Survey Research Methods (2022)
Vol. 16, No. 2, pp. 241-266
doi:10.18148/srm/2022.v16i2.7907
European Survey Research Association

Comparing Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia: A Total Survey Error Perspective

Paul John Lavrakas
The Social Research Centre
Melbourne, Australia

Darren Pennay
The Social Research Centre
Australian National University

Dina Neiger
The Social Research Centre
Australian National University

Benjamin Phillips
The Social Research Centre
Australian National University

In this paper we report the findings from our study which was undertaken to learn if the findings of Chang and Krosnick (2009), Cornesse et al. (2020), Erens et al. (2014), MacInnis, Krosnick, Ho, and Cho (2018), Yeager et al. (2011) would be replicated in Australia. Our Australian Online Panels Benchmarking Study (OPBS) involved administering the same questionnaire across eight independent national Australian samples aiming to achieve approximately 600 completed questionnaires/interviews from each sample. The questionnaire was administered by the Social Research Centre (SRC), a subsidiary of the Australian National University (ANU), to three probability samples, and to five nonprobability samples drawn from the online panels operated by five independent nonprobability online panel providers. A dual-frame telephone sampling methodology was used for two of the probability surveys and the third used an address-based sampling (ABS) frame. The target population for the OPBS was persons aged 18 years and older living in Australia who were fluent in English. The three probability sample surveys likely had little Coverage Error, a known amount of Sampling Error, a nonignorable amount of Nonresponse Error, little Adjustment Error, and a small-to-modest amount of Measurement Error. Overall, the three probability samples as a group were less biased on the substantive measures and had less variance from the benchmark values, compared to the nonprobability surveys. The five nonprobability surveys likely had a nonignorable amount of Coverage Error, an unknowable amount of Sampling Error, a nonignorable amount of Nonresponse Error, unknown Adjustment Error, and a small-to-modest amount of Measurement Error. Overall, the five nonprobability panel surveys as a group were more biased on the substantive measures and had more variance from the benchmark values, compared to the probability surveys. Our OPBS study replicates very closely with previous comparison studies conducted in other countries.

Keywords: Nonprobabilty samples; Probability samples; replication study; total survey error

1 Introduction

Around the world nowadays, online panels are routinely used as a method for gathering survey data. A precursor of these panels was established in the Netherlands in the 1980s (Saris, 1998), using a probability-based sample of approximately 1,000 citizens, and telephone as the data collection mode. But with the explosion of the Internet in the 1990s, online panels were established predominantly using nonprobability-based sampling and computer-assisted web data collection (CAWI) in the United States and Europe (Callegaro et al., 2014). And, because the business model for these nonprobability panels was shown to be profitable, panel companies started to create their own international footprint by acquiring established nonprobability panels throughout the world. Nonprobability research panels were first created in Australia in the late 1990s (Fine, 2016).

Since the start of online panels, the vast majority of them, as well as most people who participate in them, have been established/recruited via nonprobability sampling methods. This is the case worldwide (see ESOMAR, 2020; Fine, 2016). In parts of Europe and in the United States, the increased use of CAWI also resulted in the establishment of several probability-based online panels during the late 1990s and early 2000s onwards, which enabled a high level of cov-

Contact information: Paul John Lavrakas (care of Benjamin Phillips), the Social Research Centre, PO Box 13328, Law Courts VIC 8010 (E-Mail: pjlavrakas@gmail.com)

erage and the scientific and representative sampling of the population. However, the first Australian probability-based online panel, *Life in Australia*™, was not established until 2016.

Using data from the most recent ESOMAR Global Market Research Industry Report (ESOMAR, 2020, pp. 10, 163) we estimated that globally in 2019, approximately US$15.3 billion was spent on online research. The majority of the clients who use online panels for their research needs choose to use nonprobability panels instead of probability-based panels. There are many reasons for this, including nonprobability panels having much lower cost, being much quicker in providing data, being more able to generate sufficient numbers of respondents from much smaller segments of the general population, and an apparent sense that the quality of the data from these nonprobability panels is fit for the purpose(s) to which clients will put the data.

2 Previous research

Our study is positioned within the context of a growing body of international research examining the differences in the estimates produced from surveys conducted via probability sampling methods with those from surveys conducted via nonprobability online panels. Cornesse et al. (2020) describe 23 such studies from around the world. By country of origin, there have been 10 such studies reported from the USA, three from the Netherlands, two each from Germany, Belgium and the UK and one from Canada, Sweden and France, and this study from Australia. To our knowledge this study was the first of its kind undertaken outside of Europe and North America.

In reviewing this previous research, Cornesse et al. (2020, pp. 14–15) report that

- "...the higher accuracy of probability sample surveys has been demonstrated across various topics, such as voting behavior (Chang & Krosnick, 2009; Malhotra & Krosnick, 2007; Sturgis et al., 2018), health behavior (Yeager et al., 2011), consumption behavior (Szolnoki & Hoffmann, 2013), sexual behavior and attitudes (Erens et al., 2014; Legleye et al., 2018), and socio-demographics (Chang & Krosnick, 2009; Dutwin & Buskirk, 2017; Erens et al., 2014; MacInnis et al., 2018; Malhotra & Krosnick, 2007; Szolnoki & Hoffmann, 2013; Yeager et al., 2011).
- In addition, the higher accuracy of probability sample surveys has been found across a number of countries, such as Australia (Pennay, Neiger, Lavrakas, & Borg, 2018), France (Legleye et al., 2018), Germany (Blom et al., 2018; Szolnoki & Hoffmann, 2013), the Netherlands (Brüggen, van den Brakel, & Krosnick, 2016; Scherpenzeel & Bethlehem, 2011), Sweden (Sohlberg, Gilljam, & Martinsson, 2017), the United Kingdom (Sturgis et al., 2018), and the United States (Chang & Krosnick, 2009; Dutwin & Buskirk, 2017; MacInnis et al., 2018; Malhotra & Krosnick, 2007; Yeager et al., 2011).
- Furthermore, the higher accuracy of probability sample surveys has been shown over time, with the first study demonstrating higher accuracy of probability sample surveys in 2007 (Malhotra & Krosnick, 2007) to the most recent ones in 2018 (Blom et al., 2018; Legleye et al., 2018; MacInnis et al., 2018; Sturgis et al., 2018). All of these studies from different times and countries and that focused on different topics reached the same overarching conclusion that probability sample surveys led to more accurate estimates than nonprobability samples."

3 Our research in Australia

The substantive topics of interest in our Australian study were self-reported socio-demographics and health determinants and outcomes. The "health" domain was measured by questions relating to life satisfaction, general health, psychological distress, tobacco and alcohol consumption, and whether covered by private health insurance. We gathered essentially the same data from national Australian samples of adults in three probability-based surveys and in five nonprobability-based surveys. We have compared the data in these eight surveys, as presented in this article, using a Total Survey Error framework (see Groves, 1989; Groves & Lyberg, 2010) to guide our comparisons/evaluations.

Our research was undertaken to learn if the main findings from the prior international studies would be replicated in Australia. The conclusions from those studies were broadly as follows:

- Probability samples were consistently highly accurate across a set of demographics and nondemographics, especially after poststratification with primary demographics;
- Nonprobability samples done via the Internet were consistently less accurate, on average, than probability sample surveys;
- There was considerable variation in accuracy among the findings of nonprobability samples, much more so than among probability samples;
- Poststratification with demographics sometimes improved the accuracy of nonprobability sample surveys and sometimes reduced their accuracy, so this method should not be relied upon with confidence to try to repair deficiencies in such samples.

Yeager et al. (2011, p. 737) concluded that "probability samples, even ones without especially high response rates, yielded quite accurate results. In contrast, nonprobability samples were not as accurate and were sometimes strikingly inaccurate, regardless of their completion rates". Brüggen et al. (2016, p. 21) further noted "the inconsistent performance of online nonprobability panels" and that nonprobability panels that performed well for one set of variables did not necessarily do so for others.

4 Methodology

Our 2015 Australian Online Panels Benchmarking Study (OPBS) involved administering the same questions across eight independent national Australian samples to achieve approximately 600 completed questionnaires/interviews from each sample. The questionnaire was devised by researchers from the Social Research Centre (SRC), a subsidiary of the Australian National University (ANU), and was administered to three probability samples by the SRC and to five nonprobability samples each drawn from the online panels operated by five independent nonprobability online panel providers in Australia. A dual-frame telephone sampling methodology was used for two of the SRC's probability surveys and the third SRC survey used an address-based sampling (ABS) frame.

The target population for each of the eight surveys was persons aged 18 years and older living in Australia and able to self-complete the questionnaire in English or be interviewed in English.

4.1 Choosing the Online Nonprobability Panels

Eight Australian online panel providers were invited to submit a quotation to administer the questionnaire to members of their Australian nonprobability panels. These providers were not randomly selected from all panel operators in Australia but approached based on being known to the SRC either directly or by repute. Two of the companies did not submit a proposal by the given deadline and were excluded. The six remaining companies were assessed against a range of criteria, excluding price; the lowest rated panel was excluded.

Of the five companies selected, four complied with ESOMAR by addressing all of "28 Questions to Help Buyers of Online Panels;" the other company partially complied with the ESOMAR requirements. Three of the five panels were ISO 26362 accredited. The difference in cost between the lowest and highest priced panel survey was 24 percent.

Data collection for all eight surveys was undertaken between October and December 2015 with varying fieldwork periods designed to accommodate the requirements of each survey design/mode. Ethical clearance for the conduct of this research was provided by the ANU Human Research Ethics Committee.

4.2 Questionnaire

The data collection instrument that was administered to the eight samples was called the *Health, Wellbeing and Technology Questionnaire*. It included a wide range of questions about health, wellbeing, the use of technology, and demographics.

While the questions were presented in as consistent a manner as possible, there were some minor differences in presentation to accommodate the various data collection modes and formats. The precise wording used for the substantive questions of interest across the three modes of data collection used for the surveys (CATI, online, and mailback paper survey) and the wording of the questions in the benchmark surveys is provided in Appendix A.[1] The mailback paper survey version of the questionnaire is available in SRM's online appendix.

The questionnaire covered four broad topic areas:

1. *Primary demographics*—Sex, age, location, educational attainment, country of birth, and telephone status. These are the variables the SRC typically uses for poststratification weighting.

2. *Secondary demographics*—Indigenous status, citizenship, enrolled to vote, geographic mobility, employment status, language spoken at home, home ownership status, volunteerism, household composition, and earned compensation (wages/salaries).

3. *Substantive measures*—General health status, psychological distress, life satisfaction, private health insurance coverage, daily smoking status, and alcohol consumption in the last 12 months.

4. *Calibration variables*—Early adopter questions and use of information technology—accessing the internet, internet use, and online survey participation.

The selection of the calibration variables warrants some further explanation. Fahimi, Buttermore, Thomas, and Barlas (2015, p. 9) found that calibration approaches that involved additional attitudinal and behavioural benchmarks to the poststratification weight showed "great promise for reducing systematic biases" in surveys. This work demonstrated that early adopter items, internet and TV usage, and eligibility for a health-care card differentiate well between probability and nonprobability samples, thereby making them good candidates to include in calibration. Accordingly, these variables were used as the calibration variables in addition to secondary demographics in this study and the analysis of how these performed in reducing bias was the subject of a separate research project (see Neiger, Pennay, Ward, & Lavrakas, 2017).

