Who Provides Inconsistent Reports of their Health Status?

The Importance of Age, Cognitive Ability and Socioeconomic Status

Abstract

Self-assessed health (SAH) measures are widely used in models of health and health inequalities. Such models assume that SAH is a reliable measure of health status. We utilise a unique feature of a national longitudinal survey to examine the consistency of responses to a standard SAH question that is asked twice to the same individual in close temporal proximity in up to three waves (2001, 2009, and 2013). In particular, we analyse whether the consistency of responses varies with personal characteristics. The main analysis sample includes 18,834 individual-year observations. We find that 57% of respondents provide inconsistent reports at least once. Characteristics that are associated with significantly higher inconsistencies are age, education, cognitive ability, and time between responses. The results suggest that there are systematic differences in the ability of individuals' to self-evaluate and summarise their own health. Consequently, failure to account for such error may lead to large estimation biases in models of health outcomes, particularly with respect to the relationship between education, cognitive ability, and health.

Keywords: Australia; self-assessed health; health reporting; socioeconomic status; cognition; panel data; mixed-effects multinomial logit model

1. Introduction

Every year numerous published articles in epidemiology, public health and economics use self-assessed health (SAH) to provide new knowledge about the demographic and socioeconomic characteristics associated with good health and the extent of health inequalities. Such evidence is then used to motivate health and health care policy. In particular, a large literature has examined the relationship between self-assessed health and education, unemployment, household income, occupation, wealth, neighbourhood deprivation, early life circumstances and retirement (see, for example, Meer et al. 2003; Contoyannis et al. 2004; Kunst et al. 2005; Cutler and Lleras-Muney 2010; Johnson 2010; and Hu et al. 2016). In this study we examine the reliability of general SAH, arguably the most commonly analysed measure of health, where respondents are asked to rate their current general health on an ordinal scale (for example, excellent, very good, good, fair, poor). If it is the case that the willingness or capability of individuals to consistently answer this type of question is associated with certain individual-level characteristics and traits (such as education and cognitive ability), then inferences about inequalities in health may be misleading.

To address this issue we take advantage of a peculiarity in a large longitudinal survey that provides a 'quasi-experiment' for analysing reporting consistency in SAH. In three waves of the Household, Income and Labour Dynamics in Australia (HILDA) survey, respondents are asked, within a short period of time, to report their general health using two different survey modes: face-to-face interview (FFI) and self-completion questionnaire (SCQ). By analysing the variation in responses, we are able to document the extent of inconsistency in SAH and determine whether the inconsistency is influenced by individuals' characteristics and traits including their education, employment status, cognitive ability, and personality.

There are several reasons why respondents may report their health differently, even if asked on the same day when their underlying level of health arguably remains unchanged. First, it is likely that a degree of uncertainty exists in identifying underlying health levels, and therefore individuals assess their health with some "error" each time (Crossley and Kennedy 2002). Given the cognitive demands placed on respondents, such as comprehending the question, recalling information from memory, and communicating the response (Bowling, 2005), it is likely that cognitive ability will affect the consistency of responses. Similarly, certain personality traits, such as conscientiousness, may affect the effort and consideration that is taken when answering survey questions and therefore on the reliability of responses (Bertrand and Mullainathan 2001). Second, it is conceivable that the mode of data collection influences responses (Bowling 2005). Different modes require different skills (for example, verbal and listening in interviews, and reading and writing in paper questionnaires). Moreover, the social nature of interviews may induce individuals to give more positive and socially desirable responses, which is known as social desirability bias (Bowling 2005). This bias can also be associated with certain individual characteristics and traits. For example, older individuals may understate any health problems in an interview so as to appear healthy and robust in front of the interviewer. In contrast, unemployed individuals may feel socially conditioned to inflate their health problems in order to help justify their unemployment status (Anderson and Burkhauser 1985). Third, there may be "learning effects", whereby preceding questions influence or frame the respondent's perception about their health. In particular, specific health and disease questions can influence subsequent responses about general health (Bowling and Windsor 2008).

Our study complements a growing literature that examines reporting heterogeneity in SAH across different groups of individuals. These studies, which typically condition on a measure of underlying latent health, have identified a number of reporting phenomena, including justification bias (Bound 1991; Kerkhofs and Lindeboom 1995; Lindeboom and Kerkhofs 2009), reference group effects (Wiseman 1999; Groot 2000) and heterogeneity in the interpretation and use of response scales (Lindeboom and Van Doorslaer 2004; Etilé and Milcent 2006; Bago d'Uva et al. 2011). Other related studies have confirmed that survey design affects responses, and that these effects differ by certain individual characteristics (Lumsdaine and Exterkate 2013; Holford and Pudney 2015). However, to the best of our knowledge, few studies have had the ability to examine the extent, characteristics, and consequences of inconsistency in responses to two near identical questions asked in close temporal proximity to the same individual. The closest examples are Crossley and Kennedy (2002) and Clarke and Ryan (2006), which focus on survey mode and question ordering effects. Both studies find that around 30% of respondents changed their responses to the general health question asked in different survey modes. While Crossley and Kennedy (2002) conclude that both survey mode and question ordering are likely to play a role, Clarke and Ryan (2006) find that survey mode has the dominant role in response changes.

We build on both of these studies in a number of important ways. First, we examine a much wider range of individual characteristics to provide a more complete understanding of who reports their health inconsistently. Many of the characteristics we examine, such as cognitive ability and personality traits are key drivers of reporting behaviour, yet are rarely measured in household surveys. Second, our study has the advantage of having three waves of data, and we use statistical models that allow us to take into account repeated observations by individuals and households. This is important because inconsistencies in reporting are likely to be clustered at both the individual and household level; our approach enables this assumption to be tested for the first time. A further contribution of our study is the ability to separate out the influences of survey mode from those of question order, which has not been feasible in past studies.

2. HILDA and the Quasi-Experiment

Our data is drawn from the Household, Income and Labour Dynamics in Australia survey (HILDA), which is a nationally representative longitudinal study of Australian households that began in 2001. Wave 1 contained a sample of 19,914 panel members from 7,682 households, and in each subsequent year household members have been followed-up, along with any new household members resulting from changes in the composition of the original household. New households were included in the wave 11 top-up sample. The household response rates range from 87.0 per cent in wave 2 to 70.8 per cent in wave 11, while the household response rates for those households responding in the previous wave ranges from 87.0 per cent in wave 2 to 96.4 per cent in wave 11 (Summerfield et al. 2012).

Annual data is currently available from 2001 to 2014, and each year includes detailed information on income, employment, health and other demographic and socio-economic information.

In every wave, HILDA includes a confidential paper self-completion questionnaire (SCQ). The first question of the SCQ reads, "In general, would you say your health is", with respondents instructed to cross one box on a 5-point ordinal scale with the labels: Excellent, Very good, Good, Fair, and Poor. In waves 1, 9 and 13 (2001, 2009 and 2013) the face-to-face interview (FFI) includes a very similar SAH question: "In general, how would you rate your health? Is it excellent, very good, good, fair or poor?" Therefore, in three waves of HILDA, individuals are asked two near-identical SAH questions, with identical ordinal scales.

Aside from mode of administration, the FFI and SCQ health questions differ with respect to the preceding questions, which can have framing effects (Bertrand and Mullainathan 2001). The FFI question is located near the end of the interview at the beginning of a health module, and after modules on education, employment, and income, while the SCQ question is the first question on the SCQ and is also followed by a series of health questions. Therefore, the type and magnitude of priming will depend upon the ordering of the FFI and SCQ, which implies that the likelihood of inconsistency may be a function of ordering. Importantly, we are able to test for this because the ordering of FFI and SCQ was not fixed; some HILDA respondents completed the FFI before the SCQ and some after. In waves 9 and 13, the respondents record the date they complete the SCQ, and so we know that 43% of the SCQs and FFIs are completed on the same day (although we do not know the ordering). Of the remaining 57%, 80% completed the FFI before the SCQ.

In all analyses, we restrict the sample to individuals who are at least 25 years old and are not studying at the time of the interview, because completed education is one of our variables of interest. Our main analysis sample includes individuals for whom we know the date of the SCQ (i.e. who are present in waves 9 and 13) and who complete the FFI and SCQ not more than 30 days apart (97.4%). This allows us to examine the influence of days between SAH questions on response consistency and ensure that the underlying health of respondents remains unchanged between questions. The median and mean days between questions are 1 and 2.5 days, respectively. We also exclude phone interviews. The main analysis sample includes 18,834 individual-year observations.

