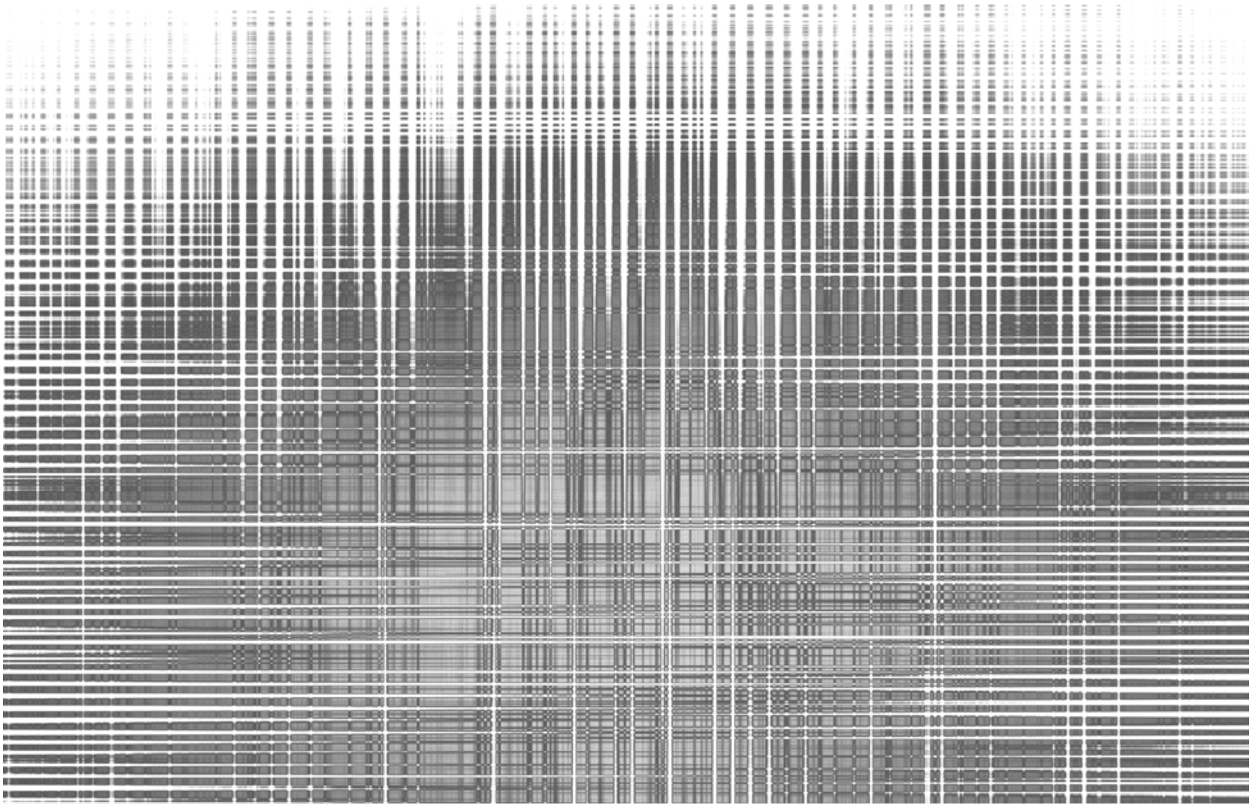
CENTRE FOR SOCIAL
RESEARCH & METHODS



The social determinants of health and subjective wellbeing: a comparison of probability and nonprobability online panels

N Biddle, J Sinibaldi and J Sheppard

CSRM & SRC METHODS PAPER
NO. 6/2018

Series note

The ANU Centre for Social Research & Methods (CSRM) was established in 2015 to provide national leadership in the study of Australian society. CSRM has a strategic focus on:

- development of social research methods
- analysis of social issues and policy
- training in social science methods
- providing access to social scientific data.

CSRM publications are produced to enable widespread discussion and comment, and are available for free download from the CSRM website (http://csrm.cass.anu.edu.au/research/publications).

CSRM is located within the Research School of Social Sciences in the College of Arts & Social Sciences at the Australian National University (ANU). The centre is a joint initiative between the Social Research Centre and the ANU. Its expertise includes quantitative, qualitative and experimental research methodologies; public opinion and behaviour measurement; survey design; data collection and analysis; data archiving and management; and professional education in social research methods.

As with all CSRM publications, the views expressed in this Methods Paper are those of the authors and do not reflect any official CSRM position.

Professor Matthew Gray
Director, ANU Centre for Social Research & Methods
Research School of Social Sciences
College of Arts & Social Sciences
The Australian National University
December 2018

Methods Paper No. 6/2018

ISSN 2209-184X
ISBN 978-1-925715-16-3

An electronic publication downloaded from http://csrm.cass.anu.edu.au/research/publications.

For a complete list of CSRM Methods Papers, see http://csrm.cass.anu.edu.au/research/publications/methods-research-papers.

ANU Centre for Social Research & Methods
Research School of Social Sciences
The Australian National University

The social determinants of health and subjective wellbeing: A comparison of probability and nonprobability online panels

N Biddle, J Sinibaldi and J Sheppard

Nicholas Biddle is Associate Director, ANU Centre for Social Research & Methods, College of Arts and Social Sciences, Australian National University.

Jennifer Sinibaldi is Assistant Research Professor, Joint Program in Survey Methodology, University of Maryland.

Jill Sheppard is a lecturer in the School of Politics and International Relations, College of Arts & Social Sciences, Australian National University.

Abstract

As response rates to surveys decline all over the world, researchers are increasingly turning to sampling frames that are easier and cheaper to reach, and that have more predictable response rates. These include nonprobability web panels (NWPs) and probability web panels (PWPs). Although generally more expensive to construct, the latter have been shown in many instances to suffer from fewer biases and deviation from benchmarks. The literature comparing NWPs with PWPs is fledgling. We add to this research area by comparing measures of the social determinants of health that were estimated from a number of NWPs and PWP equivalents with a high-quality benchmark. The analysis finds that, when looking at the distributions of self-assessed health and life satisfaction, probability panels differ less from the gold standard than do nonprobability panels. This supports previous work, although we also show that this conclusion holds when a greater range of control variables is included in the model. However, some of the predictors of health are captured better using the nonprobability panels. In particular, the relationship between area-level disadvantage and health is better captured through a pooled nonprobability sample.

Acknowledgments

The authors gratefully acknowledge the assistance of Darren Pennay and the Social Research Centre for providing data and valuable advice, and participants at the Current State and Future of Online Research in Australia workshop at the Australian National University on 14 July 2016 for their feedback.

Acronyms

CATI	computer-assisted telephone interview
CSRM	Centre for Social Research & Methods
HILDA	Household, Income and Labour Dynamics in Australia
NWP	nonprobability web panel
OPBS	Online Panels Benchmarking Study
PWP	probability web panel
RDD	random-digit dialling
SEIFA	Socio-Economic Indexes for Areas

Contents

Figures

1 Overview

Increasing nonresponse to probability-based surveys has prompted the question of whether probability-based data collection is worthwhile. Some researchers support the use of nonprobability convenience samples as a legitimate alternative, especially when budgets are small or no complete sampling frame exists (e.g. Heckathorn 1997, Rivers 2013). Specifically, nonprobability data collection has gained momentum in the area of web surveys. Web intercept surveys that rely on convenience sampling rarely provide accurate population values (Baker et al. 2010), and nonprobability web panels may be better positioned to control or correct for error because of the richer auxiliary information obtained during the recruitment survey.

The body of literature assessing the validity of estimates derived from nonprobability web panels (NWPs) compared with probability web panels (PWPs) is growing. With a few exceptions, however, most studies have focused on prevalence estimates for one or more outcomes of interest. While interest in such estimates will continue, researchers and policy makers are often as interested in the relationships between variables, or what predicts the outcomes of interest.

One such area of enquiry is the social determinants of health. Specifically, to what extent does a person's gender, ethnicity, location within or across countries, demography, and socioeconomic position predict their health outcomes? By extending beyond the medical model of disease, the literature on social determinants of health allows governments to better target interventions, and quantify the effect of inequalities within society (Marmot 2005).