All the questions used to measure primary and secondary demographic characteristics and the substantive items were

[1] Of note, there is a slight disagreement in the literature on the best means for trying to achieve data collection mode equivalence. For example, Dillman (2000) advocates unimodal design, which would use precisely the same wording and as close to the same formatting as possible across all data collection modes. On the other hand, de Leeuw (2005) writes of "generalized [data collection] mode design" where a slightly different format may be used across data collection modes to achieve cognitive equivalence of stimulus. As it applies to our study, what de Leeuw prescribes is using mode experiments and a more formal process of trying to ensure mode equivalence than we were able to deploy. Thus, we opted to follow the approach prescribed by Dillman.

adapted from measures included in the Australian 2016 Census of Population and Housing or from high quality Australian government surveys or official statistics. This is a critical part of our overall research design as it enables the accuracy of the estimates derived from the various probability and nonprobability surveys to be compared against each other and against official benchmarks. On average, the questionnaire took between 6–11 minutes to complete across the various modes (excluding the paper survey mailback mode for which the average completion time is unknown).

4.3 Probability-based Surveys: Sample Design and Recruitment

The three probability surveys all used different sampling designs.

Probability Design 1—Dual-frame Random Digit Dialling (DFRDD). This survey was a standalone DFRDD telephone survey. The samples were randomly generated landline and mobile phone telephone numbers with 50% of interviews to be completed via the landline frame and 50% via the mobile phone frame. For the landline frame, 15 probability-proportional-to-size geographic strata were established based on the capital city/noncapital city distribution of adults across the Australian states and territories, except for the Australian Capital Territory, which was treated as a single stratum. If there were two or more eligible adults in a household, when reaching residential households from the landline frame, a quasi-random allocation was undertaken to select one adult with either the "next birthday" or "most recent birthday".[2] A single national stratum was used for the mobile frame because, in Australia, mobile phone numbers do not contain any geographic markers. For the mobile phone sample, and as is the common practice throughout the world, the person answering the phone was the person invited to participate in the survey, provided s/he was eligible to participate.

Probability Design 2—Address-based Sampling (ABS). The sampling frame used for this survey was the Australian Geocoded National Address File (G-NAF). G-NAF is maintained by what is now Geoscape (formerly the Public Sector Mapping Agency, PSMA Australia), a public company owned by Australia's federal, state and territory governments, and is the authoritative national address index for Australia. The G-NAF is compiled from existing and recognized address sources from the state and territory government land records, as well as address data from Australia Post and the Australian Electoral Commission (Geoscape, N.d.). The address sample was selected from the G-NAF database using a stratified sampling design in accordance with the distribution of the Australian residential population aged 18 years and over across 15 geographic strata. To accommodate situations in which more than one person in a household was eligible, the printed instructions at the beginning of the CAWI and mailback paper questionnaires asked for the person aged 18 years or older with either "next birthday" or "most recent birthday" (randomly alternating) to complete the questionnaire (please see Footnote 3).

Probability Design 3—DFRDD ANUpoll. Participants in this survey were recruited at the conclusion of an established DFRDD survey, the 21st ANUpoll, in an ongoing series of polls being undertaken by the SRC for the ANU. Respondents who completed the October 2015 ANUpoll, which explored attitudes to aging and money, were invited to take part in "a future study about health and wellbeing." Contact details were captured for those who agreed to participate in the subsequent survey and, depending upon their preference, these sample members were either emailed a link to complete the questionnaire online or sent a paper survey of the questionnaire to return via the mail. The ANUpoll survey (i.e., the host survey used for recruitment) utilized a DFRDD sampling design with a 60:40 split in the proportion of interviews obtained via landline and mobile phone numbers. For the landline frame the same 15 geographic strata were used. For the mobile frame a single national stratum was used, for the reason noted above. When calling the landline sample the method used to select the household member to interview when there were two or more eligible adults in a household was restricted to the "next birthday" method only (as that was the standard practice for the ANUpolls conducted at that time). For the mobile phone sample, the person answering the phone was the person invited to participate in the survey, provided s/he was eligible to participate. Of the 1,200 respondents who completed the host survey and were invited to participate in a "future study about health and wellbeing," 693 (58%) agreed and provided an email address and/or a physical address for distribution of the questionnaire.

4.4 Nonprobability Surveys: Sample Design and Recruitment

The recruitment methods used to build nonprobability panels varied considerably. Common elements include banner advertising on websites, online invitations and messages embedded on webpages, partnerships, print media, online marketing, social media and referral programs (Baker et al., 2010, pp. 720–721; Callegaro et al., 2014; Kennedy et al., 2016, pp. 6–9). Offline methods are sometimes also used,

[2]Past research has shown that randomly alternating the use of both of the quasi-random "birthday" selection methods ("last" birthday and "next" birthday) yields a more representative within-unit selection of eligibles than using only one method (see Battaglia, Link, Frankel, Osborn, & Mokdad, 2008; Lavrakas, Tompson, & Benford, 2010). Each birthday selection method is biased towards a selection of people born "closest" to the date of the interview, but the use of both methods within the same survey helps to reduce the biases.

but less commonly due to their higher cost. The offline methods include mail outs, recruitment from offline panels, offline surveys, existing marketing databases of mail or email addresses, direct recruitment via telephone, mail or face-to-face and offline media exposure (see Callegaro, Lozar Manfreda, & Vehovar, 2015, pp. 48–51; see Callegaro et al., 2015, p. 207). The recruitment methods used by four of the five nonprobability panel providers that participated in this study are shown in Table 1.[3] Each panel provider was asked to conduct a "nationally representative" survey of 600 respondents from their respective panels. Of note, no instructions were provided as to how this task should be carried out, as we purposely wanted the nonprobability panel providers to generate samples that would reflect their normal sampling approaches.

Quota sampling is a very common approach used by nonprobability panels to select those panelists who will provide data for a questionnaire.[4] Four of the panel providers addressed this task by moving the age, sex, and place of residence questions to the beginning of the questionnaire and using these as screening questions. Data for these primary demographics are gathered from all their panelists by the nonprobability panels at the time the panellist initially joins the panel. These data were gathered anew in our study in the form of screening questions used to impose the respective panel's age, sex and geographic quotas on each nonprobability panel sample in order that they achieved a sample reflecting the distribution of the Australian adult population by these characteristics.

The remaining panel provider (Panel 1) drew its sample to be "Australian Bureau of Statistics representative" and applied quotas to their online survey allowing for ± 5 percentage points variation on the quota groups. To determine how much sample to draw, this panel provider assumed a within-panel 20% completion rate (based on average completion rates for their similar surveys).

Nonprobability panel vendors tend to be secretive about their specific recruitment methods because, as Callegaro et al. (2014) observed, "[they] believe that their [own proprietary] methods provide them a competitive advantage." Nevertheless, we did learn the basics of their recruitment methods. All the online panel providers approached panel members via an email to their personal email address. The common features of this invitation included a direct link to the questionnaire, a description of the length of the questionnaire, mention of the incentive for completing the questionnaire, and the questionnaire completion closing date. Two of the five panels also provided the survey title/topic. One of the providers recruited for multiple surveys at once, inviting panel members to take part in a variety of screening questions and directing to them to one of the surveys for which they were eligible. Other methods of invitation included use of SMS, emails to panel member's panel account, and social media.

4.5 Field Period

Recruitment and data collection for the eight surveys occurred between October and December 2015. Standard response maximization techniques were used for the probability surveys including advance letters, incentives (contingent and noncontingent)[5], several call/contact attempts (for the telephone surveys the maximum number of unanswered calls was capped at three for mobile phone sample records and six for landline records), reminder mailings, choice of data collection mode, and refusal conversion. For the nonprobability panels, email invitations containing a direct link to the questionnaire were sent out by panel providers using their own software. Nonprobability panel members were offered a contingent incentive for completing the questionnaire in accordance with the usual practices of the respective panel company.

4.6 Benchmarks

One of the key objectives of the study was to determine the accuracy of the respective survey estimates relative to independent population benchmarks. The benchmarks used for these purposes were from the Australian national census, high quality surveys undertaken by the Australian Bureau of Statistics, or other federal government agencies or high-quality administrative datasets as shown in Appendix B.

4.7 Response Analysis: Completion Rate

It is not possible to calculate a traditional response rate for nonprobability panels of the general population, because there is no way of knowing anything about the number of persons who actually had an opportunity to join the panel (but did not join) compared to those who did join. Thus, when trying to measure "response" to a given questionnaire such as the one used for our study, the best that can be done

[3]This information was not available for Panel 4.

[4]Quota sampling is the most commonly used method for selection of within-panel samples to complete questionnaires with nonprobability panels (see Callegaro et al., 2014, p. 12). However, YouGov and Toluna in the U.S. use different within-panel sampling approaches (see Callegaro et al., 2014, p. 12). Our experience in the Australian market is that up until quite recently even YouGov has not implemented its U.S. approach, as demonstrated by the polling they did for the 2019 national elections. Our aim in allowing the panels in 2015 to use their own standard approach to sampling and inviting their panel members to complete our questionnaire was to reflect the practices of nonprobability panels in Australia, rather than trying to impose atypical rigor.

[5]A range of noncontingent and contingent incentives were used to try to maximise response rates for the ABS, ANUpoll, and online panel surveys. No incentives were offered to sample members approached as part of the DFRDD survey.

Table 1
Original Recruitment Methods Used to Establish the Online Panels

Recruitment method	Panel 1	Panel 2	Panel 3	Panel 5
Banner advertising on websites	X	-	-	-
Online invitations and messages	X	X	-	-
Partnerships	X	X	X	X
Print media	-	X	X	-
Online marketing	-	X	-	-
Direct mail	-	X	X	-
Social media	-	X	X	-
Other ad hoc initiatives	-	X	-	-
Other survey methods (e.g., CATI, Face-to-Face)	-	-	X	X
Referral programs	-	-	X	X

Source: Compiled from publicly available information and information provided by the panel providers.

is to merely compare the number of panelists within a given nonprobability panel who were invited to complete the questionnaire against the number who actually completed it. This rate is generally known as the "completion rate," and it does not take into account the myriad layers of nonresponse that occurred while the panel was being created and maintained.

As defined by Callegaro and DiSogra (2008, p. 1021), "the completion rate is the proportion of those who completed the [questionnaire] among all the eligible [panelists or respondents] who were invited to take the questionnaire" and "[this rate] appears to be the single most informative [response] metric to report for a volunteer opt-in panel. The interpretation of this rate may reflect the [panelist's] interest in the survey [topic] and/or the ability of the survey company to maximize cooperation" (Callegaro & DiSogra, 2008, p. 1026).

Traditional response rates (c.f American Association for Public Opinion Research, 2016) can be calculated for the three probability samples we used. However, since traditional response rates could not be computed for the nonprobability samples in our study, we are reporting what could be calculated for all eight surveys in our study (i.e., the completion rate).

Table 2 shows the completion rates, following Callegaro and DiSogra (2008) definition, for the three probability surveys and four of the five nonprobability surveys. (This statistic cannot be calculated for Panel 1 as the panel provider would not disclose the number of invitations sent out.) The within-panel completion rate for the nonprobability panels ranged from three percent to 15 percent. For the probability surveys, the completion rates ranged from 15 percent for the RDD survey to 81 percent for the ANUpoll. [6]

4.8 Mode of Data Collection

In planning our study, we recognized that the ABS probability survey would have data collection conducted using self-administered modes (CAWI and mailback paper survey) and that the DFRDD study and part of the ABS and ANUpoll surveys would have interviewer-administered telephone data collection. For the nonprobability samples, all data would be gathered via the self-administered CAWI mode. In contrast, most of the benchmarks were based on data that were gathered via an in-person interviewer-administered mode. Thus, we tried to select questions for the questionnaire used in our eight surveys for which we did not expect appreciable data collection mode effects.