Table 1 provides a brief description and mean values of the individual characteristics and traits that we use in our statistical modelling of response inconsistency. Largely following the literature (Crossley and Kennedy 2002; Clarke and Ryan 2006; Lumsdaine and Exterkate 2013), our base set of control variables are gender (47% are male), age (average age is 50.5 years), marital status (73.5% are married or co-habiting), number of children (0.542), highest level of qualification (27.2% have a university degree), employment status (7.3% are unemployed or marginally attached to the labour market, 23.0% are retired, and 6% are out of labour force for other reasons), immigrant status (23.7% were born overseas), and whether a foreign language is spoken in the home (mean is 9.9%). Given our focus on identifying potential explanations for why individuals provide inconsistent SAH responses, we also examine the role of cognitive ability. This is measured by the first and only predicted factor with eigenvalue greater than one (1.602) from a principal components factor analysis of three tests conducted in wave 12 relating to memory (the Backwords Digit Span Test), cognitive function (the Symbol-Digits Modalities Test) and verbal skills (the National American Reading Test) (Wooden, 2013). Since the cognitive test scores are only available for the respondents who were present in Wave 12 and completed the tests, we include indicators of not being asked to complete the tests (8%) and refusing to complete the tests (7%) in the regressions.

In extended model specifications we test to see whether the personality trait of conscientiousness is related to inconsistent SAH responses. Conscientiousness is measured as the average value across 6 questions (each answered on a 7 point scale) regarding being orderly, systematic, inefficient, sloppy, disorganised and efficient, asked to respondents in wave 5, 9 and 13. We also calculate the percentage of all questions that are left unanswered in the SCQ (varies from 0% to 98%) as a proxy for the respondent's effort or commitment to the survey, and create a variable indicating if the interviewer deemed that the respondent was suspicious or uncooperative in the FFI (about 1% of the sample). Another covariate is a binary variable indicating if a respondent reports that he/she is always or often pressed for time (36.1%), which might indicate that less time was spent on completing the SCQ (which we do not directly observe). Indicators are also included to test whether reporting inconsistency can be explained by language difficulties (2.6%) or eyesight/hearing problems (1.8%). We further examine possible explanations by testing if there is a relationship between response consistency and the length of time respondents have been followed in HILDA (9.16 years on average); the number of times the household had to be visited to complete the interviews; whether another person is present in the household when the FFI was being conducted (38%); and the number of days between the completion of the SCQ and FFI (the gap is greater than one week for 9.1% of observations).

3. Describing the Inconsistency in Self-Assessed Health

Figure 1 presents the distributions of responses to the SAH questions from the FFI and SCQ. The differences are small. More respondents assessed their health as 'poor' and 'excellent' in the FFI than in the SCQ (differences equal 0.8 and 2.2 percentage points, respectively), and fewer respondents assessed their health as 'good' (difference equals 3.2 percentage points). These differences imply that the mean and standard deviation of responses in the FFI (3.37 and 1.02) are slightly larger than the mean and standard deviation of responses in the SCQ (3.35 and 0.97).

Figure 2 clearly highlights the extent of inconsistency between self-assessments, by presenting the frequencies of the observed differences between individual SCQ and FFI responses in the same wave. Remarkably, almost 26% of responses are inconsistent. In 94% of inconsistent cases, the difference is one category (for example, a move from 'fair' to 'good', or from 'very good' to 'excellent' health). Interestingly, the inconsistencies are nearly symmetric, with 46% of inconsistent respondents reporting better health in the SCQ ($H_{FFI} < H_{SCQ}$) and 54% reporting better health in the FFI ($H_{FFI} > H_{SCQ}$). In data from all three waves (not shown), 57% of respondents are inconsistent at least once across the three waves.

Table 2 presents the numbers and frequencies of respondents for each of the 25 possible combinations of responses to the two questions. As demonstrated in Figure 2, the bottom-left and top-right corners have few observations, indicating that few respondents completely reverse their SAH. More interesting are the statistics around the main diagonal. Firstly, they show that respondents are most inconsistent at either end of the health distribution: only 66% of people who reported 'poor' in the FFI, reported 'poor' in the SCQ; and only 68% of people who reported 'excellent' in the FFI, reported 'excellent' in the SCQ. The reporting behaviour may be partly driven by a mode effect that causes some respondents to avoid extreme categories in SCQ (Clarke and Ryan, 2006). Secondly, Table 2 demonstrates that switches occur in roughly equal frequency across the distribution of responses: 30% of respondents who choose 'fair' in the FFI, choose a different response in the SCQ; similar percentages for respondents who switch from 'good' or 'very good' in the FFI to a different response in the SCQ equal 23% and 24%, respectively.

Importantly, the substantive inconsistency between responses is not randomly distributed across the sample. Figure 3 plots non-parametric regression estimates of the relationship between two binary measures of inconsistency – whether an individual reports worse health in the SCQ than in the FFI ($H_{FFI} > H_{SCQ}$) and whether an individual reports better health in the SCQ than in the FFI ($H_{FFI} < H_{SCQ}$) – and two individual characteristics, cognitive ability and age.

Figure 3A shows that individuals with low cognitive ability test scores have an estimated 18% and 15% likelihood of reporting $H_{FFI} > H_{SCQ}$ and $H_{FFI} < H_{SCQ}$, respectively, compared with a likelihood of around 8% for individuals with high scores. A move from the top to the bottom of the cognitive ability distribution is thus estimated to increase overall inconsistency by nearly 20 percentage points; a particularly large gradient. Figure 3B demonstrates that the likelihood of inconsistency also varies substantially with age. While the relationship between cognitive ability and inconsistency is roughly linear, the relationship between age and inconsistencey is nonlinear, with middle-aged respondents being the most consistent reporters of SAH. Respondents around 80 years are particularly inconsistent, with an 18% likelihood of reporting $H_{FFI} > H_{SCQ}$.

4. Statistical Methods

We further investigate the consistency of SAH by modelling whether health reporting in the FFI is equal, higher or lower than in the SCQ. This categorical treatment of the difference between assessments broadly follows the approach used in Conti and Pudney (2001) in their comparison of job satisfaction responses across different interview modes. Given the nested structure of our data – assessments across time by individuals within households – we model the categorical outcome variable Y_{ijk} using a three-level mixed-effects multinomial logit model, where i indexes individuals, j indexes households, and k indexes time. Variable Y_{ijk} takes the value 1 if an individual reports worse health in the SCQ, the value 2 if an individual reports his/her health consistently in the FFI and SCQ, and the value 3 if an individual reports better health in the SCQ. In this model, the probability of choosing the base (second) alternative equals:

$$\Pr\left(Y_{ijk} = 2 | X_{ijk}, \eta_{ijm}, \mu_{jm}\right) = \frac{1}{1+\sum_{l=1,3}\exp(X'_{ijk}\beta_l+\eta_{ijl}+\mu_{jl})}. \quad (1)$$

where X_{ijk} includes a set of individual characteristics, η_{ijm} is an individual-level effect and μ_{jm} is a household-level effect. Individual-level and household-level effects are assumed to be normally distributed. The probabilities of choosing the other alternatives equal:

$$\Pr\left(Y_{ijk} = m | X_{ijk}, \eta_{ijm}, \mu_{jm}\right) = \frac{\exp(X'_{ijk}\beta_m+\eta_{ijm}+\mu_{jm})}{1+\sum_{l=1,3}\exp(X'_{ijk}\beta_l+\eta_{ijl}+\mu_{jl})}\ , m = 1,3. \quad (2)$$

The three-level mixed-effects multinomial logit model is estimated using the mean-and-variance adaptive 10-point Gauss–Hermite quadrature approximation, because there are no analytical formulas of the likelihood function. We use Stata’s *gsem* command to implement the estimation (see the Stata manual for technical details about the numerical integration method). In the baseline model (presented in Table 3), we allow the individual and household effects to vary across response categories. In the remaining models we restrict these effects to be the same across categories in order to help achieve convergence. For all models we present average partial effects, which are interpreted as the average change in the probabilities of the three response categories associated with a change in the individual characteristic. Ethics approval is not required for this research as it uses secondary data from the HILDA Survey.

5. Explaining Health Reporting Inconsistency

We model the FFI and SCQ combination of responses ($H_{FFI} > H_{SCQ}$, $H_{FFI} = H_{SCQ}$, $H_{FFI} < H_{SCQ}$) using two samples: (1) all respondents in waves 1, 9 and 13; and (2) respondents in waves 9 and 13, who completed the FFI and SCQ within 30 days of one another. Sample (2) (which excludes 2.6% of wave 9 and 13 respondents) allows the influence of days between SAH questions on response consistency to be examined while balancing the need to ensure that the underlying health of respondents remains unchanged between questions.

In comparing the coefficient estimates from the two samples (shown in Table 3), it is clear that our main findings are robust to the sample used, and therefore for all remaining discussions we focus on the results based on Sample (2). In Appendix Figure A1 [INSERT LINK TO ONLINE FILE A], we further demonstrate that the results are robust to the number of days between surveys. The graphs show that estimated average partial effects for the most important covariates are insensitive to whether the estimation sample includes questionnaires that were completed 1 day, 10 days, 20 days, 30 days or 40 days apart.

