C.W. (2010). Do measured and unmeasured family factors bias the association between education and self-assessed health? *Social Indicators Research*, 98, 321-336.

Monk Jr, E.P. (2015). The cost of color: Skin color, discrimination, and health among African-Americans. *American Journal of Sociology*, 121, 396-444.

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

Nouri, F., Feizi, A., Afshar, H., Keshteli, A.H., Adibi, P. (2019). How five-factor personality traits affect psychological distress and depression? results from a large population-based study. *Psychological Studies*, *64*, 59-69.

Nesson, E.T., Robinson, J.J. (2019). On the measurement of health and its effect on the measurement of health inequality. *Economics & Human Biology*, 35, 207-221.

Nutbeam, D. (2008). The evolving concept of health literacy. *Social Science & Medicine*, 67(12), 2072-2078.

Ohlsson, C., Labrie, F., Barrett-Connor, E., Karlsson, M. K., Ljunggren, O., Vandenput, L., Mellström, D., Tivesten, A. (2010). Low serum levels of dehydroepiandrosterone sulfate predict all-cause and cardiovascular mortality in elderly Swedish men. *The Journal of Clinical Endocrinology & Metabolism*, 95(9), 4406-4414.

Pearson, T. A., Mensah, G. A., Alexander, R. W., Anderson, J. L., Cannon III, R. O., Criqui, M., Fadl, Y., Fortmann, S., Hong, Y., Myers, G., Rifai, N., Smith, S., Taubert, K., Tracy, R., Vinicor, F. (2003). Markers of inflammation and cardiovascular disease: application to clinical and public health practice: a statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association. *Circulation*, 107(3), 499-511.

van Doorslaer, E., Wagstaff, A., van der Burg, H., Christiansen, T., De Graeve D., Duchesne I, Gerdtham, U.G., Gerfin, M., Geurts J, Gross L, Häkkinen U, John J, Klavus J, Leu RE, Nolan B, O'Donnell O, Propper C, Puffer, F., Schellhorn, M., Sundberg, G., Winkelhake, O. (2000). Equity in the delivery of health care in Europe and the US. *Journal of Health Economics*, 19, 553-583.

Voicu, B. (2015). Priming effects in measuring life satisfaction. Social Indicators Research, 124, 993-1013.

Wooldridge, M. J. (2002). Econometric analysis of cross section and panel data. The MIT Press, London.

Wooldridge, M. J. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20(1), 39-54.

Zajacova, A., Huzurbazar, S., Todd, M. (2017). Gender and the structure of self-rated health across the adult life span. *Social Science & Medicine*, 187, 58-66.

Ziebarth, N. (2010). Measurement of health, health inequality, and reporting heterogeneity. *Social Science & Medicine*, 71(1), 116-124.

Online Appendix

Table A1 Distribution of (in)consistent responses to SAH open interview (SAH-OI) and self-completion (SAH-SC) mode

Waves	SAH-OI > SAH-SC	SAH-OI = SAH-SC	SAH-OI < SAH-SC	Total
Wave 2	2,297 (10.1%)	17,307 (76.4%)	3,046 (13.5%)	22,650 (100%)
Wave 3	2,937 (13.0%)	17,569 (77.6%)	2,144 (9.5%)	22,650 (100%)
Wave 4	3,101 (13.7%)	17,964 (79.3%)	1,585 (7.0%)	22,650 (100%)
Wave 5	2,128 (9.4%)	18,344 (81.0%)	2,178 (9.6%)	22,650 (100%)
Total	**10,463 (11.6%)**	**71,184 (78.6%)**	**8,953 (9.9%)**	**90,600 (100%)**

Table A2. Pooled probit model and random effects (RE) probit model for inconsistent responses to SAH questions.