Specifically, two of the three probability sample surveys allowed for mixed modes of data collection. For the ABS survey, which involved initially approaching all sample members by mail, 39 percent of respondents completed the questionnaire online and a similar proportion (38%) mailed the completed questionnaire back in the reply-paid envelope provided. Further, a quarter (23%) completed by telephone in response to outbound telephone reminder efforts. Online was the preferred mode of completion for sample members recruited via the ANUpoll (52%), followed by telephone (41%) and mailback (7%). For the DFRDD probability sample, telephone interviewing (CATI) was the only mode used for data collection. For the five nonprobability samples, CAWI was the only mode used for data collection.

Despite our efforts, we acknowledge that we may have nonignorable data collection mode effects in our comparative datasets, as we discuss later. This limitation is not unique to our study. Of the 23 similar studies documented by Cornesse et al. (2020, pp. 15–17) only two (Chan & Ambrose, 2011; Steinmetz, Bianchi, Tijdens, & Biffignandi, 2014) used web as the only mode of data collection for both the probability and nonprobability samples in their research. Eleven of

[6] An adjusted completion rate for our ANUpoll follow-up survey—one that takes into account all those who were originally sampled for the initial ANUpoll survey (from which the sample was drawn for our study)—is 6.6 percent (i.e., the 560 interviews achieved divided by the 8,493 sample records released to eligible sample members for the initial ANUpoll = 6.6%).

Table 2
Completion rate by survey

	Probability Samples			Nonprobability Samples				
	RDD	ABS	ANU poll	Panel 1	Panel 2	Panel 3	Panel 4	Panel 5
Number of invitations	4,097	2,050	693	N/A	7,097	4,097	6,132	23,527
Number of invitations opened	-	-	-	N/A	2,315	1,241	684	1,314
Number of completes	601	538	560	601	600	626	630	601
Completion rate (%)	14.7	26.2	80.8	N/A	8.5	15.4	10.3	2.6
Time in field (days)	19	48	54	8	8	7	6	4

Source: Compiled from publicly available information and information provided by the panel providers.

their studies used multiple modes of data collection for their probability surveys (face-to-face [F2F]/CAWI, CATI/CAWI, F2F/CATI, CATI/paper survey and CAWI/CATI) and the remainder employed a single non-CAWI mode of data collection for the probability survey component of their studies.

4.9 The Size and Profile of the Offline Population

One of the criticisms made of nonprobability online panels is that they systematically exclude the offline population, thereby creating noncoverage and leaving open the likelihood for nonignorable noncoverage error. At the time of data collection for this study (i.e., 2015), some 86 percent of Australian households had access to the internet at home (Australian Bureau of Statistics, 2016).

In this context, it is interesting to compare the size and profile of the offline population as reflected in the composition of the three probability surveys undertaken as part of this study, all of which sampled both the online and offline populations. To do this we combined the three probability samples (in order to provide a sufficiently large offline sample for analytic purposes) and looked at the online/offline distribution of the combined samples. For the purposes of this analysis, the offline population was defined as people who were reportedly not able to access the internet at home, be it via a broadband connection, a dial-up connection or in some other way, such as through mobile phones or some other mobile device.[7] The size of the Australian offline population for the probability surveys was nine percent unweighted and eight percent weighted. This suggests that the offline population was slightly under-represented in our three probability surveys when compared to the Australian Bureau of Statistics' benchmark at the time of 14%. In contrast, nonprobability online panels completely fail to include the offline population.

4.10 Weighting

Design weight. For the DFRDD surveys (the RDD survey and the ANUpoll) the chance of selection is calculated via the following formula (Baffour et al., 2016; Best, 2010):

$$p = x\frac{S_{\mathrm{LL}}\mathrm{LL}}{U_{\mathrm{LL}}\mathrm{AD}_{\mathrm{LL}}} + \frac{S_{\mathrm{MP}}\mathrm{MP}}{U_{\mathrm{MP}}}$$

where:

- S_{LL} is the number of survey respondents contacted by landline
- U_{LL} is the estimated number of residential landline telephone numbers in Australia
- LL indicates the number of landlines in the respondent's household
- $\mathrm{AD}_{\mathrm{LL}}$ is the number of in-scope adults in the respondent's household
- S_{MP} is the number of survey respondents contacted by mobile
- U_{MP} is the estimated number of allocated mobile phone numbers in Australia
- MP indicates the presence of a mobile phone

For the ABS sample, a single frame design weight was calculated. Because there is no need to adjust for overlapping sample frames and each household has an equal chance of selection into the survey (hence an address weight was not required), only the within household chance of selection is accounted for in the weighting solution. Therefore, the design weight is equal to the number of adults in the household.

Because the probability of selection of the opt-in online nonprobability panels is unknowable, a design weight is not calculable; therefore, a design weight of 1 was assigned to each nonprobability record.

Poststratification weight. After the design weight was calculated, it was then adjusted to try to reduce possible nonresponse and noncoverage error to create a final weight (aka a poststratification weight). Raking was used for this purpose (Valliant, A., & Kreuter, 2013). Doing this enables the weighted estimates to reflect the population not only with respect to those attributes commonly adjusted for, such as age, sex and geography, but also to take into account additional

[7] This is a very conservative definition of the offline population, because not all people who "can" access the Internet do in fact access it or are willing to access it to complete a questionnaire.

parameters such as educational attainment, birthplace, and telephone status. The population benchmarks used are provided in Appendix B (and the benchmark values are shown in the "value column" in Tables C1 and C2).

4.11 Analytic Methods

As noted, a major aim for this study was to learn whether the findings from the previous similar international studies generalize to the Australian research context. Accordingly, the analytical method that we used closely follows the unweighted and weighted approaches used by Chang and Krosnick (2009) and by Yeager et al. (2011).[8] Thus, the following comparisons have been undertaken to identify the between-survey estimates and benchmarks for secondary demographics and substantive measures:

- *Secondary demographics:* Unweighted and weighted survey estimates of the modal response category compared to the corresponding benchmark (Table C2).
- *Substantive measures:* Unweighted and weighted survey estimates of the modal response category compared to the corresponding benchmark (Table C1).
- *Average absolute error:* Defined as the percentage point deviation from the benchmark between unweighted and weighted survey estimates of the modal response category and the corresponding benchmark averaged across secondary demographics (Table C3) and substantive measures (Table C4).

Standard errors of survey estimates and standard errors of average absolute errors were calculated using a bootstrapping procedure. In this instance the bootstrapping procedure as implemented in the R package "boot" ("boot: Bootstrap R (S-Plus) functions," 2015; Davison & Hinkley, 1997) is an accepted methodology for estimation of the sampling distribution of any statistics including sampling errors of probability samples (Baker et al., 2013). However, the methods that are used to estimate sampling errors with probability samples should not be used with nonprobability samples due to nonprobability samples violating key assumptions of probability sampling theory. Although there is no universally agreed method to estimate sampling errors of nonprobability samples, the AAPOR Taskforce Report on Nonprobability Sampling, Baker et al. (2013) cited bootstrapping (or resampling) as one of the acceptable methodologies for reporting the precision of nonprobability-based estimates.[9]

Independent-samples t-tests were conducted to test the null hypotheses of:

- no differences between survey estimates and benchmarks (Tables B1 and B2); and
- no differences between the average of the absolute errors for each pair of surveys includes in the study (Tables B3 and B4).

In addition, following Yeager et al. (2011), summary measures were calculated to illustrate overall accuracy of the eight surveys:

- ranking of average absolute errors across all eight surveys, with the smallest average absolute error ranked as 1.
- number of significant differences from benchmarks for each survey.
- largest percentage point absolute error for each survey.

5 Results

Figures 1 and 2 display a compilation of the results (shown in detail in Appendix C) for various accuracy-metrics in order to provide a high-level comparison between, and within, the probability and nonprobability samples. The first set of comparisons in the left-most region of the figures show the unweighted and the weighted average absolute percentage point errors for the secondary demographics (Figure 1) and for the substantive measures (Figure 2). The second (middle) region of comparisons show the number of statistically significant ($p < .05$) differences for each survey's findings from its respective benchmark, for the secondary demographics (Figure 1) and for the substantive measures (Figure 2), both weighted and unweighted. The right-most region of comparisons in Figures 1 and 2 show the size of the largest percentage point absolute error for each survey, for secondary demographics and the substantive measures, both unweighted and weighted.

[8] A reviewer suggested the use of R-indicators (J. Bethlehem, Cobben, & Schouten, 2011; Schouten & Cobben, 2007; Schouten, Cobben, & Bethlehem, 2009). The R-indicator generally relies on the availability of auxiliary variables for respondents and nonrespondents in order to estimate response propensity. For meaningful comparisons of R-indicators across surveys, we would need also need a consistent set of auxiliary variables across surveys. Minimal auxiliary information is available for the mobile phone portion of the DFRDD samples because mobile phone numbers in Australia are not associated with geography. No auxiliary information at all was available for the nonprobability panels. Although it is possible to calculate response propensities from raked weights (see J. Bethlehem, 2020), as we note in section 5.1 this is of limited utility to our study given that the nonprobability panels set quotas for the sample on primary demographics, effectively forcing the sample distribution to mirror the population. Expanding the calculation of the response propensities to secondary demographics and substantive measures would, in large measure, duplicate the analysis in this section but with less nuance.

[9] Benchmarks sourced from the Australian Census and the Australian Election Commission do not have sampling errors associated with them as these are not sample surveys. Standard errors of the remaining benchmarks were acquired directly from the Australian government agency that conducted the survey or calculated from the survey data using weights and information about sample design provided by the government agency. The size of standard errors for survey-based benchmarks is relatively small (standard errors for all but two measures were less than 1% of the benchmark estimate and the remaining two were less than 5% of the benchmark estimate).

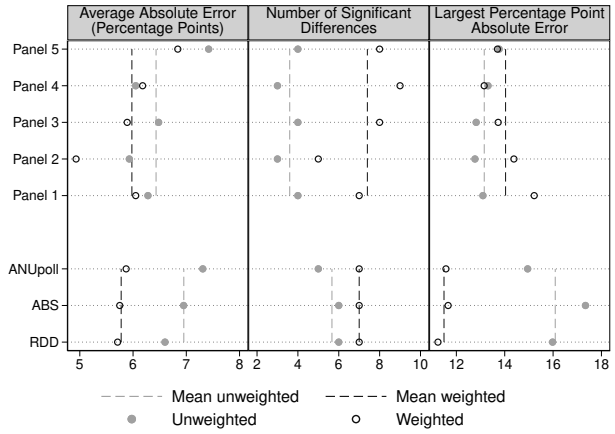

Figure 1. Average Absolute Errors, Number of Significant Differences and Largest Percentage Point Error: Secondary Demographics (Unweighted and Weighted)

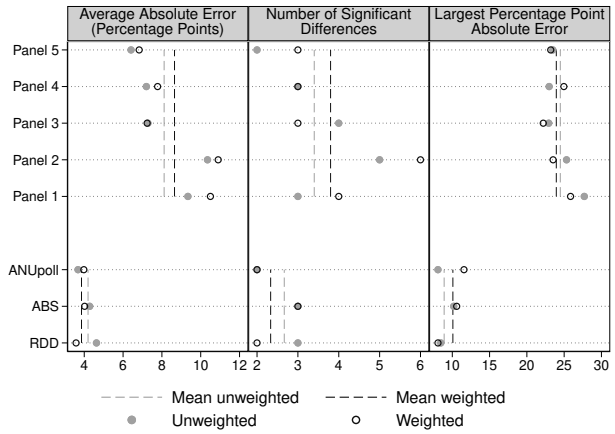

Figure 2. Average Absolute Errors, Number of Significant Differences and Largest Percentage Point Error: Substantive Measures (Unweighted and Weighted)

In the remainder of this section, we present a high-level explanation of the findings. We refer the interested reader to the tables in the Appendix C (Tables C1–C4) for the more detailed statistical results underlying these high-level findings.