5.1. Main Results

Table 3 presents estimated average partial effects from the three-level mixed-effects multinomial logit model defined in Section 4. The variances and the Intraclass Correlation Coefficients (ICC) of the random effects, provided in Appendix Table A1 [INSERT LINK TO ONLINE FILE A], justify our modelling approach and show that it is important to account for clustering at the individual and household level. The ICCs show both a moderate degree of persistence in respondents' reporting inconsistency ($ICC_{\eta 1} = 0.063$ and $ICC_{\eta 3} = 0.020$) and clustering of reporting behaviour across household members ($ICC_{\mu 1} = 0.052$ and $ICC_{\mu 3} = 0.121$). Nevertheless, the majority of variation in inconsistent reporting arises from differences across individuals within households and within a particular survey wave.

The main results reveal several important findings. First, cognitive ability is a strong predictor of inconsistency, with low ability respondents significantly more likely to both report better health in the FFI than in the SCQ ($H_{FFI} > H_{SCQ}$) and significantly more likely to report worse health in the FFI than in the SCQ ($H_{FFI} < H_{SCQ}$), compared to high ability respondents. In addition, the two average partial effects are similar in magnitude (-1.2 and -1.6 percentage points). In other words, low ability respondents more frequently switch responses, and in both directions. To further elucidate this important association, we re-fitted the model with binary variables representing deciles of the cognitive ability distribution rather than one continuous variable. The results are displayed in Figure 4. The graphs show that respondents in the tenth decile (highest cognitive ability) are significantly less likely to report both worse health in the SCQ (by 4.2 percentage points) and better health in the SCQ (by 6.1 percentage points) compared with respondents in the first decile (lowest cognitive ability). Figure 4 also highlights that cognitive ability has an increasing effect throughout its distribution, rather than having a threshold after which cognitive ability is unimportant. Results from a model that separately include test scores in memory, cognitive function and verbal skills show that all aspects have similar estimated associations with inconsistency (results available upon request).

Related to the finding on cognitive ability is the substantive association between inconsistency and a university degree level education. In particular, university graduates are more likely to report their health consistently (by 3.8 percentage points), compared with high-school dropouts. Naturally, this association is much stronger if cognitive ability is omitted from the model. Overall, the cognitive ability and university degree estimates indicate that the cognitive demands in accurately answering self-assessed health questions must be reasonably high (in at least one of the survey modes).

There is evidence that women are more likely to report their health consistently than men (by 1.4 percentage points). Another significant predictor of inconsistency is age, however, unlike cognitive ability and education, the relationship is not symmetric. Respondents aged 25-34 years are considerably more likely to report worse health in the FFI than in the SCQ: the likelihood of $H_{FFI} < H_{SCQ}$ is 3.2 percentage point higher than for 45-54 year olds. Respondents aged 35-44 years have a similar tendency (by 2.5 percentage points). Interestingly, the pattern is reversed at older ages, with respondents aged 75 years and over 4.4 percentage points more likely to report better health in the FFI than in the SCQ. A conceivable explanation for these results is social desirability bias. Younger respondents may avoid declaring to an interviewer that their health is excellent, so as not to project over-confidence; but in the confidential SCQ feel comfortable doing so. For older respondents, relatively good health might be a marker of status, and so they may overstate their health during the FFI. An alternative explanation is that older people feel less comfortable in revealing medical problems to an interviewer.

Generally, we find weak associations between the set of employment status variables and inconsistency. The justification hypothesis asserts that individuals without a paid job may overstate their disability or health problems in order to justify their non-employment status, implying in our context that non-employed respondents would be more likely to report worse health in the FFI than in the SCQ. We find marginally significant effects for retired respondents, but the results do not generally support the presence of justification bias. Notably, the finding of non-significance holds true when the models are estimated separately by gender, and also when retirement is disaggregated by early and normal-age (≥65 years) retirement.

Importantly, we can test the sensitivity of the Table 3 estimates to the order in which respondents answer the self-assessed health questions. The FFI question is located near the end of the interview after modules on education, employment, income and relationships, while the SCQ question appears at the start of the SCQ. If certain types of respondents are more readily ‘primed’ by prior survey questions, the relationship between certain traits and inconsistency may be stronger or weaker depending upon the ordering. Previous studies have been unable to separate the effects of survey mode from question order effects due to the ordering of modes being identical for all respondents. Estimated average partial effects for the sample that completed the SCQ prior to the FFI and for the sample that completed the SCQ after the FFI are presented in Appendix Table A2 [INSERT LINK TO ONLINE FILE A]. The smaller sample sizes for these two models have increased the standard errors; however, overall the pattern of estimates is similar. In particular, cognitive ability and education remain important predictors for both samples. For example, the estimated average partial effect of cognitive ability on consistent reporting ($H_{FFI} = H_{SCQ}$) equals 2.8 percentage points irrespective whether the SCQ is completed first or second.

5.2 Extended Results

Next we further expand the covariate set presented in Table 3 and re-fit the three-level mixed-effects multinomial logistic model. The aim of this additional analysis is to identify determinants of inconsistency not often considered within the reporting reliability literature and to examine the sensitivity of the Table 3 estimates to the addition of potential confounders. The estimated average partial effects of the additional covariates are presented in Table 4. We find that the average partial effects of age, gender, education, cognitive ability, and employment are insensitive to the inclusion of the additional covariates. The robustness of the results increases our confidence in the conclusion that the variation in responses to SAH questions is substantially higher for respondents with low cognitive ability and low education.

The Table 4 estimates reveal several other characteristics that increase inconsistency in a plausible and roughly symmetric way. Perhaps as expected, being a ‘conscientious’ person reduces inconsistency, with respondents who classify themselves as orderly, systematic, efficient, non-sloppy, and organised significantly less likely to switch responses. Language problems (as reported by the interviewer) increase inconsistency, with a higher probability of reporting $H_{FFI} < H_{SCQ}$ and $H_{FFI} > H_{SCQ}$. Moreover, the extent of inconsistency is considerable (average partial effects equal 3.3 percentage points and 3.9 percentage points), which suggests that more needs to be done in the survey to accommodate language difficulties. In contrast, we do not find that having eyesight or hearing problems are significantly related to inconsistent health reporting. The number of years in the panel, which naturally increases respondent experience and may improve trust in the study, increases consistency. Those reporting to be time-constrained are more likely to report $H_{FFI} < H_{SCQ}$, but not $H_{FFI} > H_{SCQ}$.

A strong predictor of inconsistency is the number of days between responses. The number of days greatly increases the likelihood of reporting better health in the FFI and of reporting better health in the SCQ, with the estimated average partial effects similar whether the FFI occurred before or after the SCQ. This provides additional evidence to suggest that the ordering of the FFI and SCQ does not greatly influence responses. Notably, the estimates are large even when the number of days between questions is few. For example, completing the SCQ 1 or 2 days after the FFI is estimated to increase the likelihood of $H_{FFI} > H_{SCQ}$ by 1.6 percentage points and to increase the likelihood of $H_{FFI} < H_{SCQ}$ by 1.2 percentage points (compared to completing both on the same day). One possibility for these large differences is that proximate contextual effects are important, such as mood, weather, and day of the week. If true, this implies that SAH measurements contain a sizeable random component. Another explanation is that SAH responses are strongly weighted by short-term acute illness that might arise in the time between the two questions, rather than being heavily weighted to chronic health conditions. A final possibility is that two responses on the same day are more consistent because the respondent recalls their initial response and replicates it; so rather than making a ‘new’ health evaluation, respondents are simply endeavouring to be consistent. Such a process would suggest that the true level of reporting heterogeneity in SAH might be larger than is observed in this quasi-experiment.

We find that the presence of others during the FFI is associated with better reported health in the SCQ: a 1.5 percentage point decrease in the likelihood of $H_{FFI} > H_{SCQ}$ and a 1.4 percentage point increase in the likelihood of $H_{FFI} < H_{SCQ}$. In other words, in the non-confidential FFI respondents tend to report worse health than in the confidential SCQ. These results are consistent with the findings of Conti and Pudney (2011), who show that the presence of others influences how individuals answer questions about their job satisfaction.

Three covariates that do not have statistically significant effects are the percent of unanswered questions in the SAH, the indicator for being suspicious about the survey or uncooperative, and the number of calls required to organise the interview. These variables were added to capture the willingness and enthusiasm of the respondent to fully participate in the survey.