	Pooled probit		Random Effects probit	
Covariates	**Coeff.**	**SE**	**Coeff.**	**SE**
Male	0.088***	0.012	0.095***	0.012
Aged 26-35	-0.074***	0.024	-0.078***	0.026
Aged 36-45	-0.061**	0.025	-0.065**	0.026
Aged 46-55	-0.007	0.025	-0.008	0.026
Aged 56-65	-0.063**	0.027	-0.065**	0.028
Aged 66-75	-0.019	0.032	-0.023	0.035
Aged 76-85	-0.019	0.039	-0.022	0.041
Aged 86 and older	0.002	0.065	-0.008	0.069
Other higher qualification	0.066***	0.018	0.072***	0.020
A-level	0.083***	0.016	0.090***	0.017
GCSE	0.136***	0.017	0.148***	0.018
Other low qualification	0.123***	0.021	0.135***	0.022
No qualification	0.116***	0.023	0.125***	0.024
Log of the HH income	-0.047***	0.012	-0.045***	0.013
Unemployed	0.009	0.027	-0.000	0.028
Retired	-0.097***	0.021	-0.103***	0.022
Other job status	-0.097***	0.017	-0.096***	0.018
Single	-0.003	0.018	-0.004	0.019
Separated/divorced	-0.016	0.020	-0.017	0.022
Widowed	-0.000	0.026	0.003	0.028
Immediate word recall†	-0.046***	0.009	-0.049***	0.009
Delayed word recall†	-0.019**	0.008	-0.020**	0.009
Number of correct subtractions†	-0.030***	0.006	-0.033***	0.006
Verbal fluency: correct words†	-0.023***	0.007	-0.026***	0.007
Verbal fluency: incorrect words†	0.014***	0.005	0.015***	0.005
Numeric ability: correct answers†	-0.065***	0.007	-0.070***	0.008
Agreeableness†	-0.010*	0.006	-0.010*	0.006
Conscientiousness†	-0.013**	0.006	-0.014**	0.006
Extraversion†	0.006	0.006	0.007	0.006
Neuroticism†	-0.026***	0.006	-0.028***	0.006
Openness to experience†	0.020***	0.006	0.022***	0.006
Multiple adult interviews/HH	-0.012	0.014	-0.013	0.015
Children (aged 10-15) interviewed/HH	0.032**	0.015	0.029*	0.016
Wave 3	-0.037***	0.013	-0.040***	0.014
Wave 4	-0.097***	0.013	-0.105***	0.014
Wave 5	-0.158***	0.013	-0.170***	0.014
Constant	0.294***	0.111	0.282**	0.117
Rho			0.121***	0.006
Sigma(u)			0.371***	0.010
Log-likelihood	-46,301		-46,063	
Sample size	90,600		90,600	

Notes: The pooled probit and RE probit models account for region fixed effects. Reference categories: age groups (age 16-25), gender (female), education (Degree), marital status (married/cohabiting), job status (employed), presence of other adults in the household during the interview (none), presence of children in the household during the interview (none), wave fixed effects (Wave 2), and region fixed effects (North East). Standard errors (SE) are clustered at individual level. Balanced sample is used for these estimations.

† Expressed in terms of deviations from their standard deviation.

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

Table A3: Fractional response model for inconsistent responses to SAH questions.

	Coeff.	SE
Male	0.086***	0.012
Aged 26-35	-0.064***	0.024
Aged 36-45	-0.031	0.024
Aged 46-55	-0.018	0.024
Aged 56-65	-0.065**	0.026
Aged 66-75	-0.040	0.033
Aged 76-85	-0.025	0.041
Aged 86-104	-0.022	0.078
Other higher qualification	0.060***	0.019
A-level	0.072***	0.017
GCSE	0.130***	0.017
Other low qualification	0.122***	0.021
No qualification	0.129***	0.022
Log of the HH income	-0.063***	0.012
Unemployed	0.026	0.026
Retired	-0.078***	0.022
Other job status	-0.099***	0.017
Single	0.005	0.018
Separated/divorced	-0.008	0.021
Widowed	-0.010	0.027
Immediate word recall†	-0.048***	0.009
Delayed word recall†	-0.018**	0.008
Number of correct subtractions†	-0.029***	0.006
Verbal fluency: correct words†	-0.025***	0.006
Verbal fluency: incorrect words†	0.017***	0.005
Numeric ability: correct answers†	-0.060***	0.007
Agreeableness†	-0.007	0.006
Conscientiousness†	-0.012**	0.006
Extraversion†	0.006	0.006
Neuroticism†	-0.025***	0.006
Openness to experience†	0.019***	0.006
Multiple adult interviews/HH	-0.003	0.015
Children (aged 10-15) interviewed/HH	0.052***	0.015
Constant	0.288***	0.109
Log-likelihood	-11,629.0	
Sample size	22,650	

Notes: The outcome is the ratio between the number of inconsistent responses (0 to 4) and the number of waves, varying from 0 to 1. Reference categories: age groups (age 16-25), gender (female), education (Degree), marital status (married/cohabiting), job status (employed), presence of other adults in the household during the interview (none), presence of children in the household during the interview (none), and region fixed effects (North East). Standard errors (SE) are clustered at individual level. Standard errors (SE) are clustered at individual level. Balanced sample is used for these estimations.

† Expressed in terms of deviations from their standard deviation.

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