5.1 Primary Demographics

As part of weighting, the probability and nonprobability samples were matched to the population distribution for primary demographics. Accordingly, weighted comparisons are not meaningful for primary demographics.

Comparison of unweighted primary demographics is also of limited interest as nonprobability sample providers use these demographics to set quotas for the sample, effectively forcing the sample distribution to mirror the population.

However, as will be seen from the comparison of secondary demographics and substantive measures, imposing population distribution of primary demographics on the sample does not guarantee a representative sample or accuracy of those estimates.

5.2 Secondary Demographics

Unweighted. As shown in Figure 1 (and Table C3), all but one of the nonprobability panels' unweighted estimates of secondary demographics were closer to the benchmarks than the unweighted estimates of the probability sample surveys on each of the three metrics shown. This is likely because the nonprobability panels all imposed age, sex and location quotas thereby also bringing the unweighted sample into closer alignment with a range of secondary demographic benchmarks. The probability sample surveys ranked 5th, 6th, and 7th, out of the eight surveys in terms of unweighted absolute average error on secondary benchmarks (see Table C3). The probability sample surveys also reported the largest percentage point absolute error for these unweighted measures (see Figure 1). The unweighted probability surveys differed significantly from the benchmarks on five to six measures (out of 13), whereas the unweighted nonprobability panels differ significantly from the benchmarks on three to four measures (see Figure 1 and Table C2).

However, when comparing average absolute errors of secondary demographics across surveys (see Table C3), there was only one significant difference between probability and nonprobability surveys: i.e., only the ANUpoll's unweighted absolute average error was significantly different at $p < .05$ from the best nonprobability panel survey (Panel 2).

As quotas were not applied to the probability sample surveys and the weighting was necessary to account for both chances of selection and poststratification adjustments, these results were not unexpected in relation to the secondary benchmarks.

Weighted. As also shown in Figure 1 (and Table C3) and as was expected, weighting improved the accuracy of the probability sample survey estimates relative to secondary benchmarks and brought the accuracy of the probability surveys more in line with that of the nonprobability surveys on the three metrics. The average absolute error for the weighted probability survey estimates of secondary demographics ranged from 5.7 to 5.9 and for the nonprobability panels from 4.9 to 6.8.

There are no significant differences between weighted probability and nonprobability surveys' average absolute errors with respect to secondary demographics. While weighting improved the accuracy across the board, the nonprobability Panel 2 remained the most accurate of all surveys, in terms of average absolute error and the number of significant differences.

5.3 Substantive Measures

For the substantive measures, the probability sample surveys were consistently more accurate when comparing both unweighted and weighted data. This comported with the findings of Kennedy et al. (2016) which showed that balancing the sample on demographic variables (as was the case for the nonprobability panels) is no guarantee of accurate measurement of the substantive/outcome variables of interest.

Unweighted. As shown in Figure 2 (and Table C4), in terms of the unweighted estimates of the substantive measures there were no significant differences in average absolute errors across the probability sample surveys, with ANUpoll having the smallest average absolute error of 3.7 and DFRDD the largest of 4.6.

The average absolute errors of the substantive measures for the unweighted nonprobability panel surveys, as a group, were almost double that of the probability sample surveys (ranging from 6.4 for Panel 5 to 10.3 for Panel 2) and significantly different from all probability surveys. Importantly, although Panel 2 was the best performing sample on demographic variables, it was the worst performing sample on the substantive variables both for weighted and unweighted data (see Table C4). This highlights again the danger of relying on a "demographically balanced" nonprobability panels to provide accurate measurement on substantive measures of interest.

Weighted. Similar to unweighted data, Figure 2 (and Table C4) also shows that for the weighted estimates of the substantive measures there were no significant differences in average absolute errors across probability sample surveys. The standalone DFRDD survey, when weighted, had the lowest average absolute error at (3.6). The biggest improvement as a result of weighting was achieved for the DFRDD estimates, with ANUpoll estimates deteriorating slightly (from 3.7 unweighted average absolute error to 4.0 weighted average absolute error).

As a result of weighting, the average absolute error increased for all nonprobability surveys except for Panel 3 which reduced marginally (from 7.3 to 7.2). Panel 1 recorded the largest increase in average absolute average error attributable to weighting, increasing from 9.3 (unweighted) to 10.5 (weighted).

Weighting brought the probability sample surveys closer together reducing the difference in average absolute error between them. However, weighting had the opposite effect on nonprobability panel surveys, slightly increasing the range of their average absolute errors. Weighting also increased the largest absolute error for ABS, ANUpoll and Panel 4 (see Table C4).

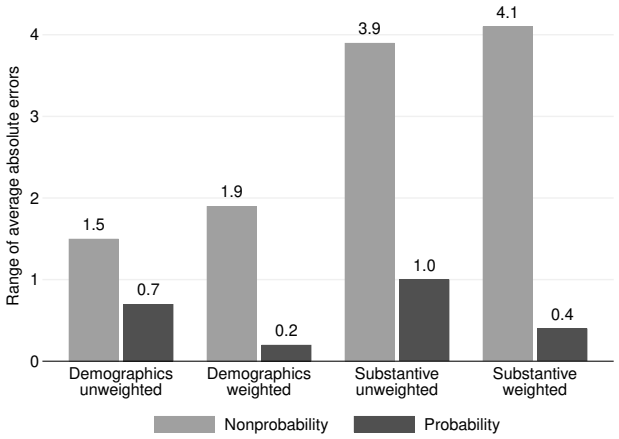

Figure 3. Range of Average Absolute Errors across Surveys.

5.4 Variance/Inconsistency of absolute errors across surveys types

Consistent with prior findings (e.g Chang & Krosnick, 2009; Walker, Pettit, & Rubinson, 2009; Yeager et al., 2011), the probability sample surveys were more consistent in their measurement of both secondary demographics and substantive measures.

As shown in Figure 3 (and Tables C3 and C4), without weighting, the average absolute error for the three probability surveys had a range of variation of 0.7 percentage points among secondary demographics and 1.0 percentage points among substantive measures. In comparison, the five nonprobability panel surveys had corresponding unweighted range of variation of 1.5 and 3.9 percentage points respectively. Corresponding ranges of variation for weighted data were 0.2 and 0.4 percentage points for probability surveys' secondary demographic and substantive measures respectively versus 1.9 and 4.1 percentage points, respectively, for the nonprobability panel surveys' secondary demographics and substantive measures respectively.

6 Discussion

In this section we discuss our findings from a Total Survey Error (TSE) perspective (see Groves, 1989; Groves & Lyberg, 2010), with special attention to different errors associated with the substantive variables that were gathered in the eight surveys.

6.1 Coverage and Coverage Errors Related to Probability Samples and Online Nonprobability Panel Samples

Probability Sample Surveys. Studies that utilize probability samples typically are careful in choosing their sampling frame(s) so as to minimize noncoverage and the possi-

bility of nonignorable coverage error. The three probability sample surveys conducted for this study used frames with extremely high coverage of the Australian residential population. There is no reason to expect that any nonignorable coverage error was present in these three surveys.

Nonprobability Panel Surveys. The five surveys that were fielded on nonprobability-based panels used a variety of convenience frames from which they built their respective panel. They did not cover the Australian residential population well because anyone who did not use the Internet at the times that these panels were undertaking online recruitment would not have been able to consider joining the panel. The noncoverage that was inherent in all these panels is undoubtedly very large and is differential (nonrandom) in nature. It is differential because those who are exposed to an invitation to join nonprobability panels are undoubtably different in many nonignorable ways from those not exposed to such invitations. And these differences are expected to often be correlated with what is being measured in surveys, such as the substantive measures gathered in this study. For example, Fahimi and colleagues (Fahimi et al., 2015) identified significantly different responses between members of probability and nonprobability online panels, after controlling for confounding effects, in relation to factors such as social engagement, self-assertion, shopping habits, happiness and security, politics, sense of community, altruism, survey participation, and internet and social media usage.

Comparison. Uncorrected coverage error in the nonprobability panels was a probable contributing factor to the level of bias in the estimates they generated for the substantive variables measured in the study. It also was likely to be a reason that the nonprobability panel surveys showed considerably more variation in the accuracy of their substantive measures than did the probability sample surveys.

6.2 Sampling and Sampling Errors Related to Probability Sample Surveys and Online Nonprobability Panel Sample Surveys

Probability Sample Surveys. Surveys that use probability samples often take care in deciding how they draw their initially designated sample from their sampling frame(s). This was the case for the three probability sample surveys conducted for our study. Using probability sampling allows the users of such surveys to have a known degree of statistical confidence (associated with sampling error) in those findings. Confidence intervals also can be computed when using probability sampling, which can be used with a known degree of confidence to make decisions about the reliability of findings, including point estimates and differences between sampled subgroups. The extent of the error (variance) associated with sampling can be stated precisely with probability sampling and its meaning is readily understood. Each of the substantive findings generated from the three probability sample surveys in this study can be assigned confidence intervals.

Nonprobability Panel Surveys. There is no single universally accepted measure of "sampling error" for nonprobability samples and there is some debate as to whether the concept of sampling error can even be applied to such samples (see Baker et al., 2010, p. 773). Leaving that aside, methods that have been described in the literature include resampling (as used by Yeager et al. (2011), and in this manuscript), for the use of model-based and pseudo-design-based estimation (see, e.g. Dever & Valliant, 2014), and Bayesian methods such as credibility intervals (see Baker et al., 2013, pp. 63–64), among others. In practice, for the purpose of confidence interval calculation, most nonprobability panel samples are treated as if the sample was drawn from the population of interest using a simple random sample. As noted by Baker et al. (2013), this results in biased estimates of precision and invalid confidence intervals.

Comparison. As shown by the much greater variation across the five nonprobability panel surveys in their estimates of the substantive measures generated by each survey, compared to the much less variation associated with the substantive measures from the probability surveys, the findings for the substantive measures from the set of three probability samples were much more consistent (reliable) than were the findings of the five nonprobability surveys.

6.3 Nonresponse and Nonresponse Errors Related to Probability Sample Surveys and Nonprobability Panel Surveys

Probability Sample Surveys. The size of the nonresponse that occurs can be readily calculated in a probability sample survey. Even when a survey is of members of a probability-based panel, such calculations are easy to make and are essentially the product of the response rate that was achieved when building the panel, the retention rate within the panel, and the completion rate for the particular questionnaire for which panel members were sampled. For probability sample surveys, including those conducted within a probability panel, a number of approaches can be pursued to estimate the extent of nonresponse bias. This also is a function of the nature of the nonresponse that occurred when building the panel, the nonresponse from panel attrition, and the nature of the nonresponse that occurred within the sample/panel for a particular questionnaire. The three probability-based surveys that were conducted as part of this study suffered from considerable nonresponse, albeit at typical levels for these types of surveys in Australia at the time the surveys were conducted. If the nonresponse that occurred was at least in part of a differential nature, as it likely was, poststratification weighting may well have reduced the size of the nonresponse biases that were present in the probability sample data.