The aim of this paper is to contribute to two sets of literature (determinants of health and survey methodology) by evaluating the results from the Online Panels Benchmarking Study (OPBS) (Pennay et al. 2018). The OPBS contained data from three surveys based on probability samples of the Australian population (including one PWP) and five surveys administered to members of nonprobability online panels. This is the first study to compare the predictive relationships found in PWPs and NWPs in Australia, and the only study (as far as the authors are aware) to focus on the social determinants of health more broadly. The results will be of use to researchers of the social determinants of health (those using PWPs or NWPs) as well as data collectors, and will inform the emerging literature on the representativeness of PWP versus NWP data collection.

Section 2 provides some background about existing research on panel comparisons, including a brief summary of the literature on the social determinants of health. Section 3 presents our data and methods, and Section 4 reports results from our analysis. Section 5 provides some concluding comments.

2 Background

2.1 Probability and nonprobability panels

The standard assumption used in much survey data analysis is that we can estimate the probability that an individual is selected into our survey sample. Following the total survey error approach (Groves et al. 2011), there exists a target population of interest from which a sampling frame can be drawn. From this sample frame, we can draw a sample. Differences between the target population and sampling frame are referred to as coverage error, and differences between the sampling frame and sample are referred to as sampling error. Importantly, these errors are assumed to be known or quantifiable.

In many instances, sampling frames do not exist, are too expensive or difficult to access, or cannot be generated in a timely manner. These instances encourage researchers to use nonprobability samples. By definition, coverage and sampling error in such surveys are harder to estimate. NWPs can use detailed quotas or purposive sampling, poststratification or propensity score adjustment, and innovative modelling to develop weights, all of which can significantly improve the accuracy of the resulting statistics (Dever et al. 2008, Lee & Valliant 2009, Valliant & Dever 2011). However, the levels of implementation, disclosure (Baker et al. 2010, 2013) and performance (Kennedy et al. 2016) of these methods vary by panel and statistic.

The advantages of web panels include low cost, speed, and convenience of a readily available pool of potential respondents. NWPs more readily harness these advantages than PWPs, which are more costly to establish and maintain (including panel attrition and refreshment), and require more time in the sampling and recruitment phase (see Liu [2016] for costs of NWPs of different lengths). The key question is whether the results from an NWP are comparable to those from a PWP, or possibly more accurate. Increasingly, studies are comparing the accuracy of the data from NWPs with a probability-based survey to inform the debate about the legitimacy of NWP data collection.

Assessment of NWPs includes comparisons with both PWPs and other probability-based data collections that do not use a web panel (e.g. random-digit dialling [RDD] telephone studies). Erens et al. (2014) compared a computer-administered self-interview and computer-administered personal interview mixed-mode probability-based survey with four quota-based NWPs. They found that all of the NWPs were demographically less representative of the general population than the probability-based data, and no single NWP consistently performed better than the others. The authors poststratified both the nonprobability and probability-based data to the population before comparing, and United Kingdom census data and other survey-generated United Kingdom official statistics served as ‘gold standards’.

While not comparing web panel with web panel, and therefore confounding the results by mode, the methodology and findings of the above study accord with the literature that compares NWPs with PWPs. Yeager et al. (2011) compared six NWPs, a PWP and an RDD study with various government benchmarks. After poststratification, all panels were found to match the population’s basic demographic distribution, but many NWPs were less accurate on additional characteristics such as household size, income, home ownership and having a passport. All panels were significantly biased on smoking and alcohol consumption. Like Erens et al. (2014), Yeager et al. (2011) concluded that all NWPs in their sample were inferior to the PWP but that none was consistently better than the others.

Chang and Krosnick (2009) also compared an NWP with a PWP and an RDD study. Their analysis assessed bias in the respondent composition using the Current Population Survey[1] as a gold standard, but also examined measurement error in the responses. While the NWP had the lowest measurement error (attributed to self-administration and topic interest), the less balanced demographic distribution made the NWP statistics more biased in total than the PWP data.

Pew Research Center researchers (Kennedy et al. 2016) compared nine NWPs (conducted by eight agencies) with Pew's own RDD-recruited PWP and 20 benchmarks, mostly from the Current Population Survey, the National Health Interview Survey[2] and the American Community Survey.[3] They found that the quality was variable across the nine NWPs, but generally those who participated in the NWPs differed from the general population on important demographic characteristics such as education, income and household size (similar to the conclusions of others). Although the distribution of race, ethnicity and gender in the NWPs often matched the population, the distributions of those who participated from these subgroups were significantly different from those for the general subpopulation, even after adjustment.

Unlike older studies, the Pew study did not conclude that the PWP was superior to the NWPs, because one NWP consistently outperformed the PWP. The authors noted that differences between NWPs seemed to be correlated with the rigour of the methodology. This emphasis on methodology is echoed throughout the comparison literature (e.g. Lee & Valliant 2009, Baker et al. 2013, Rivers 2013), and encourages both balanced sample selection and weighting after data collection, particularly to accommodate joint rather than marginal distributions.

The literature comparing NWPs with PWPs is fledgling. As survey research environments change and NWPs become more sophisticated, there is room and need to enrich and update the existing evidence. High-quality comparisons such as those mentioned use government data as a benchmark for key variables of interest, use weighted data, and sometimes explore multivariate relationships. Interestingly, however, all of the studies compared each NWP separately with the probability survey.

Considering these characteristics, we add to this research area by maintaining the established methodological procedures but also adding new developments. Our analysis has the following features:

- compares probability with nonprobability survey data in the same mode (web panel)
- uses a gold standard to assess the accuracy of both the PWP and NWP statistics
- emphasises the biases found in the relationships between variables (i.e. regression coefficients) rather than focusing on univariate comparisons
- combines the data from all the NWPs for comparison with the PWPs.

Combining the nonprobability data before comparing the statistics with those of the PWPs is intended to assess whether the aggregation decreases or increases bias. We hypothesise that the coverage and nonresponse errors in nonprobability panels are different and to some degree will cancel out when combined. To our knowledge, this assumption has been proposed (Rivers 2013) but not tested. We take this exploration a step further by looking beyond the comparison of probability with nonprobability and instead evaluate whether the combination of the PWP and NWP data results in a value even closer to the gold standard. We hypothesise that this combination will provide the most accurate estimate because measurement and nonresponse errors in the PWP are somewhat balanced by the NWPs.

2.2 Social determinants of health

We use as our substantive topic of interest outcomes on the social determinants of health. We selected this topic for three main reasons:

- It is of considerable interest for policy makers.
- There is a large body of literature in Australia and abroad against which we can compare our findings.
- High-quality population-level benchmark data are freely available.

In Wilkinson and Marmot (2003), two internationally leading researchers on the social determinants of health state that 'even in the most affluent countries, people who are less well-off have substantially shorter life expectancies and more illnesses than the rich'. In terms of causal mechanisms, the authors state that 'disadvantage has many forms and may be absolute or relative. It can include having few family assets, having a poorer education during adolescence, having insecure employment, becoming stuck in a hazardous or dead-end job, living in poor housing, trying to bring up a family in difficult circumstances and living on an inadequate retirement pension'.

There is also a large literature on the social determinants of health in Australia, highlighting similar causal factors (AIHW 2016, Fisher et al. 2016), as well as factors specific to the Australian context. Two of these Australian-specific factors are:

- a high level of international migration, with significant health differences between the Australian-born and overseas-born populations, as well as within the overseas-born population (Kennedy et al. 2015)
- an Indigenous population with significantly worse health outcomes than the non-Indigenous population and other comparable Indigenous populations (Cooke et al. 2007, Marmot 2011).

3 Data and methods

3.1 Data

Data for this paper come from two data sources. The comparison data come from the OPBS collected throughout late 2015 (Pennay et al. 2018). The benchmark data come from the Household, Income and Labour Dynamics in Australia (HILDA) survey. The compiled OPBS dataset is available for free download through the Australian Data Archive, and HILDA data are available to approved researchers through the same source (Pennay et al. 2018).