6. Discussion and Conclusion

Self-assessed measures of health are widely collected in survey data and regularly used by researchers across many disciplines to study the extent of socioeconomic inequalities in health. This use of SAH is often justified because it has been found by many studies to be a significant predictor of mortality (Idler and Benyamini, 1997; Jylhä, 2009). However, despite the popularity, there remains considerable debate about what SAH actually measures (see, for example, Dowd and Zajacova, 2010; Gunasekara et al., 2012; Layes et al. 2012; Suziedelyte and Johar, 2013; Au and Johnston, 2014). Important issues are the existence of systematic differences in how individuals assess their health and report it on an ordinal scale, and how this reporting heterogeneity could obscure true health inequalities (Kerkhofs and Lindeboom, 1995; Lindeboom and Van Doorslaer, 2004; Etilé and Milcent, 2006; Bago d'Uva et al., 2011; Layes et al., 2012; Greene et al. 2015). For example, Bago d'Uva et al. (2011) find no socioeconomic inequality in visits to the doctor when SAH is used to measure health need, but after correcting SAH for reporting heterogeneity, that inequality favours the more-educated.

In this paper we build upon previous studies, most notably Crossley and Kennedy (2002) and Clarke and Ryan (2006), by providing new insights into SAH reporting behaviour. We document the consistency of a nationally representative sample of survey respondents in reporting their own health, and identify the characteristics and traits of those individuals who provide inconsistent reports. A rare feature of our data means that the same individual is asked to rate his or her health twice in close temporal proximity. The two responses are elicited through two survey modes: face-to-face interview (FFI) and self-completion questionnaire (SCQ), and nearly half of respondents provide both their health assessments on the same day. We also exploit the multilevel structure of our data and provide evidence of modest persistence in respondents' reporting inconsistency and clustering of reporting behaviour across household members.

We find a number of salient findings that have implications for the use of SAH by researchers. First, there is considerable inconsistency in reporting SAH between the two survey modes, with 57% of respondents reporting inconsistently at least once across the three waves of panel data. This indicates that there is considerable uncertainty and measurement error in individual self-assessments of health. Second, we find that age, education and cognitive ability, are each strong predictors of reporting inconsistency. In particular, respondents with a university degree and high ability respondents are significantly less likely to report their health inconsistently. These findings are the first that we are aware of to demonstrate the substantive cognitive demands in reporting SAH reliably. Although a seemingly simple question, SAH involves an "active cognitive process that is not guided by formal, agreed rules or definitions" (Jylhä, 2009, p.308), and individuals must not only recognise the meaning of 'health', but identify and evaluate components of their own health status, and decide which response option best summarises it. The higher level of inconsistency in SAH responses among low ability individuals is consistent with Krosnick (1991), who proposed that individuals *satisfice* when coping with the cognitive demands of survey questions. Third, we find that respondents' personality is a significant predictor of reporting consistency, with those who are 'conscientious' being significantly less likely to give inconsistent SAH reports. Similarly, our results reveal that individuals who are 'time-poor' provide more inconsistent SAH responses, and that building the trust of respondents in the survey improves reporting consistency. Fourth, while inconsistent reporting is substantial for individuals completing the FFI and SCQ questions on the same day, the number of days between responses greatly increases the likelihood of inconsistency. This is consistent with previous research which finds that transitory factors, such as mood and vitality, are key drivers of SAH (Au and Johnston, 2014). Finally, our analysis highlights the importance of language problems in reducing the quality of survey data, suggesting that more needs to be done to accommodate such respondents in surveys.

Overall, our findings demonstrate that there is considerable measurement error in SAH responses and that this error systematically differs by individual characteristics, including education, cognitive ability and personality. This non-classical measurement error has the potential to be problematic for analyses that use SAH as either an independent or dependent variable, because it can lead to bias in key estimates, with the direction of the bias being unclear (Hyslop and Imbens, 2001). We therefore provide some additional support for the advice of Layes et al. (2012) to researchers that, "For this popular measure to continue to play an important role in population health research and policy development, its users must acknowledge and understand the determinants of self-rated health, including reporting behaviour." Moreover, it is important that researchers explicitly recognise the limitations of commonly used SAH questions, are particularly careful when interpreting the estimated effects of SAH, and highlight possible biases caused by measurement error.

In addition, it is important that future studies continue to explore the cognitive demands placed on respondents when they are asked to evaluate their health. This type of methodological research can aid the development of less cognitively demanding SAH questions. Other key principals of survey design may also be revealed. For instance, consistently priming respondents with more specific health questions or with a vignette of a hypothetical individual's health and their SAH rating, may be a cost-effective strategy to reduce cognitive demands, and consequently the estimation biases associated with measurement error.

We also leave to future research the important issue of whether SAH is better elicited through face-to-face interviews or self-completion questionnaires. We do not have any clinical or objective measures of contemporaneous health that allow us to examine whether the responses given in the FFI or the SCQ are closer to true underlying health. Ideally, future experiments will randomise survey mode among respondents, correlating the obtained self-assessments with objective measurements. Such studies will be better placed to comment on the optimal survey mode for different types of health questions.

Our study has several other limitations. While we have shown strong correlations between reporting inconsistency and individuals' characteristics and traits, most notably education and cognitive ability, we are not able to rule out within our modelling framework that both these characteristics and reporting consistently are determined by other unobserved factors. Nor are we able to identify whether proximate contextual effects, or the incidence of short-term acute illness, differ by education or cognitive ability, which could explain some of the high reporting inconsistency we find for low educated and low ability respondents. It is also the case that our cognitive ability measure does not cover all cognitive aspects, and is only collected in one wave of data.

References

Anderson, K. H. and R. V. Burkhauser (1985). "The retirement-health nexus: A new measure of an old puzzle." Journal of Human Resources **20**(3).

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

Bago d'Uva, T., M. Lindeboom, O. O'Donnell and E. van Doorslaer (2011). "Education-related inequity in healthcare with heterogeneous reporting of health." Journal of the Royal Statistical Society Series a-Statistics in Society **174**(3): 639-664.

Bertrand, M. and S. Mullainathan (2001). "Do people mean what they say? Implications for subjective survey data." American Economic Review **91**(2): 67-72.

Bijwaard, G. E., H. van Kippersluis and J. Veenman (2015). "Education and health: the role of cognitive ability." Journal of Health Economics: **42**, 29–43.

Bound, J. (1991). "Self-reported versus objective measures of health in retirement models." Journal of Human Resources **26**(1): 106-138.

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

Bowling, A. and J. Windsor (2008). "The effects of question order and response-choice on self-rated health status in the English Longitudinal Study of Ageing (ELSA)." Journal of epidemiology and community health **62**(1): 81-85.

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

Conti, Gabriella, James Heckman, and Sergio Urzua (2010) "The education-health gradient." The American Economic Review **100**(2): 234-238.

Conti, G. and S. Pudney (2011). "Survey design and the analysis of satisfaction." Review of Economics and Statistics **93**(3): 1087-1093.

Contoyannis, P., A. M. Jones and N. Rice (2004). "The dynamics of health in the British Household Panel Survey." Journal of Applied Econometrics **19**(4): 473-503.

Crossley, T. F. and S. Kennedy (2002). "The reliability of self-assessed health status." Journal of Health Economics **21**(4): 643-658.

Cutler, D. M. and A. Lleras-Muney (2010). "Understanding differences in health behaviors by education." Journal of Health Economics **29**(1): 1-28.

Dowd, J. B. and A. Zajacova. (2010). "Does self-rated health mean the same thing across socioeconomic groups? Evidence from biomarker data". Annals of Epidemiology **20**(10): 743-749.

Etilé, F. and C. Milcent (2006). "Income-related reporting heterogeneity in self-assessed health: evidence from France." Health Economics **15**(9): 965-981.

Groot, W. (2000). "Adaptation and scale of reference bias in self-assessments of quality of life." Journal of Health Economics **19**(3): 403-420.

Greene, W. H., Harris, M. N. and B. Hollingsworth. (2015). Inflated responses in measures of self-assessed health. American Journal of Health Economics **1**(4): 461-493.