Nonprobability Panel Surveys. For nonprobability panel surveys, it is essentially impossible to compute a response rate for the time when the panel was established. That is because it is not known how many persons were exposed to invitations to join the panel. It is commonly understood, however, that far less than one percent of all persons who were exposed to invitations to join a nonprobability panel end up joining (Tourangeau, Conrad, & Couper, 2013, p. 42). Although a completion rate can be calculated for nonprobability panel surveys, this rate does not account for the "response rate" that was experienced when the panel was established or for the attrition rate that occurred during the life of the panel. The completion rates for the questionnaires from four of the nonprobability sample surveys in our study that reported such information to us were in line with what is common for such approaches. As such, with opt-in nonprobability panel surveys there is no well-accepted scientific approach to account for the amount or nature of the nonresponse biases that may have occurred for a given survey.

Comparison. In addition to the large amount of noncoverage associated with nonprobability online panel surveys, they also have an appreciably (nonignorably) higher level of nonresponse than do probability sample surveys. The much greater amount of nonresponse for nonprobability panel surveys, compared to surveys using probability samples/panels, occurs at the stages when the panels are built, during the lifetime of the panel (i.e., panel attrition), and each time that panelists are invited to complete a questionnaire. The latter form of nonresponse can occur due to problems associated with the contact of panelists when email invitations to complete a questionnaire are sent out and with gaining cooperation from contacted panellists (see Callegaro et al., 2015, pp. 132–135). Although a considerable amount of nonresponse in probability samples occurs when contact is first made with sample members, this amount is miniscule when compared to that which occurs in the recruitment of nonprobability samples/panels. Nonresponse at the retention and questionnaire-completion stages within probability samples and probability panel surveys is also far less than with nonprobability samples and nonprobability panel surveys. Differential nonresponse, which is one of the two primary mechanisms that makes nonresponse bias possible, is likely to be more of a problem with nonprobability surveys than with probability surveys for many reasons. For example, little or no effort is made in nonprobability panels to try to motivate ongoing cooperation among those exposed to the initial recruitment invitation for a particular questionnaire or among those who join but attrite from the panel. In contrast, with probability samples/panels, considerable resources are typically committed to counter differential nonresponse, in an effort to minimize its effects on nonresponse biases. Therefore, in the case of the probability sample surveys in our study, their nonresponse errors are likely much smaller than for the nonprobability panel surveys.

6.4 Weighting and Adjustment Errors Related to Probability Samples and Nonprobability Panel Samples

Probability Sample Surveys. Our probability sample surveys were designed to be adjusted for selection probability and to conform to population distributions via weighting. On the whole, and as expected, weighted estimates for these surveys are more accurate than unweighted estimates. However, the weighting methodology for the ANUpoll did not make any adjustment for the two-stage selection process for these respondents. This process has potential to introduce additional nonresponse and noncoverage errors that are not as effectively corrected for by the poststratification weights as the single stage probability sample surveys. Among the three probability samples, the DFRDD sample has the best weighting efficiency and the ABS has the least.

Nonprobability Panel Surveys. Ideally, weighting for nonprobability panel surveys should correct for biases that are present in these panels, such as overrepresentation of "early adopters," those with reduced social engagement, a lower rate of volunteering, a greater engagement with internet and social media, etc. (see Fahimi et al., 2015). However, this is rarely done in practice. Instead, all units are usually given a design weight of 1.0 and standard poststratification adjustments are applied. This ignores the enforcement of quotas on these variables and a variety of proprietary (aka "secret") mechanisms used by nonprobability panel providers to try to have their samples resemble the population on standard poststratification dimensions. This practice often results in reduced rather than improved accuracy (Kennedy et al., 2016), as illustrated by the reduced accuracy for substantive measures for all but one nonprobability panel surveys in this study.

Comparison. Weighting generally reduces TSE for probability samples, with the probability sample surveys' estimates having lower average absolute deviations from the substantive benchmarks than unweighted probability samples, as well as lower average deviations than weighted and unweighted nonprobability panels. Without accounting for adjustment errors that are known to be present in nonprobability panel surveys, the "probability-style weighting" of nonprobability samples often is more likely to increase rather than decrease TSE.

6.5 Measurement and Measurement Errors related to Probability Sample Surveys and Nonprobability Panel Surveys

As noted previously, the items in our questionnaire were chosen for several reasons including our desire to minimize data collection mode effects across the eight surveys.

The questions that were used in the eight surveys were almost identical, so we believe there is little reason to expect

any differential questionnaire-related measurement errors associated with whether probability sample surveys were used to gather the data or the data were from online nonprobability panel surveys. Therefore, questionnaire-related error will not be addressed further.

But there are other forms of measurement error that can be addressed, in particular respondent-related measurement error, interviewer-related measurement error, and data collection mode measurement error.

Probability Sample Surveys. It is common with surveys that are based on probability samples for considerable care to be given to data quality. This includes attention to interviewer training and monitoring when using interviewer-administered data collection. It also includes attention to the manner in which respondents may create measurement error in the form of bias and variance. The data in the ANUpoll and the ABS survey (both of which used probability samples) are likely affected by using/combining data from the different data collection modes (online, paper survey, and telephone) that were used in those surveys, as past work has suggested that for some questions the mode of data collection will affect the answers provided (see Kreuter, Presser, & Tourangeau, 2008). This can be a disadvantage in mixed-mode data collection surveys. One reason for this is that it is difficult to sort out and correct for potential differential data collection mode effects. Furthermore, many probability sample surveys that are interviewer-administered suffer from the joint interviewer-related and respondent-related error of social desirability, especially when sensitive questions are asked (Kreuter et al., 2008). In addition, respondents and interviewers also may contribute to measurement error in the form of recency effects whereby response alternatives that are heard most recently by the respondent are more likely to be chosen than those heard earlier (Holbrook, 2008). However, when using self-administered data collection with probability sample surveys, primacy effects—whereby answers read first by the respondent (i.e., at the beginning of a list of response choices) are more likely to be chosen than those read last (i.e., at the end of the list of choices; see, Scanlan (2008). Thus, primacy effects appear to occur more frequently with self-administered data collection than with interviewer-administered data collection modes.

Nonprobability Panel Surveys. It is generally accepted that members of general population online nonprobability panels, as a group, tend be more likely to provide certain respondent-related measurement errors, than do respondents in probability surveys and probability online panels (see Baker et al., 2014; Greszki, Meyer, & Schoen, 2014; Hillygus, Jackson, & Young, 2014). The five nonprobability surveys in our study all used self-administered online data collection (aka CAWI). Given the acknowledged data quality problems that arise from the behaviours of some opt-in panelists, our panel providers exercised what have become standard practices for them and took steps to exclude "poor quality" responses from the final data provided to us. These steps included removing "straight-liners," removing "junk"/poor quality responses to open ended questions and removing speeders. Speeders were variously defined by our nonprobability panel providers as completing the questionnaire in less than three minutes, completing the questionnaire within an unspecified departure from the average completion time, and completing the questionnaire in a time of one third or more below median completion times. One panel provider also gave extra scrutiny to panelists who were flagged in their panel data base as having previously provided poor quality responses.

On the other hand, social desirability and recency effects seem to occur less frequently with self-administered (e.g., CAWI) data collection of the type that is used for online nonprobability panel surveys compared to interviewer-administered data collection which is generally not used with nonprobability panel samples. Yet, self-administered data collection modes such as CAWI (which was used to gather data in all our nonprobability panels) seem more likely than interviewer-administered data collection to contribute to primacy effects, whereby answers read first (i.e., at the beginning of a list of choices) by the respondent are more likely to be chosen than those read last (i.e., at the end of the list of choices); see, Scanlan (2008).

Comparison. Measurement errors associated with interviewers, respondents, and/or mode of data collection are likely present in both our probability sample surveys and our nonprobability panel surveys. The probability sample surveys used interviewers to gather at least some of the data. Thus interviewer-related error may be present in these surveys but would not be present in any of the nonprobability panel surveys since no interviewers were involved in data collection. In particular, the self-administered data collection mode used for all the nonprobability panel surveys should have reduced social desirability error associated with questions that respondent might deem sensitive. Other respondent-related measurement errors are likely present in all eight surveys regardless of their sample type, but for different reasons. For example, speeding and satisficing are reasoned to be more a problem for the nonprobability panel data that were collected via a self-administered online mode and not as likely for the probability sample data. Primacy effects are more likely to be present for the nonprobability panel surveys, whereas recency effects are more likely to be a problem for the probability surveys. Whether the quality of the data that respondents provided was higher in the probability sample surveys or in the nonprobability panel surveys is not possible for us to be confident about. However, given that the probability sample surveys had less error (bias and variance) for our substantive measures, it is possible that higher data quality in the probability surveys was part of the explanation

for those findings.

7 Conclusions

7.1 Strengths and Limitations of Probability Samples

The three probability sample surveys in this study likely had little Coverage Error, a known amount of Sampling Error, a nonignorable amount of Nonresponse Error, little Adjustment Error, and a small-to-modest but ultimately unknown amount of Measurement Error.

Overall, the three probability sample surveys, as a group, were found to be less biased on the substantive measures and had less variance from the benchmark values, compared to the nonprobability panel surveys.

7.2 Strengths and Limitations of Nonprobability Panels

The five nonprobability surveys in this study likely had a nonignorable amount of Coverage Error, an unknowable amount of Sampling Error, a nonignorable amount of Nonresponse Error, an unknown amount of Adjustment Error, and a small-to-modest amount but ultimately unknown amount of Measurement Error.

Overall, the five nonprobability panel surveys, as a group, were found to be more biased on the substantive measures and had more variance from the benchmark values, compared to the probability sample surveys. However, as a group, they were essentially comparable to the probability samples (as a group) in their overall accuracy for the demographic measures.

7.3 Implications

The Australian results that we have presented and discussed are in close agreement with the findings reported in the past decade+ in several European countries and the United States about comparisons between the statistics generated by probability sample surveys and those generated by nonprobability panel surveys. That is, our probability sample surveys were consistently found to be more accurate than our nonprobability panel surveys in measuring substantive variables. We also showed what others have found about variability in accuracy across surveys by sample type: that is, different nonprobability panel surveys tend to vary much more from each other in their accuracy than do different probability sample surveys from each other. Furthermore, when we recall that (a) four of the five nonprobability panel companies used in this study complied with all of the ESOMAR "28 Questions to Help Buyers of Online Panels;" (b) the other company partially complied, and (c) three of the five nonprobability panel companies were ISO 26362 accredited, it seems that compliance with these standards is no guarantee of less bias and/or reduced variability.

We found that the average size in percentage points of the differences in average absolute error (see Table 3) for the three probability sample surveys from their respective benchmarks vs. the five nonprobability panel surveys from their respective benchmarks was 0.5 percentage points for the unweighted secondary demographics, −0.2 percentage points for the weighted secondary demographics, −3.9 percentage points for the unweighted substantive measures, and −4.8 percentage points for weighted substantive measures. That is, the difference between the average errors for secondary demographics was less than 1 percentage point for both unweighted and weighted results in both types of samples. For the substantive measures, however, the average differences in accuracy for the nonprobability sample surveys were essentially four to five percentage points worse than for the probability surveys.