The OPBS samples were generated via three surveys based on probability samples of the Australian population and five surveys administered to members of nonprobability online panels. The three probability samples use:

- RDD, dual-frame computer-assisted telephone interview (CATI) (*n* = 601)
- address-based sampling (*n* = 538)
- telephone recruit to online survey (*n* = 560), a replication of the design for a PWP.

The first sampling frame was generated from randomly generated phone numbers, with 50% from known landline banks and 50% from known mobile number banks. The response rate, AAPOR RR3 (AAPOR 2015), among this sample was 17.9%, for an effective *n* of 601.

The address-based recruitment was conducted using the Geo-coded National Address File (GNAF). The GNAF frame collates address data from state and territory land records, the national postal authority (Australia Post) and the Australian Electoral Commission. The GNAF frame includes 14 million addresses within Australia, representing substantial coverage of the residential population. While precise coverage information is not known, more than 90% of addresses in the GNAF sample match to a registered Australia Post address, suggesting that only a small percentage of GNAF address records contain errors (Pennay et al. 2018). The response rate (AAPOR RR3) for this sample was 26.5%, for an effective *n* of 538.

The telephone-to-online sample was recruited on the back of a separate RDD-sampled CATI survey (from the ANUPoll series of public opinion surveys conducted by the Social Research Centre, an Australian National University company). This particular ANUPoll sampled a 60:40 split between landline and mobile phone numbers, and measured respondents' participation in a range of social class–related activities. At the conclusion of the interviews, respondents were asked whether they would participate in a future study on health and wellbeing issues. Of the ANUPoll respondents, 58% (*n* = 693) agreed to future contact. Of that 693, 560 participated in the subsequent OPBS survey. The net response rate (AAPOR RR3) for the final sample was 12.4%.

The numbers of completed interviews for the five nonprobability panels were 601, 600, 626, 630 and 601, respectively. Each of the selected panel providers administered the survey to their respondent pools, applying their usual recruitment and administration methods. Panel providers were instructed to draw a nationally representative sample from their respondent pools using 'non interlocking quotas set by state, region, age and gender' (Pennay et al. 2018). Additional information on the nonprobability panel recruitment, refreshment and administration methods is available in Pennay et al. (2018), although only sparse methodological information was provided to the client, as is industry standard.

The questionnaire administered to these samples – the Health, Wellbeing and Technology Survey – was designed by researchers at the Social Research Centre, and included a wide range of demographic measures and questions about health, wellbeing and the use of technology. Data collection for all eight iterations of the Health, Wellbeing and Technology Survey was undertaken between October and December 2015, with varying fieldwork periods designed to accommodate the particular requirements of each survey. All the questions used to measure primary and secondary demographic characteristics and the substantive items were adapted from high-quality Australian Government surveys.

We compare results from the OPBS with results from HILDA (Wooden & Watson 2007). HILDA is made up of a representative panel of Australian households that have been followed longitudinally from 2001. The panel was boosted in 2011, and we use data from wave 13. Given the focus of this paper, we do not exploit the longitudinal aspect of HILDA, but instead use one wave of data only. Several important positive aspects of HILDA are relevant to this paper:

- The demographic, geographic and socioeconomic controls on HILDA are very similar to those in the OPBS.
- Unlike surveys conducted by the Australian Bureau of Statistics (such as the National Health Surveys or General Social Surveys), it is possible to obtain unit record data from HILDA to include in a pooled model.
- HILDA is commonly used in Australia to understand health outcomes (see, for example, Butterworth et al. [2011]), with a robust and validated data collection methodology.

HILDA has two important limitations when used for such analysis. First, as a longitudinal panel survey, the representativeness of HILDA will have diminished by wave 13 as a result of nonrandom attrition. We control for this using population weights, but attrition due to unobservable characteristics is always difficult to control for. The second limitation is that HILDA is predominantly conducted using face-to-face interviewing, whereas the surveys in the OPBS are either completed online or via telephone interviewing. This introduces the potential for mode effects to be the cause of differences between our benchmark and comparison data, as opposed to errors in representation.

According to Crossley and Kennedy (2002), 'there is a literature which suggests that people respond more candidly to sensitive questions when self-completing a form as opposed to being personally interviewed'. Crossley and Kennedy did not find a statistically significant difference in the mean responses across five categories of self-assessed health, but did find a significant difference in the distribution of responses across categories – a fattening of the tails for interviewer compared with self-completed responses. As far as we know, no literature looks explicitly at the effect of mode of response on the apparent social determinants of health, as opposed to the levels of health outcome. This is an area of potential future research.

Given the existing literature, we do not expect mode effects to be driving our results. However, we do recommend that the predictions from HILDA be taken as indicative, and the relative difference between HILDA and the results from the OPBS be the focus, rather than taking the coefficients at face value.

3.2 Methods

The analysis of these datasets focuses on the following three research questions:

1. Are there differences in health of respondents between the OPBS surveys once social determinants are controlled for?
2. Is the social gradient the same within the OPBS surveys and relative to the benchmark?
3. Is the relationship between health and life satisfaction the same within the OPBS surveys and relative to the benchmark?

To answer the first question, we use a dataset that combines respondents from the eight surveys in the OPBS and in-scope respondents from HILDA. Our outcome of interest is self-assessed health, which falls into one of five categories: excellent (1), very good (2), good (3), fair (4) and poor (5). We model self-assessed health using an ordered probit model, which assumes a continuous, but unobserved, latent health variable that follows the standard normal distribution. We observe whether or not an individual is above or below four cutoffs and hence model the probability of the individual being in one of the five categories mentioned above.

The explanatory variables in our model are:

- sex (male is omitted category, with dummy variable for female)
- age (45–54 years is omitted category, with separate dummy variables for ages 18–24, 25–34, 35–44, 55–64, 65–74 and 75+)
- high-school education (completing Year 12 is the omitted category, with dummy variable for not having completed Year 12)
- post-school education (no qualifications is the omitted category, with separate dummy variables for having a certificate, diploma, bachelor's degree or postgraduate degree as highest level of qualifications)
- employment status (being employed is the omitted category, with a dummy variable for not employed – that is, unemployed and not in the labour force combined)
- country of birth (born in Australia is the omitted category, with a dummy variable for born overseas)
- language spoken at home (speaks English only is the omitted category, with a dummy variable for those who speak a language other than English)
- socioeconomic status of the area (measured by the 2011 Socioeconomic Indexes for Areas [SEIFA] advantage and disadvantage variable, with the omitted category being those who live in the least disadvantaged quintile, and separate dummy variables for those who live in the remaining four quintiles).[4]

We estimate three sets of models using this pooled dataset. The first set does not control for any observable characteristics, and can be interpreted as a weighted estimate of self-assessed health. The second model includes the demographic characteristics only (age and sex) and can be interpreted as a relatively simple age-standardised estimate of self-assessed health. The final set of models includes the full set of explanatory variables discussed above.

Within each of these models, the subsamples are treated in two ways. First, we use a separate dummy variable for whether or not the individual was in each of the eight subsamples from the OPBS, with the omitted category being those who were in HILDA. This is essentially a like-for-like comparison in terms of sample size.

The second specification that we estimate has a separate dummy variable for the three probability samples from the OPBS, but pools the samples from the five nonprobability panels. Given the much larger costs associated with developing or sourcing data from probability samples, this is closer to a like-for-like comparison in terms of budget. For all six combinations of probability survey and pooled NWPs, we test whether the coefficient on the dummy variables for the particular samples is statistically significant.

To answer the second research question, we run the same model as above, but we do so separately for the different samples. We run separate models for HILDA, for each of the three probability samples and for each of the five nonprobability samples. The final model is for the pooled nonprobability samples. This gives 10 models estimated in total.

For this part of the analysis, we test whether the 95% confidence intervals for the coefficients found from the probability and nonprobability panels overlap with the 95% confidence interval for the HILDA sample, as well as whether substantive conclusions would vary across the sample – for example, 'does characteristic *x* have a statistically significant association with self-assessed health?'