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

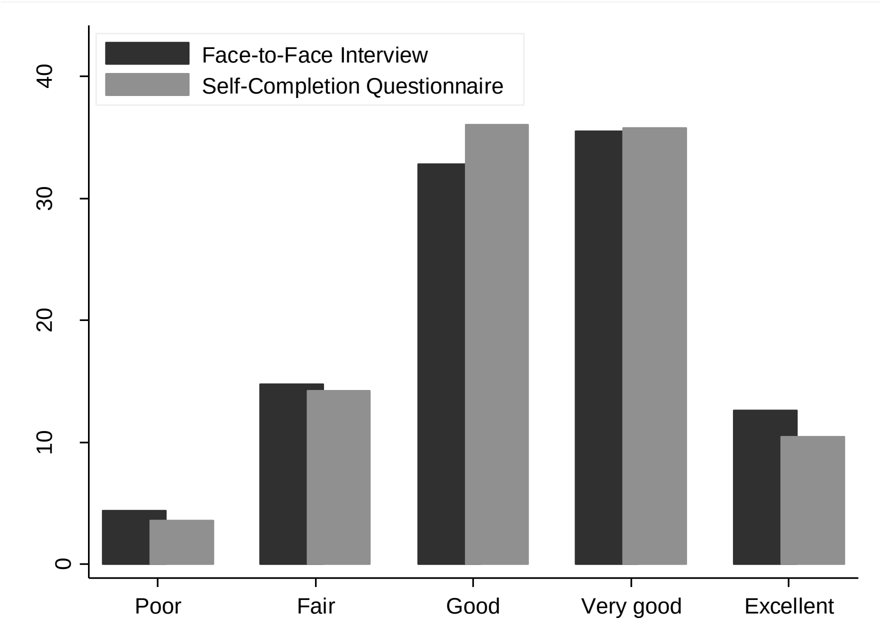
Holford, A. and S. Pudney (2015). "Survey design and the determinants of subjective wellbeing: an experimental analysis." IZA DP No. 8760.

Hu, Y. and F. J. van Lenthe et al (2016). "Trends in socioeconomic inequalities in self-assessed health in 17 European countries between 1990 and 2010." Journal of Epidemiology and Community Health, forthcoming.

Hyslop, R. and G.W. Imbens (2001). "Bias from classical and other forms of measurement error." Journal of Business and Economic Statistics 19(4): 475-481.

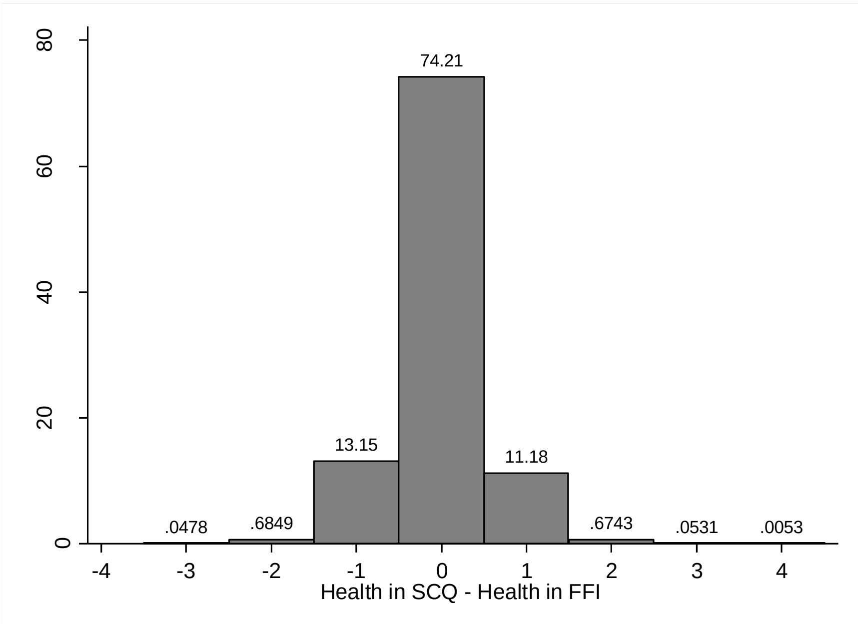
Idler, E. L. and Y. Benyamini. (1997). "Self-rated health and mortality: A review of twenty-seven community studies." Journal of Health and Social Behavior **38**: 21-37.
Johnston, D. W, C. Propper and M.A. Shields (2009). "Comparing subjective and objective health measures: Implications from hypertension for the estimated income/health gradient." Journal of Health Economics **28**: 540-552.
Johnson, R. C. (2010). "The health returns of education policies from preschool to high School and beyond." The American Economic Review **100**(2): 188-194.
Jylhä, M. (2009). "What is self-rated health and why does it predict mortality? Towards a unified conceptual model." Social Science and Medicine **69**(3): 307-316.
Kerkhofs, M. and M. Lindeboom (1995). "Subjective health measures and state dependent reporting errors." Health Economics **4**(3): 221-235.
Krosnick, J. A. (1991). "Response strategies for coping with the cognitive demands of attitude measures in surveys". Applied Cognitive Psychology **5**(3): 213-236.
Kunst, A. E. and V. Bos. et al (2005). "Trends in socioeconomic inequalities in self-assessed health in 10 European countries." International Journal of Epidemiology **34**(2): 295-305.
Layes, A., Y. Asada and G. Kepart. (2012). "Whiners and deniers - What does self-rated health measure?" Social Science and Medicine **75**(1): 1-9.
Lindeboom, M. and M. Kerkhofs (2009). "Health and work of the elderly: subjective health measures, reporting errors and endogeneity in the relationship between health and work." Journal of Applied Econometrics **24**(6): 1024-1046.
Lindeboom, M. and E. Van Doorslaer (2004). "Cut-point shift and index shift in self-reported health." Journal of Health Economics **23**(6): 1083-1099.
Lumsdaine, R. L. and A. Exterkate (2013). "How survey design affects self-assessed health responses in the Survey of Health, Ageing, and Retirement in Europe (SHARE)." European Economic Review **63**: 299-307.
Meer, J., D. L. Miller and H. S. Rosen (2003). "Exploring the health-wealth nexus." Journal of Health Economics **22**(5): 713-730.
Serper, M., R. E. Patzer, M. S. Curtis, S. G. Smith, R. O'Conor, D. W. Baker and M. S. Wolfe (2014). "Health literacy, cognitive ability, and functional health status among older adults." Health Service Research **49**(4): 1249-1267.
Summerfield, M., S. Freidin, M. Hahn, et al. (2012). "HILDA User Manual – Release 1." Melbourne Institute of Applied Economic and Social Research. University of Melbourne.
Suziedelyte, A. and M. Johar (2013). "Can you trust survey responses? Evidence using objective health measures." Economic Letters **121**(2): 163-166.
Viinikainen, J., K. Kokko, L. Pulkkinen and J. Pehkonon (2010). "Personality and labour market income: Evidence from longitudinal data." Labour **24**(2): 201-220.
Wiseman, V. L. (1999). "Culture, self-rated health and resource allocation decision-making." Health Care Analysis **7**(3): 207-223.
Wooden, M. (2013). "The measurement of cognitive ability in wave 12 of the HILDA survey". HILDA Survey Discussion Paper Series No. 1/13. Melbourne Institute of Applied Economic Research, University of Melbourne.

Figure 1: Distributions of Self-Assessed Health from the Face-to-Face Interview and Self-Completion Questionnaire



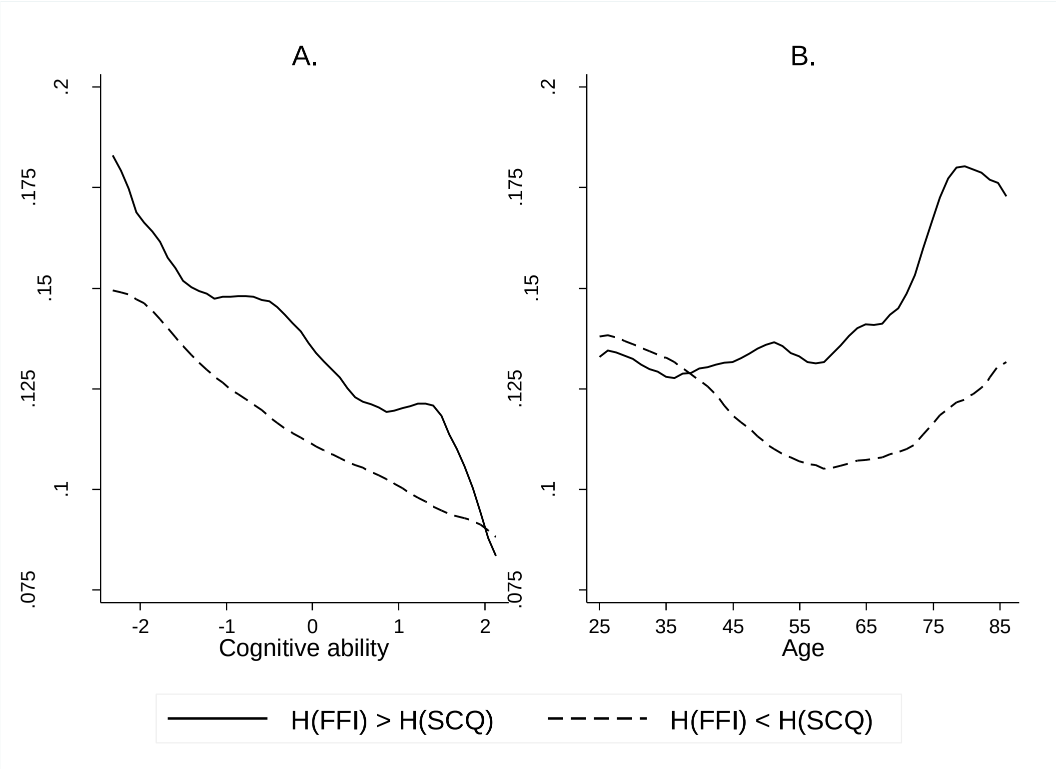
Notes: Sample consists of wave 9 and 13 HILDA respondents aged 25 years or more. Sample size equals 18,834 individual-year observations.