We acknowledge that the size of these average absolute error differences may not matter in many instances for those who are funding surveys. That is, a nonprobability sampling approach when measuring the variable domains that we measured, may be fit for purpose for many who are funding surveys. However, a very real problem for those choosing a nonprobability panel approach is that they cannot be as confident about the accuracy of the specific nonprobability panel survey provider's data that they receive as they could be about the accuracy of the data that they would receive were they funding a probability sample survey. For example, and as shown in Figure 2, two of the nonprobability panels had average absolute errors of greater than nine percentage points for both their unweighted and weighted estimates of the substantive measures. And for particular individual statistics, the errors were even larger (see Table C1, where the largest error in accuracy from its respective benchmark among the five nonprobability samples' statistics was greater than 20 percentage points). We believe that errors of these magnitudes are likely to lead the entity funding surveys that are fielded on such a nonprobability panel to draw nonignorably incorrect conclusions from at least some of the data they would receive.

Therefore, those trying to decide whether to fund a probability sample survey or a nonprobability sample survey should recognize that a probability sample survey they fund would (a) likely be more accurate in measuring their substantive statistics and (b) that they can be more confident of that, than if they funded a nonprobability panel survey to measure the same statistics. In the latter case, we recognize that a survey funder/client might merely by luck/chance choose a nonprobability panel survey that is close enough in accuracy to a probability survey's accuracy, and close enough to the true values of what they are measuring, to meet their needs. However, they may unluckily choose a nonprobability panel survey that falls woefully short of the accuracy levels they require.

Thus, a very important problem that those choosing to fund nonprobability sample surveys are left with is that they

Table 3
Average Absolute Error

Summary Metric	Probability Sample Survey Average pp error	Nonprobability Sample Surveys Average pp error	Difference between Probability and Nonprobability Sample Surveys pp
Secondary Demographics			
Unweighted	6.9	6.4	0.5
Weighted	5.8	6.0	−0.2
Substantive Measures			
Unweighted	4.2	8.1	−3.9
Weighted	3.9	8.7	−4.8

will not likely know whether they have funded a good (accurate) nonprobability panel survey or a bad (inaccurate) one for their particular purposes/needs. Simply put, the funding entity for a given survey needs to decide if they trust their luck enough when they choose a nonprobability sampling method, ostensibly because of its lower cost and quicker turnaround time. The funder also needs to decide "How great a risk do we run if the nonprobability panel survey that we are funding is not accurate enough for our needs," and "Is the cost of that risk greater than the added cost that we would face by instead funding a probability survey—one that we can be more confident will be accurate enough for our needs?"

Acknowledgement

The authors gratefully acknowledge the contributions of the Social Research Centre and the ANU Centre for Social Research and Methods for supporting this study as well as the *Life in Australia*™ panelists for their ongoing participation in social science survey research of this nature.

References

American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th ed.). American Association for Public Opinion Research. Retrieved from https://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx

Australian Bureau of Statistics. (2016). Household use of information technology, Australia, 2014–15. Catalogue number 8146.0. Canberra, Australian Bureau of Statistics. Retrieved from https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8146.0Main+Features12014-15?OpenDocument%5C#:%5C~%7B%7D:text=In%5C%202014%5C%E2%5C%80%5C%9315%5C%20for%5C%20those,parts%5C%20of%5C%20Australia%5C%20(79%5C%25)

Baffour, B., Haynes, M., Western, M., Pennay, D., Misson, S., & Martinez, A. (2016). Weighting strategies for combining data from dual-frame telephone surveys: Emerging evidence from australia. *Journal of Official Statistics*, *32*(3). doi:10.1515/jos-2016-0029

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., … Lavrakas, P. J. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4). doi:10.1093/poq/nfq048

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J., … Tourangeau, R. (2013). *Report of the AAPOR task force on nonprobability sampling*. American Association for Public Opinion Research. Retrieved from https://www.aapor.org/AAPOR%5C_Main/media/MainSiteFiles/NPS%5C_TF%5C_Report%5C_Final%5C_7%5C_revised%5C_FNL%5C_6%5C_22%5C_13.pdf

Baker, R., Miller, C., Kachhi, D., Lange, K., Wilding-Brown, L., & Tucker, J. (2014). Validating respondents' identity in online samples. In M. Callegaro, R. Baker, J. Bethlehem, A. Göritz, J. A. Krosnick, & P. Lavrakas (Eds.), *Online panel research: A data quality perspective*. John Wiley & Sons.

Battaglia, M. P., Link, M. W., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). An evaluation of respondent selection methods for household mail surveys. *Public Opinion Quarterly*, *72*(3). doi:10.1093/poq/nfn026

Best, J. (2010). *First-stage weights for overlapping dual frame telephone surveys*. Paper presented at the 65th Annual Conference of the American Association of Public Opinion Research, Chicago, IL, May 15.

Bethlehem, J. (2020). Working with response probabilities. *Journal of Official Statistics*, *36*(3). doi:10.2478/jos-2020-0033

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. John Wiley & Sons.

Blom, A. G., Ackermann-Piek, D., Helmschrott, S., Cornesse, C., Bruch, C., & Sakshaug, J. (2018). *An evaluation of sample accuracy in probability-based and nonprobability surveys*. Under review.

boot: Bootstrap R (S-Plus) functions. (2015). R package version 1.3-17. Retrieved from https://cran.r-project.org/package=boot

Brüggen, E., van den Brakel, J., & Krosnick, J. (2016). *Establishing the accuracy of online panels for survey research*. Discussion Paper 2016-04. Amsterdam: Statistics Netherlands. Retrieved from https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A., Krosnick, J., & Lavrakas, P. J. (2014). *Online panel research: A data quality perspective*. John Wiley & Sons.

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, *72*(5). doi:10.1093/poq/nfn065

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. Sage.

Chan, P., & Ambrose, D. (2011). Canadian online panels: Similar or different? *Vue Magazine*, *January/February*. Retrieved from https://inthevue.com/magazine/

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4). doi:10.1093/poq/nfp075

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., de Leeuw, E. D., Legleye, S., … Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1). doi:10.1093/jssam/smz041

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge University Press.

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2).

Dever, J. A., & Valliant, R. (2014). *Estimation with nonprobability surveys and the question of external validity*. Paper presented at the 2014 International Methodology Symposium, Statistics Canada, Gatineau, QC, Canada, October 29. Retrieved from https://www.statcan.gc.ca/eng/conferences/symposium2014/program/14288-eng.pdf

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). Hoboken, NJ: Wiley.

Dutwin, D., & Buskirk, T. D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, *81*(S1). doi:10.1093/poq/nfw061

Erens, B., Burkill, S., P., C. M., Conrad, F., Clifton, S., Tanton, C., … J., C. A. (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. doi:https://doi.org/10.2196/jmir.3382

ESOMAR. (2020). *Global market research: An ESOMAR industry report*. ESOMAR. Retrieved from https://ana.esomar.org/documents/global-market-research-2020

Fahimi, M., Buttermore, N., Thomas, R. K., & Barlas, F. M. (2015). Scientific surveys based on incomplete sampling frames and high rates of nonresponse. *Survey Practice*, *8*(6). doi:10.29115/SP-2015-0031

Fine, B. (2016). *Online research panels around the world: The situation in australia*. Paper presented at the Current State and Future of Online Research in Australia Workshop of the ANU Centre for Social Research and Methods and the Social Research Centre, Canberra, Australia, July 14.

Geoscape. (N.d.). G-NAF: The geocoded address database for australian businesses and governments. Retrieved from https://geoscape.com.au/data/g-naf/

Greszki, R., Meyer, M., & Schoen, H. (2014). The impact of speeding on data quality in nonprobability and freshly recruited probability-based online surveys. In M. Callegaro, R. Baker, J. Bethlehem, A. Göritz, J. A. Krosnick, & P. Lavrakas (Eds.), *Online panel research: A data quality perspective*. John Wiley & Sons.

Groves, R. (1989). *Survey errors and survey costs*. Wiley.

Groves, R., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5). doi:10.1093/poq/nfq065

Hillygus, S., Jackson, N., & Young, M. (2014). Professional respondents in nonprobability online surveys. In M. Callegaro, R. Baker, J. Bethlehem, A. Göritz, J. A. Krosnick, & P. Lavrakas (Eds.), *Online panel research: A data quality perspective*. John Wiley & Sons.

Holbrook, A. (2008). Recency effect. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (2nd ed.). Thousand Oaks, CA: Sage Pub.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). *Evaluating online nonprobability surveys*. Pew Research Center. Retrieved from http://www.pewresearch.org/2016/05/02/evaluating-online-nonprobability-surveys/

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5). doi:10.1093/poq/nfn063

Lavrakas, P. J., Tompson, T., & Benford, R. (2010). *Investigating the errors that occur with within-unit respondent selection*. Paper presented at the 65th annual conference of the American Association for Public Opinion Research, Chicago, IL, May 12.

Legleye, S., Charrance, G., Razafindratsima, N., Bajos, N., Bohet, A., & Moreau, C. (2018). The use of a nonprobability internet panel to monitor sexual and reproductive health in the general population. *Sociological Methods and Research*, *47*(2). doi:10.1177\%2F0049124115621333

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, *82*(4). doi:10.1093/poq/nfy038

Malhotra, N., & Krosnick, J. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with nonprobability samples. *Political Analysis*, *15*(3). doi:10.1093/pan/mpm003

Neiger, D., Pennay, D., Ward, A., & Lavrakas, P. J. (2017). Inference from nonprobability samples. Paper presented at the Joint Conference of Survey Research Methods (SRM), the European Survey Research Association (ESRA) and Étude Longitudinal par Internet Pour les Sciences Sociales (ELIPSS), March 16–17, Paris. Retrieved from https://www.europeansurveyresearch.org/news/non-prob/INPS_05_Neiger.pptx

Pennay, D. W., Neiger, D., Lavrakas, P. J., & Borg, K. (2018). The online panels benchmarking study: A total survey error comparison of findings from probability-based surveys and nonprobability online panel surveys in australia. CSRM & SRC Methods Paper No. 2/2018. Canberra, Australia: Australian National University, Centre for Social Research & Methods. Retrieved from https://csrm.cass.anu.edu.au/research/publications/online-panels-benchmarking-study-total-survey-error-comparison-findings

Saris, W. E. (1998). Ten years of interviewing without interviewers: The telepanel. In M. P. Couper, R. P. Baker, C. Z. F. Clark, W. L. Nicholls II, & J. M. Reilly (Eds.), *Computer assisted survey information collection* (pp. 409–429). New York: John Wiley & Sons.

Scanlan, C. R. (2008). Primacy effect. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (2nd ed.). Thousand Oaks, CA: Sage Pub.

Scherpenzeel, A. C., & Bethlehem, J. G. (2011). How representative are online panels? problems of coverage and selection and possible solutions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies*. Routledge.

Schouten, B., & Cobben, F. (2007). R-indexes for the comparison of different fieldwork strategies and data collection modes. Discussion Paper 07002. Voorburg, Netherlands: Statistics Netherlands. Retrieved from http://hummedia.manchester.ac.uk/institutes/cmist/risq/schouten-cobben-2007-a.pdf

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1).