To answer the third and final research question, we change the dependent variable of interest. Specifically, we add life satisfaction as the dependent variable, which ranges in value from 0 to 10. We model this using the ordered probit model. Not only do we replicate the above analysis, but we also run a final set of models with a set of dummy variables for self-assessed health (the omitted category is excellent health, with separate dummy variables for poor, fair, good and very good).

For all of the model estimates, we use sample weights provided by the relevant survey company.

4 Results

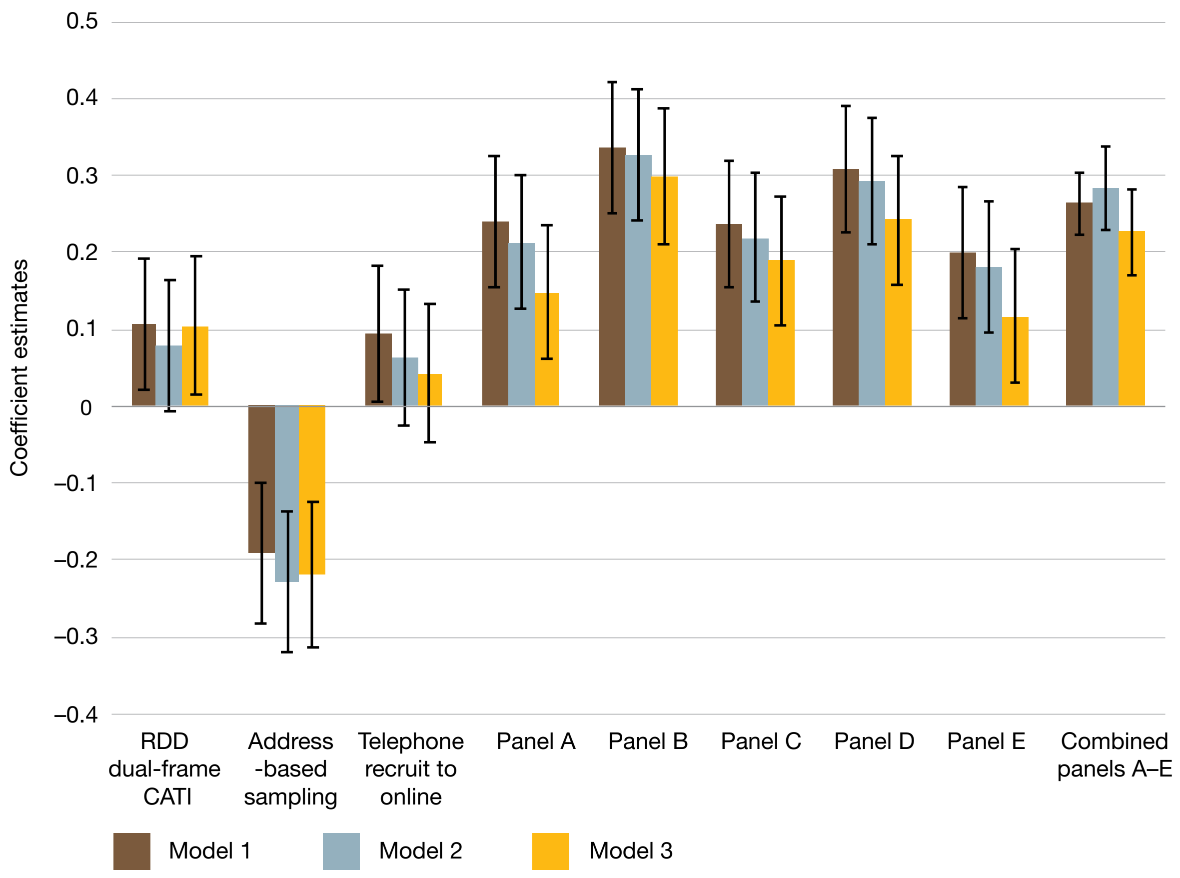
4.1 Predicting health using probability and nonprobability panels

In the first set of results, summarised in Figure 1, we show the extent to which self-assessed health varies across eight probability and nonprobability surveys relative to a gold-standard survey, while holding an increasing range of other characteristics constant (model 1 uses no controls, model 2 controls for age and gender, and model 3 controls for all variables).

The coefficient estimates give the variation of self-assessed health from the HILDA standard, as it depends on the type of panel used, estimated using the three models. The error bars noted on the graphs give the 95% confidence intervals for the coefficients.

The main finding from the analysis is that the three probability panels agree more closely with HILDA than the nonprobability panels (A–E), with no statistically significant difference in predicted self-assessed health for telephone recruit to online (the replication of the PWP). Keeping in

Figure 1 **Coefficient estimates for ordered probit model of self-assessed health**



CATI = computer-assisted telephone interview; HILDA = Household, Income and Labour Dynamics in Australia; OPBS = Online Panels Benchmarking Study; RDD = random-digit dialling

Source: Customised calculations based on the OPBS and HILDA

mind that the self-assessed health variable is coded such that higher values represent worse health, the results show that the nonprobability panels predict worse health outcomes than HILDA.

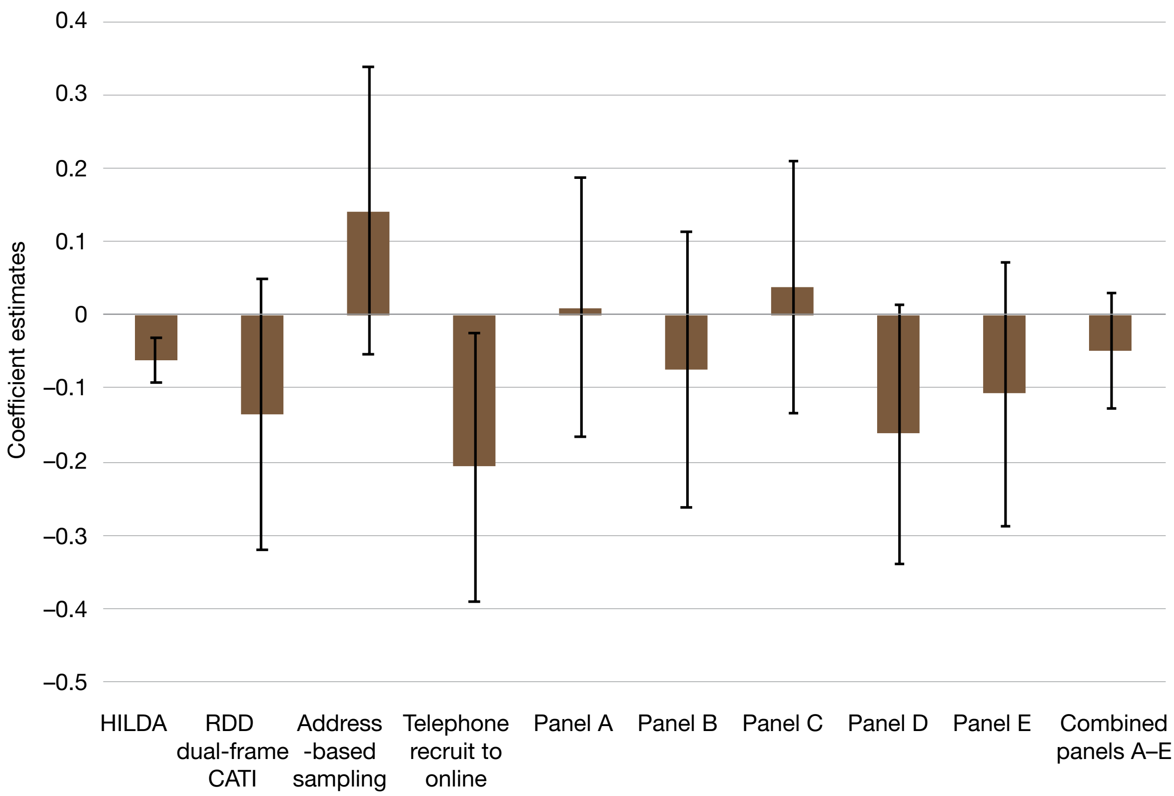
Within the probability panels, some differences in results occur, depending on which characteristics are controlled for. Without controlling for any characteristics, Panel 1 (RDD dual-frame CATI) overestimates poor health, and the address-based RDD underestimates poor health (predicts lower scores, which correspond to better health). Controlling for the full set of characteristics, there is no significant difference in self-assessed health (compared with HILDA) for two of the panels, with the address-based sample having slightly better self-assessed health than the HILDA sample.