Figure 2: Distribution of the Differences in Self-Assessed Health from the Same Survey Wave



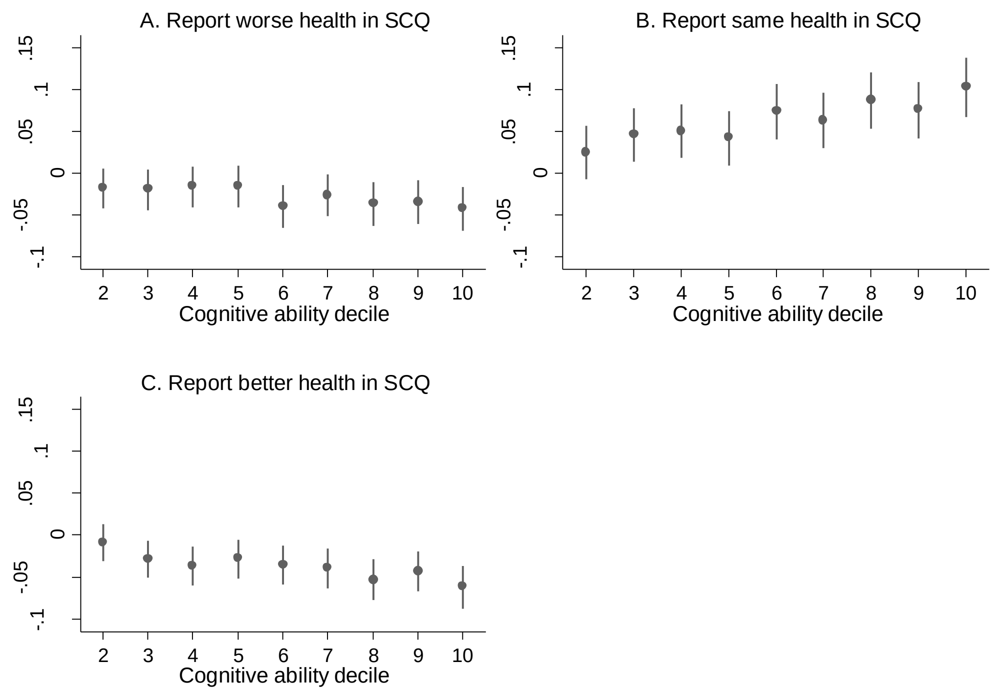
Notes: Sample consists of wave 9 and 13 HILDA respondents aged 25 years or more. Sample size equals 18,834 individual-year observations.

Figure 3: Non-Parametric Regression Estimates of the Relationship between Inconsistent Reporting and Cognitive Ability and Age



Notes: Sample consists of wave 9 and 13 HILDA respondents aged 25 years of age or more. Sample sizes in graphs A and B are 15,597 and 18,656, respectively. Bottom 1% and top 1% of cognitive ability distribution and top 1% of age distribution are excluded.

Figure 4: Estimated Average Partial Effects of Cognitive Ability Deciles



Notes: Sample size is 18,834. Regression also controls for gender, 10-year age categories, marital status, number of children, educational attainment, employment status, country of birth, foreign language, and year effects. The dots represent average partial effects from multinomial logit model with random individual and household effects, and the vertical lines represent 95% confidence intervals. The 1st cognitive ability factor decile is the omitted category.

Table 1: Variable Descriptions and Summary Statistics

Variable	Description	Mean (S.D.)
Basic Controls		
Male	= 1 if male	0.470
25-34 years old	= 1 if aged 25-34 years	0.190
35-44 years old	= 1 if aged 35-44 years	0.204
55-64 years old	= 1 if aged 55-64 years	0.178
65-74 years old	= 1 if aged 65-74 years	0.127
>=75 years old	= 1 if aged 75 years or more	0.083
Married/Partnered	= 1 if married or living with partner	0.735
No of children	Number of children under 15 years of age	0.542 (0.959)
Cognitive ability	Factor of 3 cognitive test scores: memory, cognitive function, and verbal skills	0.032 (0.992)
High school	= 1 if has high school degree	0.111
Vocational qual.	= 1 if has vocational qualification	0.335
University degree	= 1 if has university degree	0.272
Unemployed	= 1 if unemployed or marginally attached to labour market	0.073
OLF non-retired	= 1 if out of labour force, but not retired	0.060
OLF retired	= 1 if out of labour force and retired	0.230
Foreign born: ES	= 1 if born in foreign English speaking country	0.115
Foreign born: NES	= 1 if born in foreign non-English speaking country	0.122
Foreign language	= 1 if speaks foreign language at home	0.099
Extended Controls		
Conscientiousness	Average of 6 traits (orderly, systematic, inefficient (reversed), sloppy (reversed), disorganised (reversed), and efficient, each measured on 7 point scale) across waves 5, 9, and 13	5.180 (0.932)
Always/often pressed for time	= 1 if always or often feels pressed for time or rushed	0.361
Suspicious/uncooperative	= 1 if suspicious of study or cooperation fair or poor	0.009
Language problems	=1 if language problems during interview or interview completed with the assistance of an interpreter or family member	0.026
Eyesight/hearing problems	= 1 if eyesight, hearing, or reading problems	0.018
Years in panel	Years in HILDA panel	9.164 (3.732)
Percent of SCQ unanswered/10	Percent of SCQ questions left unanswered (divided by 10)	0.132 (0.419)
4-6 calls to complete interviews	Took 4-6 calls to the HH to complete all interviews	0.352
>6 calls to complete interviews	Took 6 or more calls to the HH to complete all interviews	0.125
Others present	= 1 if others were present during FFI	0.380
SCQ >= 8 days before FFI	= 1 if SCQ completed 8 days before FFI or earlier	0.020
SCQ 3-7 days before FFI	= 1 if SCQ completed 3-7 days before FFI	0.042
SCQ 1-2 days before FFI	= 1 if SCQ completed 1-2 days before FFI	0.084
SCQ 1-2 days after FFI	= 1 if SCQ completed 1-2 days after FFI	0.181
SCQ 3-7 days after FFI	= 1 if SCQ completed 3-7 days after FFI	0.134
SCQ >=8 days after FFI	= 1 if SCQ completed 8 days after FFI or later	0.071

Notes: Sample consists of wave 9 and 13 HILDA respondents aged 25 years of age or more. Students and respondents who completed the personal interview by phone are not included in the sample. Sample size is 18,834. The omitted category for age, education, employment, country of birth, number of calls, and day difference between SCQ and FFI is “45-54 years old”, “Less than high school”, “Employed”, “Born in Australia”, “1-3 calls”, and “SCQ and FFI on same day”, respectively.

Table 2: Reporting of Self-Assessed Health in Face-to-Face and Self-Completion Questionnaires

FFI\SCQ	Poor	Fair	Good	Very good	Excellent	Total
Poor	547	240	30	8	1	826
	[66.22]	[29.06]	[3.63]	[0.97]	[0.12]	[100.00]
Fair	100	1,936	680	59	2	2,777
	[3.60]	[69.72]	[24.49]	[2.12]	[0.07]	[100.00]
Good	21	460	4,782	874	38	6,175
	[0.34]	[7.45]	[77.44]	[14.15]	[0.62]	[100.00]
Very good	3	42	1223	5,102	311	6,681
	[0.04]	[0.63]	[18.31]	[76.37]	[4.65]	[100.00]
Excellent	0	6	66	693	1610	2375
	[0.00]	[0.25]	[2.78]	[29.18]	[67.79]	[100.00]
Total	671	2,684	6,781	6,736	1962	18,834
	[3.56]	[14.25]	[36.00]	[35.77]	[10.42]	[100.00]

Notes: Sample consists of wave 9 and 13 HILDA respondents aged 25 years of age or more. Row percentages are reported in the brackets.