Sohlberg, J., Gilljam, M., & Martinsson, J. (2017). Determinants of polling accuracy: The effect of opt-in internet surveys. *Journal of Elections, Public Opinion and Parties*, *27*(4). doi:10.1080/17457289.2017.1300588

Steinmetz, S., Bianchi, A., Tijdens, K., & Biffignandi, S. (2014). Improving web survey quality: Potentials and constraints of propensity score adjustments. In M. Callegaro, R. Baker, J. Bethlehem, A. Göritz, J. A. Krosnick, & P. Lavrakas (Eds.), *Online panel research: A data quality perspective*. John Wiley & Sons.

Sturgis, P., Kuha, J., Baker, N., Callegaro, M., Fisher, S., Green, J., … Smith, P. (2018). An assessment of the causes of the errors in the 2015 UK general election opinion polls. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(3). doi:10.1111/rssa.12329

Szolnoki, G., & Hoffmann, D. (2013). Online, face-to-face and telephone surveys: Comparing different sampling methods in wine consumer research. *Wine Economics and Policy*, *2*(2). doi:10.1016/j.wep.2013.10.001

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.

Valliant, R., A., D. J., & Kreuter, F. (2013). *Practical tools for designing and weighting sample surveys*. Springer.

Walker, R., Pettit, R., & Rubinson, J. (2009). The foundations of quality initiative. *Journal of Advertising Research*, *49*(4). doi:10.2501/S0021849909091089

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and nonprobability samples. *Public Opinion Quarterly*, *75*(4). doi:10.1093/poq/nfr020

Appendix A
Question wording

(Appendix A follows on next page)

Table A1
Question wording

CATI survey	Online self-complete	Hard copy self-complete	Benchmark
This section helps us to understand more about the lifestyle, health and wellbeing of Australians. Firstly, on a scale from 0 to 10, where zero means “not at all satisfied” and 10 means “completely satisfied”, overall, how satisfied are you with life as a whole these days?	This section helps us to understand more about the lifestyle, health and wellbeing of Australians. Firstly, on a scale from 0 to 10, where zero means “not at all satisfied” and 10 means “completely satisfied”, overall, how satisfied are you with life as a whole these days? Please enter a number between 0 and 10	This section helps us to understand more about the lifestyle, health and wellbeing of Australians. On a scale from 0 to 10, where zero means “not at all satisfied” and 10 means “completely satisfied”, overall, how satisfied are you with life as a whole these days?	General Social Survey, 2014: The following question asks how satisfied you feel, on a scale from 0 to 10. Zero means you feel “not at all satisfied” and 10 means “completely satisfied”. Overall, how satisfied are you with life as a whole these days?
And in general would you say that your health is excellent, very good, good, fair, or poor.	And in general would you say that your health is excellent, very good, good, fair, or poor.	And in general would you say that your health is excellent, very good, good, fair, or poor.	National Health Survey, 2011–12, General Social Survey: In general would you say that your health is excellent, very good, good, fair or poor?
The following questions are about how you have been feeling over the past 4 weeks. In the past 4 weeks, about how often did you feel <A>? Would you say <B >? A: Nervous, hopeless, restless or fidgety, so depressed that nothing could cheer you up, that everything was an effort, worthless. B: All of the time, most, some, a little, or, none of the time.	In the past 4 weeks, about how often did you feel <A>? Would you say <B >? A: Nervous, hopeless, restless or fidgety, so depressed that nothing could cheer you up, that everything was an effort, worthless. B: All of the time, most, some, a little, or, none of the time.	In the past 4 weeks, about how often did you feel <A>? Would you say <B >? A: Nervous, hopeless, restless or fidgety, so depressed that nothing could cheer you up, that everything was an effort, worthless. B: All of the time, most, some, a little, or, none of the time.	The following questions are about your feelings in the past 4 weeks. In the past 4 weeks, about how often did you feel <A>? A: All of the time, most, some, a little, or, none of the time

Continues on next page

Table A1 (Continued from previous page

CATI survey	Online self-complete	Hard copy self-complete	Benchmark
I would now like to ask you some questions about smoking and alcohol consumption. How often do you now smoke cigarettes, pipes or other tobacco products? Would you say daily, at least weekly but not daily, less often than weekly, not at all, but I have smoked in the last 12 months, or, not at all, and I have not smoked in the last 12 months.	The next few questions are about smoking and alcohol consumption. How often do you now smoke cigarettes, pipes or other tobacco products? Daily, At least weekly but not daily , Less often than weekly, Not at all, but I have smoked in the last 12 months, Not at all, and I have not smoked in the last 12 month.	The next few questions are about smoking and alcohol consumption. How often do you now smoke cigarettes, pipes or other tobacco products? Daily, At least weekly but not daily, Less often than weekly, Not at all, but I have smoked in the last 12 months, Not at all, and I have not smoked in the last 12 months.	National Drug Strategy Household Survey, 2013: How often do you now smoke cigarettes, pipes or other tobacco products? Daily At least weekly (but not daily), Less often than weekly, Not at all, but I have smoked in the last 12 months, Not at all and I have not smoked in the last 12 months.
Have you had an alcoholic drink of any kind in the last 12 months?	Have you had an alcoholic drink of any kind in the last 12 months?	Have you had an alcoholic drink of any kind in the last 12 months?	National Drug Strategy Household Survey, 2013: Have you had an alcoholic drink of any kind in the last 12 months?
How often did you have an alcoholic drink of any kind in the last 12 months, was it every day, 5 to 6 days a week, 3 to 4 days a week, 1 to 2 days a week, 2 to 3 days a month, about 1 day a month, less often, or do you, no longer drink.	How often did you have an alcoholic drink of any kind in the last 12 months? Every day, 5 to 6 days a week, 3 to 4 days a week, 1 to 2 days a week, 2 to 3 days a month, About 1 day a month, Less often, No longer drink	How often did you have an alcoholic drink of any kind in the last 12 months? Every day, 5 to 6 days a week, 3 to 4 days a week, 1 to 2 days a week, 2 to 3 days a month, About 1 day a month, Less often, No longer drink	National Drug Strategy Household Survey, 2013: In the last 12 months, how often did you have an alcoholic drink of any kind? Every day, 5 to 6 days a week, 3 to 4 days a week, 1 to 2 days a week, 2 to 3 days a month, About 1 day a month, Less often, No longer drink

Continues on next page

Table A1 (Continued from previous page

CATI survey	Online self-complete	Hard copy self-complete	Benchmark
On a day that you (have/had) an alcoholic drink, how many standard drinks (do/did) you usually have? A standard drink is equal to 1 pot of full strength beer, 1 small glass of wine or 1 pub-size nip of spirits. Would it be 20 or more drinks, 16–19 drinks, 13–15 drinks, 11–12 drinks, 9–10 drinks, 7–8 drinks, 5–6 drinks, 3–4 drinks, 2 drinks, 1 drink, or, half a drink.	On a day that you have an alcoholic drink, how many standard drinks do you usually have? A standard drink is equal to 1 pot of full strength beer, 1 small glass of wine or 1 pub-size nip of spirits. 20 or more drinks, 16–19 drinks, 13–15 drinks, 11–12 drinks, 9–10 drinks , 7–8 drinks, 5–6 drinks, 3–4 drinks, 2 drinks, 1 drink, half a drink	On a day that you have an alcoholic drink, how many standard drinks do you usually have? A standard drink is equal to 1 pot of full strength beer, 1 small glass of wine or 1 pub-size nip of spirits. 20 or more drinks, 16–19 drinks, 13–15 drinks, 11–12 drinks, 9–10 drinks , 7–8 drinks, 5–6 drinks, 3–4 drinks, 2 drinks, 1 drink, half a drink	National Drug Strategy Household Survey, 2013: On a day that you have an alcoholic drink, how many standard drinks do you usually have? (see the coloured "Standard Drinks/Instruction Card" provided to you, or the chart on page 16). 20 or more drinks, 16–19 drinks, 13–15 drinks, 11–12 drinks, 9–10 drinks , 7–8 drinks, 5–6 drinks, 3–4 drinks, 2 drinks, 1 drink, Half a drink

Appendix B
Sources used for independent benchmark measures

Table B1
Source list

Measure	Source
Primary demographics	
Sex	Australian Bureau of Statistics, Estimated Resident Population June 2015, Cat. 3101.0
Age	Australian Bureau of Statistics Estimated Resident Population June 2015, Cat. 3101.0
Region	Australian Bureau of Statistics, Census 2011,Table Builder (2011) and Australian Bureau of Statistics Estimated Resident Population June 2015, Cat. 3101.0
Educational Attainment	Australian Bureau of Statistics, Census 2011,Table Builder (2011), QAL LP Non-School Qualification: Level of Education by AGEP, Persons aged 18 years and over, Place of Usual Residence
Country of birth	Australian Bureau of Statistics, Table Builder (2011), BPLP - 4 Digit Level by AGEP, Persons aged 18 years and over, Place of Usual Residence
Telephone status	Australian Communications and Media Authority (2015), Communications Report, 2014-15
Secondary demographics	
Australian citizenship	Australian Bureau of Statistics, Census 2011, Table Builder (2011), CITP by AGEP, Persons aged 18 years and over, Place of Usual Residence.
Enrolled to vote	Australian Electoral Commission, 31 December 2015, http://www.aec.gov.au/Enrolling_to_vote/Enrolment_stats/index.htm
Indigenous status	Australian Bureau of Statistics, Census 2011, Table Builder (2011), INGP by AGEP, Persons aged 18 years and over, Place of Usual Residence.
Language other than English at home	Australian Bureau of Statistics, Census 2011, Table Builder (2011), LANP - 2 Digit Level by AGEP, Persons aged 18 years and over, Place of Usual Residence.
Geographic mobility	Australian Bureau of Statistics, Census 2011, Table Builder (2011), UAI5P by AGEP, Persons aged 18 years and over, Place of Usual Residence.
Remoteness	Australian Bureau of Statistics, Census 2011, Table Builder (2011), RA by AGEP, Persons aged 18 years and over, Place of Usual Residence.
Employment status	Australian Bureau of Statistics, Census 2011, Table Builder (2011), EMTP by AGEP, Persons aged 18 years and over, Place of Usual Residence.
Wage and salary income	Australian Bureau of Statistics, National Health Survey, 2014-15, Persons aged 18 years and over, employed income groups.
Household tenure	Australian Bureau of Statistics, Census 2011, Table Builder (2011), TEND, Dwellings: Location on Census Night
Household composition	Australian Institute of Health and Welfare, National Drug Strategy Household Survey, 2013.
Socio-economic status	Australian Bureau of Statistics, Socio-Economic Indexes for Areas, 2011.
Substantive measures	
Life satisfaction	Australian Bureau of Statistics, General Social Survey, Summary Results Australia, 2014,

Continued on next page

Continued from previous page

Measure	Source
Psychological distress (Kessler 6)	Australian Bureau of Statistics, National Health Survey, 2014-15. Persons aged 18 years and over, psychological distress, Australia.
General health	Australian Bureau of Statistics, National Health Survey, 2014-15. Persons aged 18 years and over, self-assessed health status, Australia.
Private Health Insurance	Australian Bureau of Statistics, National Health Survey, 2014-15. Persons aged 18 years and over, private health insurance, Australia
Daily smoker	Australian Institute of Health and Welfare, National Drug Strategy Household Survey, 2013
Alcoholic drink of any kind in the last 12 months	Australian Institute of Health and Welfare, National Drug Strategy Household Survey, 2013