4.2 The social health gradient

The next set of results looks at the predictors of self-assessed health, and how they vary across the different samples. Figure 2 looks at the extent to which females have different self-assessed health from males while holding other characteristics constant. The gold-standard data from HILDA suggest that females have slightly better health outcomes than males (remembering that the categorical variable is reverse coded). Only one of the samples predicts a statistically significant (and negative) sex–health gradient: the telephone recruit to online. By comparison, in no nonprobability samples (panels A–E) nor the combined panel was sex statistically significant. The wider confidence intervals compared with Figure 1 are because of smaller sample sizes.

The next set of summary results looks at the relationship between employment and self-assessed health (Figure 3). From a substantive perspective, it is not clear whether poor health predicts lower employment or whether lower employment predicts worse health (Bartley 1994). However, both effects are likely to be in the same direction, so the coefficients can be interpreted as the combined effect. This interpretation issue

Figure 2 Relationship between sex of respondent and self-assessed health



CATI = computer-assisted telephone interview; HILDA = Household, Income and Labour Dynamics in Australia; OPBS = Online Panels Benchmarking Study; RDD = random-digit dialling

Source: Customised calculations based on the OPBS and HILDA

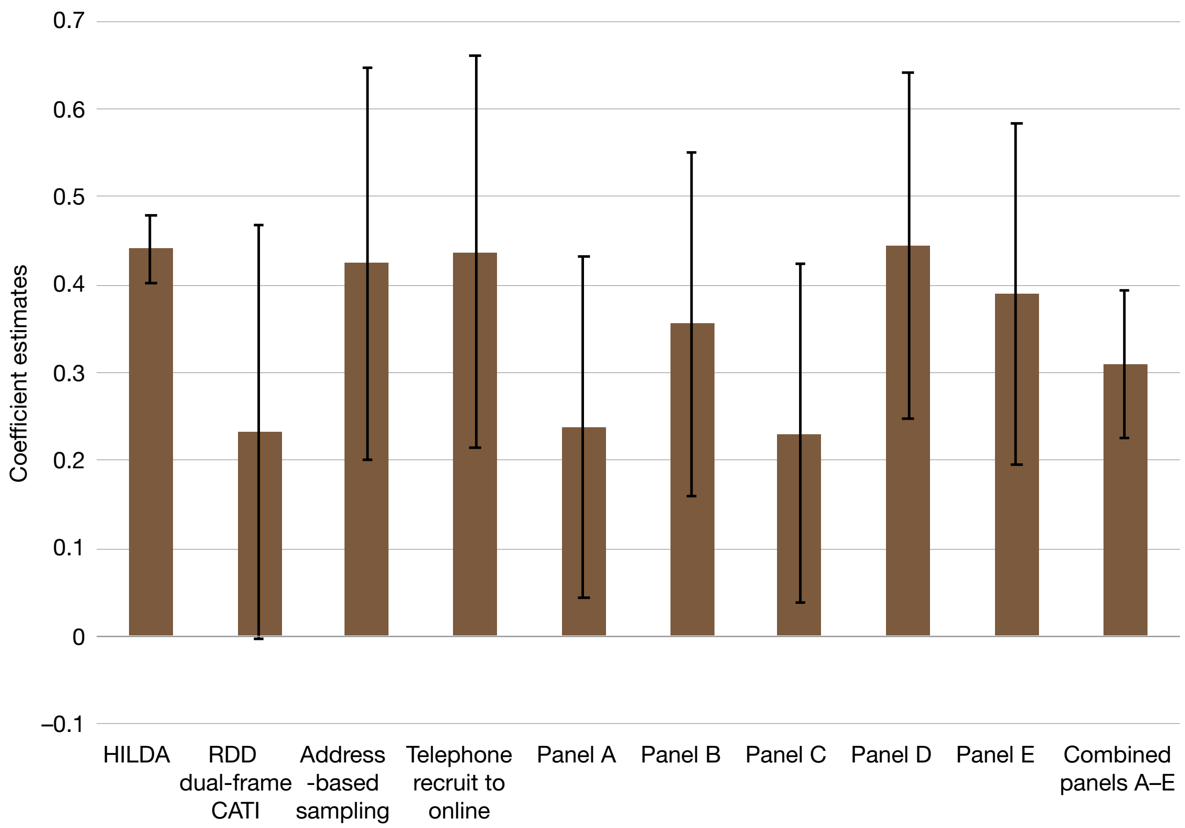
aside, seven of the eight samples (as well as the combined A–E sample) predict worse health outcomes for those not employed. The only exception is the RDD dual-frame CATI, which has a *P* value of 0.053.

The magnitude of the coefficient varies. All but one of the five nonprobability panels have a coefficient estimate that is outside the confidence interval for HILDA (although the confidence intervals from the panels overlap the HILDA confidence interval because of very imprecise estimates), and the combined A–E panel has a coefficient estimate and confidence interval that are outside the confidence interval for the HILDA estimate.

The final set of individual results from the models of self-assessed health relates to the coefficient estimates for high-school completion. For this variable, results from HILDA suggest that those who have not completed Year 12 have significantly worse health outcomes than those who have (Figure 4). For the RDD dual-frame CATI and the address-based sample, the difference is statistically significant, and the confidence intervals overlap the HILDA confidence intervals. For the telephone recruit to online sample, the coefficient is not statistically significant, although it is similar in magnitude to HILDA. The combined NWPs have a coefficient that is similar in magnitude to HILDA (with a *P* value of 0.052), whereas none of the individual NWPs is close to significant. In other words, if a researcher were to use the replication of the PWP or any of the NWP surveys to understand the social determinants of health, they would conclude that high-school education is not a significant factor, a conclusion that is quite different from that derived from the gold-standard survey.

The final set of results in this subsection looks at the relationship between the socioeconomic

Figure 3 Relationship between employment status of respondent and self-assessed health



CATI = computer-assisted telephone interview; HILDA = Household, Income and Labour Dynamics in Australia; OPBS = Online Panels Benchmarking Study; RDD = random-digit dialling