Table 3: Predictors of Inconsistency: Average Partial Effects from Three-Level Mixed-Effects Multinomial Logistic Models

	Waves 1, 9 and 13			Waves 9 and 13		
	$H_{FFI} > H_{SSQ}$	$H_{FFI} = H_{SSQ}$	$H_{FFI} < H_{SSQ}$	$H_{FFI} > H_{SSQ}$	$H_{FFI} = H_{SSQ}$	$H_{FFI} < H_{SSQ}$
Male	-0.004	-0.008	0.012***	0.007	-0.014**	0.007
	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)
25-34 years old	-0.001	-0.019**	0.020***	-0.003	-0.029***	0.032***
	(0.007)	(0.008)	(0.006)	(0.008)	(0.010)	(0.007)
35-44 years old	-0.007	-0.005	0.011**	-0.007	-0.018*	0.025***
	(0.007)	(0.008)	(0.006)	(0.008)	(0.010)	(0.007)
55-64 years old	0.003	0.007	-0.010	-0.003	0.010	-0.007
	(0.007)	(0.009)	(0.006)	(0.008)	(0.010)	(0.008)
65-74 years old	0.030***	-0.013	-0.018**	0.012	0.011	-0.023**
	(0.009)	(0.011)	(0.008)	(0.011)	(0.014)	(0.010)
>=75 years old	0.051***	-0.040***	-0.011	0.044***	-0.027*	-0.017
	(0.010)	(0.013)	(0.009)	(0.012)	(0.016)	(0.012)
Married/Partnered	0.006	0.003	-0.009**	-0.001	0.012	-0.011**
	(0.005)	(0.006)	(0.004)	(0.006)	(0.007)	(0.005)
No of children	0.005*	0.001	-0.005**	0.003	-0.000	-0.003
	(0.002)	(0.003)	(0.002)	(0.003)	(0.004)	(0.003)
Cognitive skills	-0.012***	0.029***	-0.016***	-0.012***	0.028***	-0.016***
	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)	(0.003)
High school	-0.014*	0.019**	-0.005	-0.016*	0.017	-0.001
	(0.007)	(0.009)	(0.006)	(0.009)	(0.011)	(0.008)
Vocational qual.	-0.015***	0.019***	-0.004	-0.016**	0.019**	-0.002
	(0.005)	(0.006)	(0.005)	(0.006)	(0.008)	(0.006)
University degree	-0.035***	0.047***	-0.012**	-0.027***	0.038***	-0.011
	(0.006)	(0.008)	(0.005)	(0.008)	(0.010)	(0.007)
Unemployed	0.002	-0.008	0.006	0.002	-0.007	0.005
	(0.008)	(0.009)	(0.007)	(0.010)	(0.012)	(0.009)
OLF non-retired	-0.013	-0.000	0.013*	-0.006	-0.000	0.006
	(0.008)	(0.010)	(0.007)	(0.011)	(0.014)	(0.010)
OLF retired	-0.022***	0.010	0.012*	-0.019**	0.002	0.016*
	(0.008)	(0.009)	(0.007)	(0.009)	(0.012)	(0.009)
Foreign born: ES	0.014**	-0.013	-0.002	0.014*	-0.011	-0.003
	(0.006)	(0.008)	(0.006)	(0.008)	(0.010)	(0.007)
Foreign born: NES	0.023***	-0.040***	0.017**	0.005	-0.020	0.015
	(0.008)	(0.010)	(0.007)	(0.010)	(0.013)	(0.009)
Foreign language	-0.002	-0.008	0.010	0.003	-0.011	0.007
	(0.009)	(0.011)	(0.007)	(0.011)	(0.014)	(0.010)
Sample size	30,452			18,834		

Notes: The presented figures are average partial effects from multinomial logit model with random individual and household effects. Standard errors (clustered at household level) are presented in parentheses. The omitted category for age, education, employment, and country of birth is "45-54 years old", "Less than high school", "Employed", and "Born in Australia", respectively. Regressions also control for missing cognitive ability and year effects. *, ** and *** denote statistical significance at the 0.10, 0.05 and 0.01 level, respectively.

Table 4: Average Partial Effects from an Expanded Multinomial Logistic Model

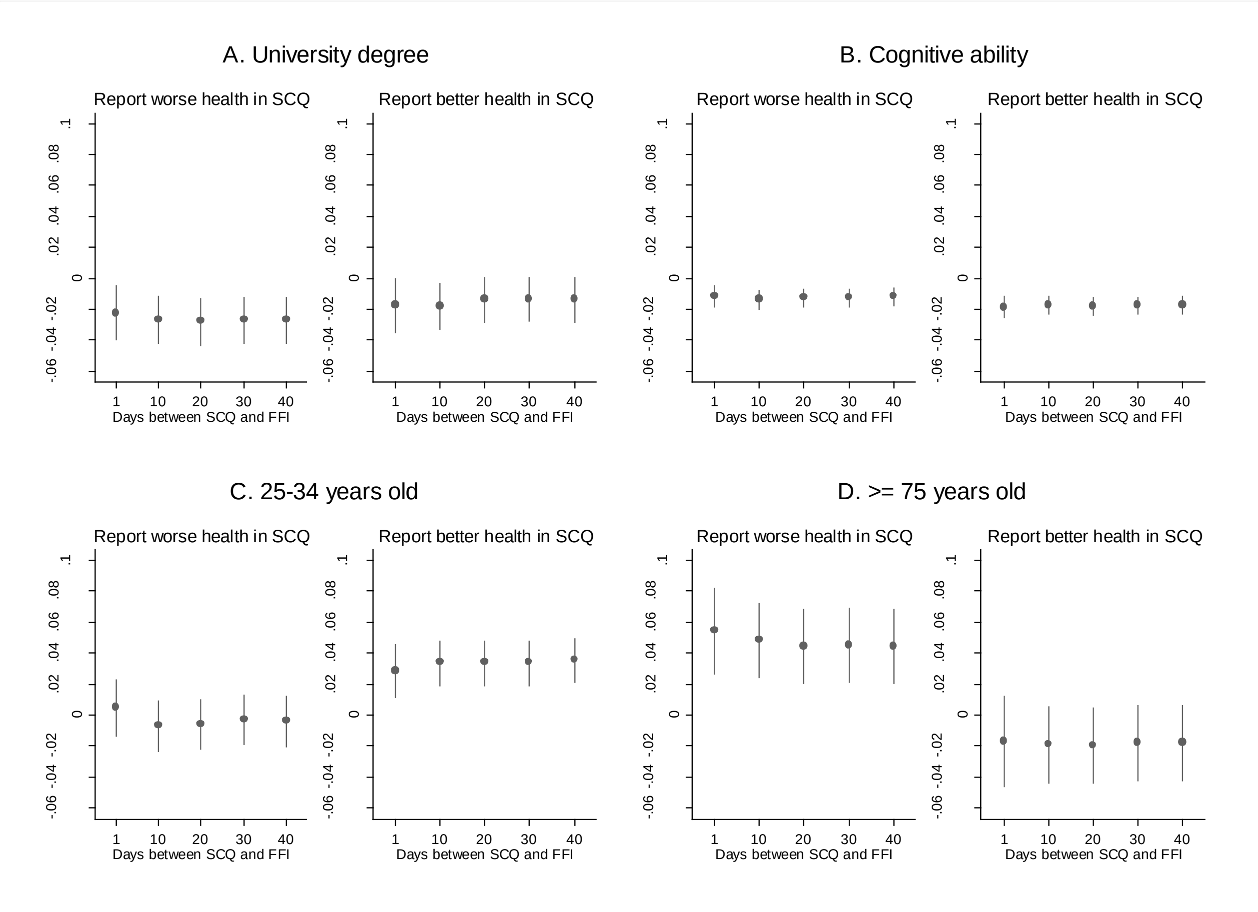
	$H_{FFI} > H_{SSQ}$		$H_{FFI} = H_{SSQ}$		$H_{FFI} < H_{SSQ}$	
Male	0.005	(0.005)	-0.012*	(0.007)	0.007	(0.005)
25-34 years old	-0.003	(0.009)	-0.023**	(0.011)	0.026***	(0.008)
35-44 years old	-0.006	(0.008)	-0.013	(0.011)	0.019**	(0.008)
55-64 years old	-0.003	(0.009)	0.007	(0.011)	-0.004	(0.008)
65-74 years old	0.017	(0.011)	0.001	(0.015)	-0.018	(0.011)
>=75 years old	0.048***	(0.013)	-0.038**	(0.018)	-0.010	(0.014)
Married/Partnered	0.008	(0.006)	0.014*	(0.008)	-0.022***	(0.006)
No of children	0.002	(0.003)	0.000	(0.004)	-0.002	(0.003)
Cognitive skills	-0.011***	(0.003)	0.027***	(0.004)	-0.016***	(0.003)
High school	-0.015	(0.009)	0.013	(0.012)	0.002	(0.009)
Vocational qual.	-0.018***	(0.007)	0.020**	(0.009)	-0.002	(0.006)
University degree	-0.028***	(0.008)	0.039***	(0.010)	-0.011	(0.008)
Unemployed/OLF attached	0.003	(0.010)	-0.006	(0.013)	0.003	(0.009)
OLF non-retired	-0.006	(0.011)	0.002	(0.015)	0.005	(0.011)
OLF retired	-0.016*	(0.009)	0.003	(0.013)	0.014	(0.009)
Foreign born: ES	0.015*	(0.008)	-0.012	(0.011)	-0.003	(0.008)
Foreign born: NES	-0.000	(0.011)	-0.011	(0.014)	0.011	(0.010)
Foreign language	-0.005	(0.012)	0.005	(0.016)	-0.000	(0.011)
Conscientiousness	-0.009***	(0.003)	0.013***	(0.004)	-0.003	(0.003)
Always/often pressed for time	-0.001	(0.006)	-0.011	(0.007)	0.013**	(0.005)
Suspicious/uncooperative	0.025	(0.025)	-0.018	(0.035)	-0.007	(0.026)
Language problems	0.033**	(0.016)	-0.072***	(0.021)	0.039***	(0.014)
Eyesight/hearing problems	0.019	(0.017)	-0.032	(0.031)	0.013	(0.025)
Years in panel	-0.000	(0.001)	0.003***	(0.001)	-0.003***	(0.001)
Percent non-answered/10	0.010*	(0.006)	-0.006	(0.008)	-0.004	(0.006)
4-6 calls to complete interviews	0.007	(0.006)	-0.012*	(0.007)	0.005	(0.005)
>6 calls to complete interviews	-0.009	(0.008)	0.003	(0.011)	0.006	(0.008)
Others present	-0.015***	(0.006)	0.001	(0.007)	0.014***	(0.005)
SCQ >= 8 days before FFI	0.084***	(0.016)	-0.120***	(0.022)	0.036**	(0.016)
SCQ 3-7 days before FFI	0.073***	(0.012)	-0.098***	(0.016)	0.025**	(0.012)
SCQ 1-2 days before FFI	0.034***	(0.010)	-0.048***	(0.013)	0.015*	(0.009)
SCQ 1-2 days after FFI	0.016**	(0.007)	-0.028***	(0.009)	0.012*	(0.007)
SCQ 3-7 days after FFI	0.049***	(0.008)	-0.064***	(0.010)	0.015**	(0.007)
SCQ >=8 days after FFI	0.078***	(0.010)	-0.108***	(0.013)	0.031***	(0.009)