Appendix C
Tables

Table C1
Survey estimates of modal response category and the corresponding benchmark for substantive measures

		Probability Samples			Nonprobability Sample Internet Surveys				
Substantive Measures Benchmark comparison	Value	RDD	ABS	ANU poll	Panel 1	Panel 2	Panel 3	Panel 4	Panel 5
Life satisfaction (8 out of 10)	32.60								
Unweighted Estimate		34.61	30.11	31.25	21.80*	20.17*	27.48*	23.81*	24.63*
Weighted Estimate		34.50	30.58	30.60	20.67*	21.03*	28.11	23.38*	24.72*
Psychological distress—Kessler 6 (Low)	82.20								
Unweighted Estimate		73.97*	76.05*	75.59*	54.50*	56.88*	59.27*	59.21*	58.72*
Weighted Estimate		74.12*	71.61*	70.63*	56.34*	58.68*	60.00*	57.24*	59.00*
General Health Status (SF1) (Very good)	36.20								
Unweighted Estimate		30.62	34.39	33.75	33.28	32.67	32.59	32.38	36.94
Weighted Estimate		33.55	36.55	34.20	32.06	30.36*	30.89	31.24	37.73
Private Health Insurance	57.10								
Unweighted Estimate		65.56*	67.29*	65.18*	53.08	49.00*	53.35	59.52	58.57
Weighted Estimate		60.35	60.48	59.05	48.22*	44.59*	53.42	56.46	54.54
Daily smoker	13.52								
Unweighted Estimate		10.32*	9.11*	12.50	21.80*	20.17*	17.25*	14.76	15.64
Weighted Estimate		15.12	9.37*	17.03*	23.33*	20.21*	17.41*	16.19	17.84*
Consumed alcohol in the last 12 months	81.87								
Unweighted Estimate		82.20	82.53	84.46	79.53	75.83*	77.32*	77.94*	79.20
Weighted Estimate		85.87*	85.48*	84.75	79.49	76.61*	77.99*	77.66*	80.38

Note: All errors are deviations from the benchmark.
* $p < 0.05$

Table C2
Survey estimates of modal response category and the corresponding benchmark for secondary demographics

		Probability Samples			Nonprobability Sample Internet Surveys				
Secondary Demographics Benchmark comparison	Value	RDD	ABS	ANU poll	Panel 1	Panel 2	Panel 3	Panel 4	Panel 5
Indigenous status (Non-Indigenous)	98.10								
Unweighted Estimate		98.84	97.96	98.39	97.50	96.50*	98.40	98.41	98.84
Weighted Estimate		98.76	98.40	98.42	97.90	96.49	98.09	98.27	98.83
Australian citizen	83.93								
Unweighted Estimate		91.01*	94.42*	92.32*	93.01*	90.50*	93.13*	90.63*	94.68*
Weighted Estimate		86.60	92.00*	86.56	91.81*	88.05	91.04*	90.76*	92.95*
Enrolled to vote	78.47								
Unweighted Estimate		88.19*	92.57*	90.36*	86.86*	86.00*	88.50*	86.83*	91.51*
Weighted Estimate		83.06	88.68*	83.02	84.95*	80.18	85.59*	84.75*	89.21*
Living at current address 5 years ago	54.80								
Unweighted Estimate		69.55*	69.14*	67.50*	61.56*	61.00*	64.22*	63.81*	68.55*
Weighted Estimate		62.10*	54.68	58.44	59.79*	58.12	61.80*	63.66*	65.89*
Currently employed	59.39								
Unweighted Estimate		58.24	57.43	60.54	51.08*	54.33*	53.99*	55.71	50.25*
Weighted Estimate		69.34*	64.60	66.43*	49.14*	53.33*	53.11*	51.00*	49.04*
Voluntary work (No)	74.22								
Unweighted Estimate		58.24*	60.78*	60.18*	72.55	73.83	71.09	68.89*	71.05
Weighted Estimate		62.65*	62.99*	62.56*	74.51	77.14	71.46	69.86*	70.65
Language other than English (No)	75.72								
Unweighted Estimate		84.19*	81.23*	86.96*	82.70*	84.17*	85.62*	80.00*	84.03*
Weighted Estimate		85.45*	80.38	84.47*	85.09*	85.37*	87.53*	84.75*	85.30*
Most disadvantaged area-based socioeconomic status) (quintile)	20.00								
Unweighted Estimate		14.98*	14.50*	11.79*	16.81*	14.67*	14.38*	13.49*	13.81*
Weighted Estimate		13.76*	15.08*	10.27*	16.97	14.75*	14.85*	14.52*	14.14*
Resident of a major city	70.22								
Unweighted Estimate		69.05	72.68	69.11	76.04*	61.83*	68.05	77.30*	75.37*
Weighted Estimate		69.03	72.92	69.76	73.15	69.88	68.31	72.63	71.81
Access the internet from home	85.90								
Unweighted Estimate		86.86	87.92	92.50*	99.00*	98.67*	98.72*	99.21*	99.67*
Weighted Estimate		89.57*	91.81*	93.07*	98.76*	98.24*	97.91*	99.05*	99.60*
Home ownership with a mortgage	29.61								
Unweighted Estimate		30.95	32.34	33.57	31.78	30.17	33.87*	33.81*	31.61
Weighted Estimate		33.75	39.96*	37.40*	28.56	30.92	30.41	29.91	28.00
Couple with dependent children	38.35								
Unweighted Estimate		22.80*	21.00*	23.39*	26.79*	25.83*	27.00*	29.21*	29.95*
Weighted Estimate		27.90*	28.19*	26.97*	23.12*	23.97*	24.62*	25.46*	28.11*
Wage and salary income $ 1,000–$ 1,249 p/wk	13.80								
Unweighted Estimate		9.97	14.14	14.33	9.76	12.06	13.17	14.54	11.76
Weighted Estimate		11.78	12.81	15.04	9.73	12.77	12.96	15.93	12.90

Note: All errors are deviations from the benchmark.
* $p < 0.05$

Table C3
Pairwise t-tests Comparing Average Absolute Errors on Secondary Demographics using Bootstrapped Standard Errors

	Probability Samples			Nonprobability Sample Internet Surveys				
Secondary Demographics Survey	RDD	ABS	ANU poll	Panel 1	Panel 2	Panel 3	Panel 4	Panel 5
Unweighted								
Average Absolute Error	6.60	6.95	7.31	6.28	5.93	6.48	6.05	7.42
Pairwise differences								
RDD	-	−0.35	−0.72	0.32	0.67	0.12	0.55	−0.82
ABS	0.35	-	−0.36	0.67	1.02	0.47	0.9	−0.47
ANUpoll	0.72	0.36	-	1.03	1.38*	0.83	1.26^{+}	−0.11
Panel 1	−0.32	−0.67	−1.03	-	0.35	−0.2	0.23	−1.14^{+}
Panel 2	−0.67	−1.02	−1.38*	−0.35	-	−0.55	−0.12	−1.49*
Panel 3	−0.12	−0.47	−0.83	0.20	0.55	-	0.43	−0.94
Panel 4	−0.55	−0.90	−1.26^{+}	−0.23	0.12	−0.43	-	−1.37*
Panel 5	0.82	0.47	0.11	1.14^{+}	1.49*	0.94	1.37*	-
Ranking	5	6	7	3	1	4	2	8
Weighted								
Average Absolute Error	5.71	5.75	5.87	6.05	4.93	5.89	6.18	6.84
Pairwise differences								
RDD	-	−0.05	−0.17	−0.34	0.78	−0.18	−0.47	−1.14
ABS	0.05	-	−0.12	−0.29	0.83	−0.13	−0.42	−1.09
ANUpoll	0.17	0.12	-	−0.17	0.95	−0.01	−0.30	−0.97
Panel 1	0.34	0.29	0.17	-	1.12	0.16	−0.13	−0.80
Panel 2	−0.78	−0.83	−0.95	−1.12	-	−0.96	−1.25^{+}	−1.92*
Panel 3	0.18	0.13	0.01	−0.16	0.96	-	−0.29	−0.96
Panel 4	0.47	0.42	0.30	0.13	1.25^{+}	0.29	-	−0.67
Panel 5	1.14	1.09	0.97	0.8	1.92*	0.96	0.67	-
Ranking	2	3	4	6	1	5	7	8

$^{+}$ $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table C4
Pairwise t-tests Comparing Average Absolute Errors on Substantive Measures using Bootstrapped Standard Errors

	Probability Samples			Nonprobability Sample Internet Surveys				
Substantive Measures Survey	RDD	ABS	ANU poll	Panel 1	Panel 2	Panel 3	Panel 4	Panel 5
Unweighted								
Average Absolute Error	4.63	4.28	3.68	9.34	10.35	7.28	7.20	6.41
Pairwise differences								
RDD	-	0.35	0.95	-4.71^{***}	-5.71^{***}	-2.65^{*}	-2.57^{**}	-1.77^{*}
ABS	−0.35	-	0.60	-5.06^{***}	-6.06^{***}	-3.00^{**}	-2.92^{**}	-2.12^{*}
ANUpoll	−0.95	−0.60	-	-5.66^{***}	-6.66^{***}	-3.60^{**}	-3.52^{***}	-2.72^{**}
Panel 1	4.71^{***}	5.06^{***}	5.66^{***}	-	−1	2.06^{+}	2.14^{+}	2.94^{**}
Panel 2	5.71^{***}	6.06^{***}	6.66^{***}	1	-	3.06^{*}	3.14^{**}	3.94^{***}
Panel 3	2.65^{*}	3.00^{**}	3.60^{**}	-2.06^{+}	-3.06^{*}	-	0.08	0.88
Panel 4	2.57^{**}	2.92^{**}	3.52^{***}	-2.14^{+}	-3.14^{**}	−0.08	-	0.79
Panel 5	1.77^{*}	2.12^{*}	2.72^{**}	-2.94^{**}	-3.94^{***}	−0.88	−0.79	-
Ranking	3	2	1	7	8	6	5	4
Weighted								
Average Absolute Error	3.58	4.02	3.98	10.50	10.90	7.24	7.78	6.83
Pairwise differences								
RDD	-	−0.44	−0.40	-6.92^{***}	-7.32^{***}	-3.67^{**}	-4.20^{**}	-3.25^{**}
ABS	0.44	-	0.03	-6.49^{***}	-6.88^{***}	-3.23^{*}	-3.76^{**}	-2.81^{*}
ANUpoll	0.40	−0.03	-	-6.52^{***}	-6.92^{***}	-3.26^{*}	-3.80^{**}	-2.85^{*}
Panel 1	6.92^{***}	6.49^{***}	6.52^{***}	-	−0.40	3.26^{*}	2.72^{+}	3.67^{**}
Panel 2	7.32^{***}	6.88^{***}	6.92^{***}	0.40	-	3.66^{*}	3.12^{*}	4.07^{**}
Panel 3	3.67^{**}	3.23^{*}	3.26^{*}	-3.26^{*}	-3.66^{*}	-	−0.54	0.41
Panel 4	4.20^{**}	3.76^{**}	3.80^{**}	-2.72^{+}	-3.12^{*}	0.54	-	0.95
Panel 5	3.25^{**}	2.81^{*}	2.85^{*}	-3.67^{**}	-4.07^{**}	−0.41	−0.95	-
Ranking	1	3	2	7	8	5	6	4

$^{+}$ $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$