Source: Customised calculations based on the OPBS and HILDA

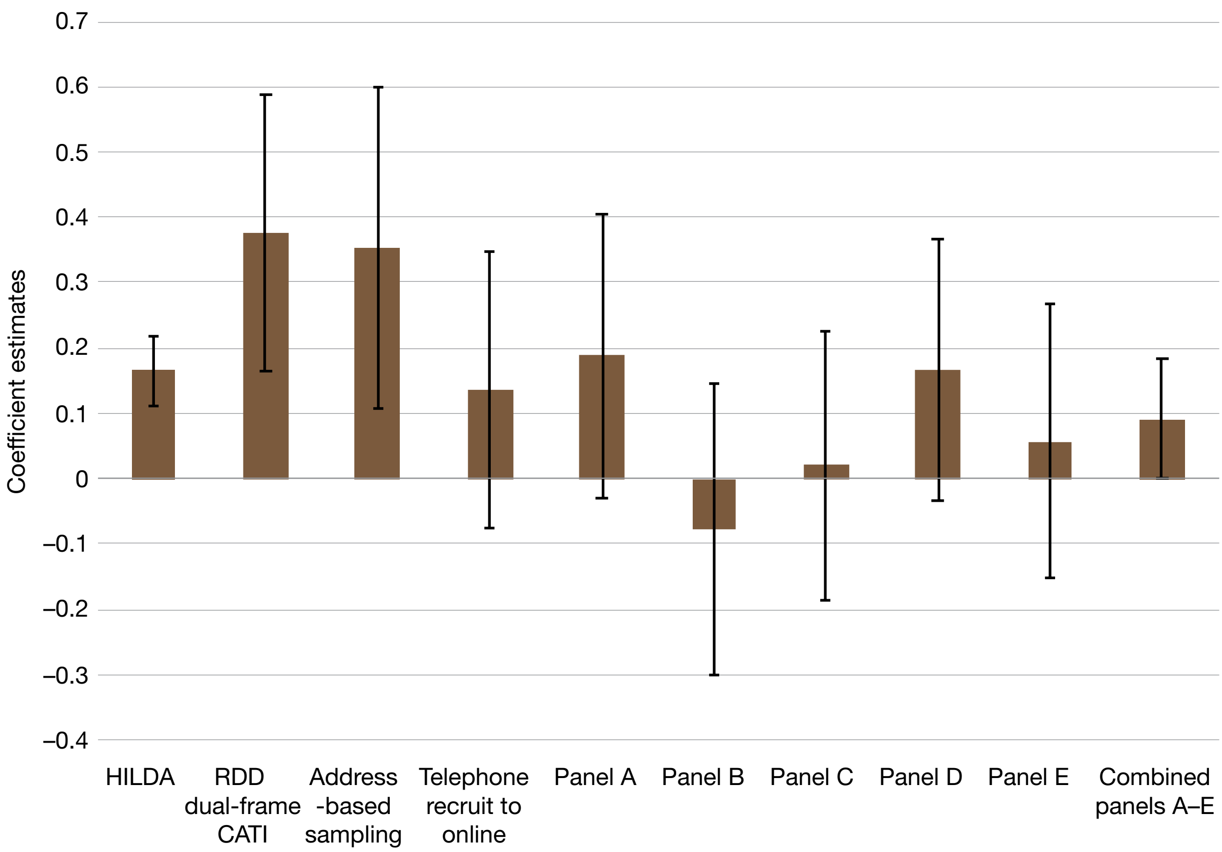
status of the area in which a person lives and their health status. This is a key area of policy interest, as many health policies intervene at a geographic rather than an individual level, and therefore the extent to which health outcomes vary by geography influences the extent to which such interventions can be tightly targeted to those most in need. This is the variable most poorly captured by the online surveys.

When looking at self-assessed health as it depends on the socioeconomic characteristics of the area in which a person lives, the results from the HILDA gold standard are reasonably clear. Those who live in relatively disadvantaged areas have significantly and substantially worse health outcomes than those who live in relatively advantaged ones, with the difference reasonably consistent across the distribution (Figure 5).

None of the probability panels, however, matches that distribution. The RDD dual-frame CATI shows a significant difference between quintile 5 (the omitted category and most advantaged set of areas) and quintiles 1, 2 and 4, and a lower probability for quintile 3. For the remaining probability surveys and for the five NWPs in isolation, there were no significant differences between the omitted category and the other quintiles, apart from quintile 4 for the address-based sampling, which is only just statistically significant. The distributions for some of the nonprobability panels have a similar shape to the HILDA gold standard. However, because of the large standard errors, none of the differences are statistically significant.

The only sample that has a similar distribution to the HILDA distribution is the combined nonprobability panels. For this sample, those in

Figure 4 Relationship between high-school completion and self-assessed health



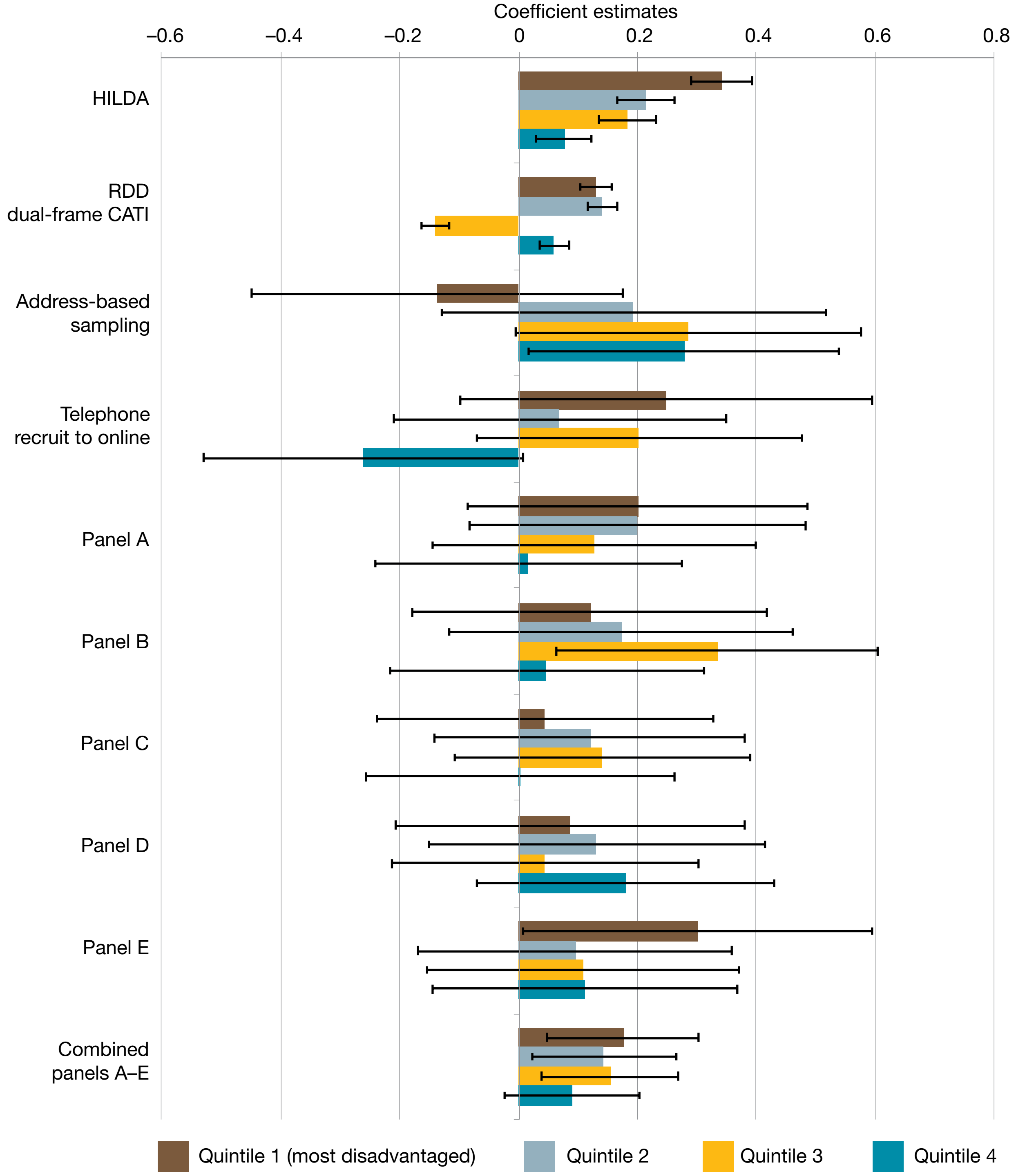
CATI = computer-assisted telephone interview; HILDA = Household, Income and Labour Dynamics in Australia; OPBS = Online Panels Benchmarking Study; RDD = random-digit dialling

Source: Customised calculations based on the OPBS and HILDA

the first three quintiles have health outcomes that are significantly different from the base case at the 5% level of significance, whereas the *P* value for quintile 4 is 0.115.

One reason for the null findings for the SEIFA variables is that most of the samples in the OPBS have far fewer individuals in the lower quintiles than they would if they were representative

Figure 5 Relationship between self-assessed health and area-level disadvantage



CATI = computer-assisted telephone interview; HILDA = Household, Income and Labour Dynamics in Australia; OPBS = Online Panels Benchmarking Study; RDD = random-digit dialling

Source: Customised calculations based on the OPBS and HILDA

(geographically) of the population as a whole. Within HILDA, each quintile contains 19.1–20.6% of the sample – roughly the same as Australia as a whole (by definition). By comparison, the three probability surveys have 11.8–15.4% of their samples in the bottom quintile. Despite being probability surveys, these samples do not pick up the geographic distribution of health outcomes.