Notes: Sample size equals 17,964. The presented figures are average partial effects from multinomial logit model with random individual and household effects. Sample includes waves 9 and 13. Standard errors (clustered at household level) are presented in parentheses. The omitted category for age, education, employment, country of birth, number of calls and day difference between SCQ and FFI is "45-54 years old", "Less than high school", "Employed", "Born in Australia", "1-3 calls" and "SCQ and FFI on same day", respectively. Regression controls for missing cognitive ability and year effects. *, ** and *** denote statistical significance at the 0.10, 0.05 and 0.01 level, respectively.

SUPPLEMENTARY ONLINE FILE A

Appendix

Figure A1: Variation in Relationship between Individual Characteristics and Self-Assessed Health: Average partial effects by the absolute difference between FFI and SCQ dates



Notes: The number of individuals who answered both questionnaires within 1, 10, 20, 30, and 40 days is 12150, 17682, 18407, 18834, and 19102, respectively. All regressions also control for gender, marital status, number of children, employment status, country of birth, foreign language, and year effects. The dots represent average partial effects from multinomial logit model with random individual and household effects. The vertical lines represent 95% confidence intervals. Omitted category for the dependent variable is “Reported health consistently in FFI and SCQ”. The omitted category for education and age is “Less than high school” and “45-54 years old”, respectively.

Table A1: Variances and Intraclass Correlation Coefficients from the Three-Level Mixed-Effects Multinomial Logistic Models

	Waves 1, 9 and 13	Waves 9 and 13
Variances		
$var(\mu_{j1})$	0.183	0.192
$var(\eta_{ij1})$	0.068	0.233
$var(\mu_{j3})$	0.289	0.465
$var(\eta_{ij3})$	0.100	0.075
Intraclass Correlation Coefficients		
$ICC_{\mu 1}$	0.052	0.052
$ICC_{\eta 1}$	0.019	0.063
$ICC_{\mu 3}$	0.079	0.121
$ICC_{\eta 3}$	0.027	0.020
Sample size	30,452	18,834

Notes: The presented variances and ICCs come from multinomial logit model with random individual (η_{ijm}) and household (μ_{jm}) effects (m = 1 if $H_{FFI} > H_{SCQ}$, and m = 3 if $H_{FFI} < H_{SCQ}$). The omitted category of the dependent variable is $H_{FFI} = H_{SCQ}$. All regressions control for gender, 10-year age categories, marital status, number of children, cognitive ability, educational attainment, employment status, country of birth, foreign language, and year effects.

Table A2: The Ordering of Survey Responses: Relationship between individual characteristics and Self-Assessed Health

	SCQ before FFI			SCQ after FFI		
	$H_{FFI} > H_{SSQ}$	$H_{FFI} = H_{SSQ}$	$H_{FFI} < H_{SSQ}$	$H_{FFI} > H_{SSQ}$	$H_{FFI} = H_{SSQ}$	$H_{FFI} < H_{SSQ}$
	(1)	(2)	(3)	(4)	(5)	(6)
Male	0.006	-0.030	0.024*	0.005	-0.017	0.012
	(0.015)	(0.019)	(0.013)	(0.009)	(0.011)	(0.008)
25-34 years old	0.019	-0.073**	0.054***	-0.026*	-0.005	0.031**
	(0.024)	(0.030)	(0.020)	(0.014)	(0.019)	(0.013)
35-44 years old	0.035	-0.066**	0.031	-0.016	0.000	0.016
	(0.024)	(0.030)	(0.021)	(0.013)	(0.018)	(0.013)
55-64 years old	-0.021	0.037	-0.016	-0.014	0.021	-0.006
	(0.024)	(0.030)	(0.021)	(0.014)	(0.019)	(0.013)
65-74 years old	-0.022	0.028	-0.006	0.000	0.020	-0.020
	(0.032)	(0.039)	(0.028)	(0.019)	(0.024)	(0.018)
>=75 years old	0.043	-0.018	-0.025	0.010	0.027	-0.037*
	(0.033)	(0.044)	(0.034)	(0.021)	(0.028)	(0.020)
Married/Partnered	-0.008	0.014	-0.006	0.002	0.006	-0.007
	(0.017)	(0.022)	(0.015)	(0.009)	(0.013)	(0.009)
No of children	-0.016*	0.027**	-0.010	0.002	0.001	-0.003
	(0.010)	(0.012)	(0.008)	(0.005)	(0.007)	(0.005)
Cognitive skills	-0.014	0.028**	-0.014*	-0.015***	0.028***	-0.014***
	(0.009)	(0.011)	(0.007)	(0.005)	(0.007)	(0.005)
High school	-0.019	0.033	-0.014	-0.012	0.014	-0.002
	(0.026)	(0.034)	(0.025)	(0.015)	(0.020)	(0.014)
Vocational qual.	-0.053***	0.056**	-0.004	-0.021*	0.039***	-0.018*
	(0.019)	(0.024)	(0.016)	(0.011)	(0.015)	(0.010)
University degree	-0.038*	0.049*	-0.011	-0.032**	0.050***	-0.017
	(0.022)	(0.028)	(0.019)	(0.013)	(0.017)	(0.012)
Unemployed	0.012	-0.020	0.008	0.019	-0.010	-0.009
	(0.026)	(0.034)	(0.024)	(0.017)	(0.023)	(0.017)
OLF non-retired	-0.046	0.053	-0.006	0.003	-0.030	0.027*
	(0.035)	(0.042)	(0.029)	(0.018)	(0.024)	(0.016)
OLF retired	-0.011	0.006	0.005	-0.002	-0.021	0.023
	(0.026)	(0.033)	(0.023)	(0.016)	(0.021)	(0.015)
Foreign born: ES	0.025	0.022	-0.047**	0.022	-0.028	0.006
	(0.022)	(0.029)	(0.023)	(0.013)	(0.018)	(0.013)
Foreign born: NES	0.019	-0.019	0.000	0.011	-0.030	0.018
	(0.030)	(0.036)	(0.023)	(0.017)	(0.023)	(0.016)
Foreign language	-0.005	-0.024	0.029	-0.006	-0.007	0.013
	(0.033)	(0.039)	(0.025)	(0.019)	(0.024)	(0.016)
Sample size	2,726			7,312		

Notes: Respondents who completed SCQ and FFI on the same day are excluded. The presented figures are average partial effects from multinomial logit model with random individual and household effects. Standard errors (clustered at household level) are presented in parentheses. The omitted category for age, education, employment, and country of birth is "45-54 years old", "Less than high school", "Employed", and "Born in Australia", respectively. Regressions also control for year effects. *, ** and *** denote statistical significance at the 0.10, 0.05 and 0.01 level, respectively.