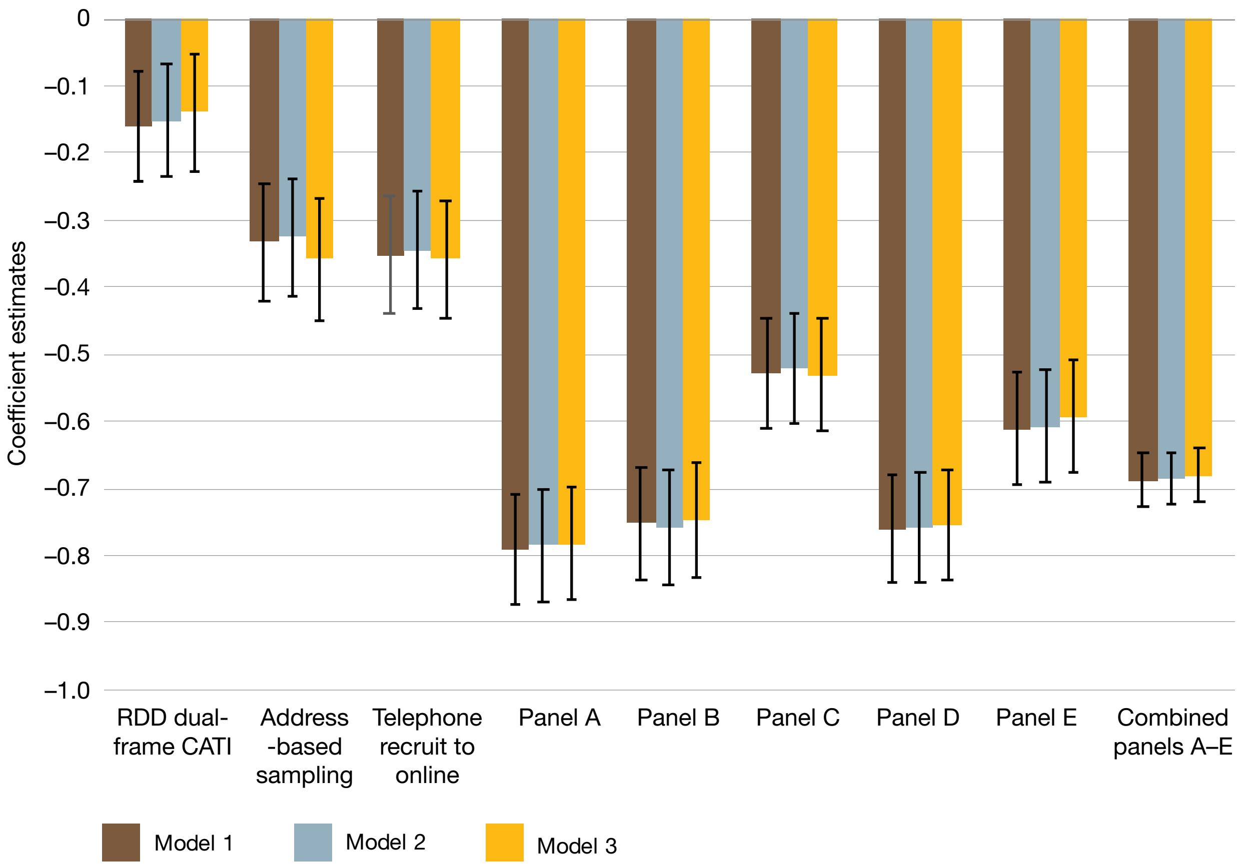
4.3 Determinants of life satisfaction

In this final subsection of results, we look at the factors associated with life satisfaction. This variable is important in and of itself, given the increasing interest in wellbeing as an area of policy concern (Easterlin 2010). However, it is of additional relevance because of the consistent finding that poor health is a key predictor of low wellbeing (Krueger & Stone 2014).

We begin the analysis by looking at the extent to which life satisfaction is different in the various subsamples compared with the HILDA gold standard. Unlike the results for self-assessed health, all eight of the subsamples and the combined nonprobability panels have significantly and substantially different values from HILDA. It is true that the differences are greatest for the nonprobability panels compared with the probability samples, but even the latter have lower values (Figure 6).

The OPBS samples show significant differences in life satisfaction from HILDA, but the relationship between health and life satisfaction is very similar. Not surprisingly, those with worse health report lower life satisfaction in the HILDA sample. The direction and shape of this relationship are very similar in each of the OPBS samples, both individually and combined (Figure 7).

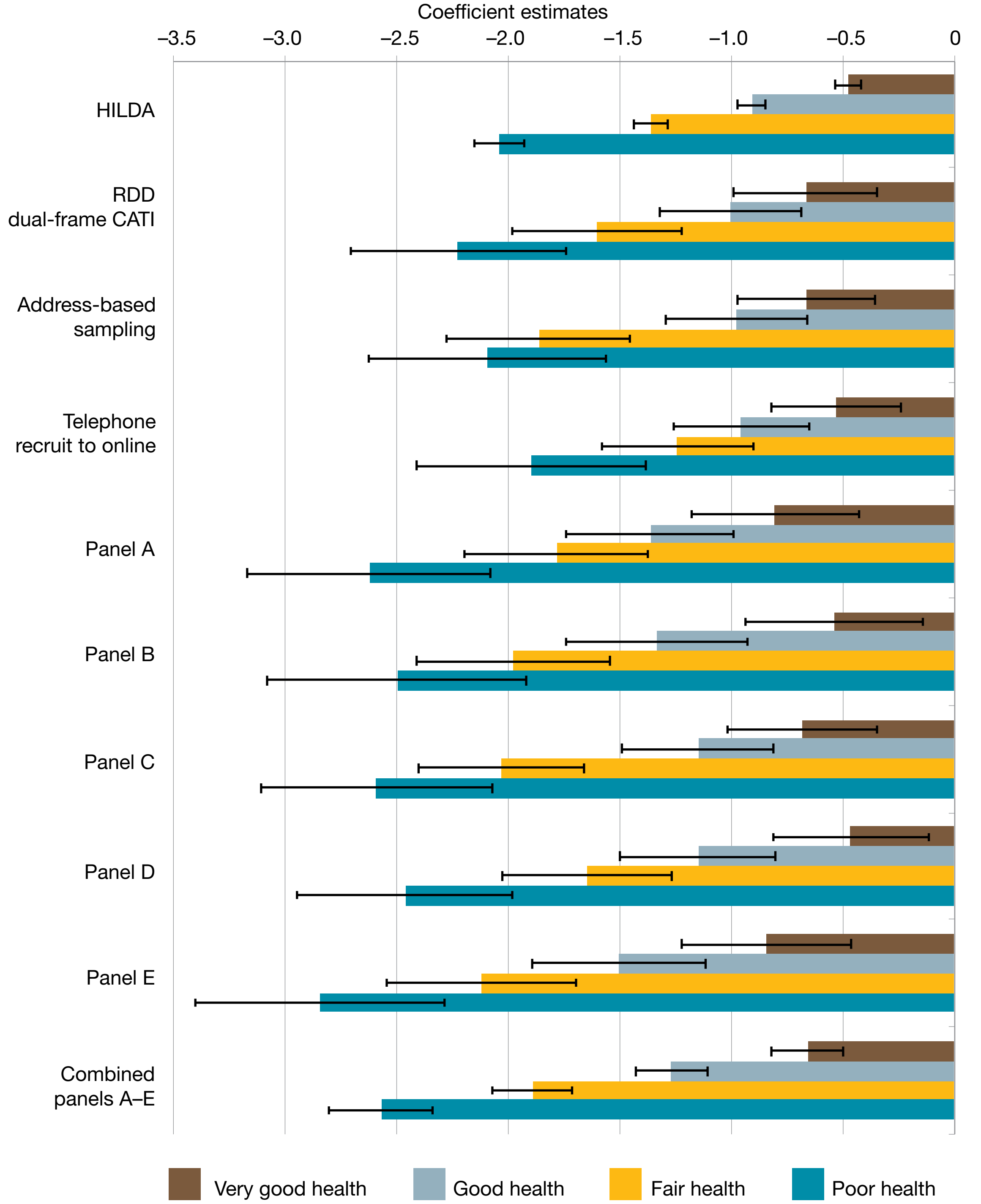
Figure 6 Coefficient estimates for ordered probit model of life satisfaction



CATI = computer-assisted telephone interview; HILDA = Household, Income and Labour Dynamics in Australia; OPBS = Online Panels Benchmarking Study; RDD = random-digit dialling

Source: Customised calculations based on the OPBS and HILDA

Figure 7 **Relationship between self-assessed health and life satisfaction**



CATI = computer-assisted telephone interview; HILDA = Household, Income and Labour Dynamics in Australia; OPBS = Online Panels Benchmarking Study; RDD = random-digit dialling

Source: Customised calculations based on the OPBS and HILDA

5 Summary and concluding comments

Use of online panels for research and policy purposes is increasing, in part as a response to the increasing cost of, and declining response rates to, more traditional data collection methods. An additional factor is the ease of access and relatively quick turnaround in data collection. The view (or fear) is that such samples are less representative of the population than samples obtained using more traditional methods.

To test for this, the OPBS collected a range of information from eight separate samples. Three were created probabilistically, whereas the remaining five were from commercial, nonprobabilistic panels.

Much of the benchmarking research, including that undertaken on the OPBS, has focused on the prevalence of particular outcomes, and whether weighting based on observable demographic characteristics gets one closer to the gold-standard benchmarks such as HILDA.

Prevalence of given outcomes (and how they might be changing through time) is clearly of ongoing policy interest. However, policy makers are often as interested in the relationships between variables. If factor *x* predicts outcome *y*, and factor *x* is amenable to policy intervention or targeting, then this suggests potential policy levers. One very important set of relationships is the social and economic determinants of health and subjective wellbeing.

As far as the authors are aware, this is the first paper to focus on the extent to which probability and nonprobability samples capture the social determinants of health in the same way as a high-quality, interview-based sample.

The analysis finds that, when looking at the distributions of self-assessed health and life satisfaction, probability panels differ less from the gold standard than do the nonprobability panels. This supports previous work, although we also show that this conclusion holds when a greater range of control variables is included in the model.

We also show that some of the predictors of health are captured better using the nonprobability panels. However, this is not universally the case. In particular, the relationship between area-level disadvantage and health is better captured through a pooled nonprobability sample. It is true that the sample size is much larger for this pooled panel than for the individual probability samples. However, the cost of the pooled sample is probably not too dissimilar from that of the individual probability samples.

Ultimately, the results show that high-quality, official surveys (such as HILDA) still have a place. Probability panels would appear to better approximate these with regard to predictive relationships. However, serious questions remain about the geographic distribution of these samples.

Notes

1. https://www.census.gov/programs-surveys/cps.html
2. https://www.cdc.gov/nchs/nhis/
3. https://www.census.gov/programs-surveys/acs/
4. 'SEIFA is a suite of four indexes that have been created from social and economic census information. Each index ranks geographic areas across Australia in terms of their relative socio-economic advantage and disadvantage. The four indexes each summarise a slightly different aspect of the socio-economic conditions in an area.' (ABS 2016)

References

AAPOR (American Association for Public Opinion Research) (2015). *Standard definitions: final dispositions of case codes and outcome rates for surveys*, rev. April 2015, AAPOR, Oakbrook Terrace, Illinois.

ABS (Australian Bureau of Statistics) (2016). *Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016*, ABS cat. no. 2033.0.55.001, ABS, Canberra.

AIHW (Australian Institute of Health and Welfare) (2016). *Australia's health 2016*, Australia's Health Series no. 15, cat. no. AUS 199, AIHW, Canberra.

Baker R, Blumberg SJ, Brick JM, Couper MP, Courtright M, Dennis JM, Dillman D, Frankel MR, Garland P, Groves RM, Kennedy C, Krosnick J, Lavrakas PJ, Lee S, Link M, Piekarski L, Rao K, Thomas RK & Zahs D (2010). AAPOR report on online panels. *Public Opinion Quarterly* 74(4):711–781.

——— , Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, Gile KJ & Tourangeau R (2013). Summary report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology* 1(2):90–143.

Bartley M (1994). Unemployment and ill health: understanding the relationship. *Journal of Epidemiology & Community Health* 48(4):333–337.

Butterworth P, Leach LS, Rodgers B, Broom DH, Olesen SC & Strazdins L (2011). Psychosocial job adversity and health in Australia: analysis of data from the HILDA survey. *Australian and New Zealand Journal of Public Health* 35(6):564–571.

Chang L & Krosnick JA (2009). National surveys via RDD telephone interviewing versus the internet: comparing sample representativeness and response quality. *Public Opinion Quarterly* 73(4):641–678.

Cooke MF, Mitrou F, Lawrence D, Guimond E & Beavon D (2007). Indigenous well-being in four countries: an application of the UNDP's human development index to Indigenous peoples in Australia, Canada, New Zealand, and the United States. *BMC International Health and Human Rights* 7(1):9.

Crossley TF & Kennedy S (2002). The reliability of self-assessed health status. *Journal of Health Economics* 21(4):643–658.

Dever JA, Rafferty A & Valliant R (2008). Internet surveys: can statistical adjustments eliminate coverage bias? *Survey Research Methods* 2(2):47–60.

Easterlin RA (2010). Well-being, front and center: a note on the Sarkozy report. *Population and Development Review* 36(1):119–124.

Erens, B, Burkill S, Couper MP, Conrad F, Clifton S, Tanton C, Phelps A, Datta J, Mercer CH, Sonnenberg P, Prah P, Mitchell KR, Wellings K, Johnson AM & Copas AJ (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: a comparison with a probability sample interview survey. *Journal of Medical Internet Research* 16(12):e276.

Fisher M, Baum FE, MacDougall C, Newman L & McDermott D (2016). To what extent do Australian health policy documents address social determinants of health and health equity? *Journal of Social Policy* 45(3):545–564.

Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E & Tourangeau R (2011). *Survey methodology*, 2nd edn, John Wiley & Sons, Hoboken, New Jersey.

Heckathorn DD (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems* 44:174–199.

Kennedy C, Mercer A, Keeter S, Hatley N, McGeeney K & Gimenez A (2016). *Evaluating online nonprobability surveys*, Pew Research Center, Washington, DC.

Kennedy SM, Kidd MP, McDonald JT & Biddle N (2015). The healthy immigrant effect: patterns and evidence from four countries. *Journal of International Migration and Integration* 16(2):317–332.

Krueger AB & Stone AA (2014). Progress in measuring subjective well-being. *Science* 346(6205):42–43.

Lee S & Valliant R (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research* 37(3):319–343.

Liu M (2016). Comparing data quality between online panel and intercept samples. *Methodological Innovations* 9:1–11.

Marmot M (2005). Social determinants of health inequalities. *Lancet* 365(9464):1099–1104.

——— (2011). Social determinants and the health of Indigenous Australians. *Medical Journal of Australia* 194(10):512–513.

Pennay D, Borg K, Neiger D, Misson S, Honey N & Lavrakas P (2018). *Online Panels Benchmarking Study, 2015*, Australian Data Archive Dataverse, doi:10.4225/87/FSOYQI.

Rivers D (2013). Comment. *Journal of Survey Statistics and Methodology* 1(2):111–117.

Valliant R & Dever JA (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research* 40(1):105–137.

Wilkinson RG & Marmot M (2003). *Social determinants of health: the solid facts*, 2nd edn, Centre for Urban Health, World Health Organization, Copenhagen.

Wooden M & Watson N (2007). The HILDA survey and its contribution to economic and social research (so far). *Economic Record* 83(261):208–231.

Yeager DS, Krosnick JA, Chang L, Javitz HS, Levendusky MS, Simpser A & Wang R (2011). Comparing the accuracy of RDD telephone surveys with internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly* 75(4):709–747.

CENTRE FOR SOCIAL
RESEARCH & METHODS

+61 2 6125 1279
csrm.comms@anu.edu.au

The Australian National University
Canberra ACT 2601 Australia

www.anu.edu.au