Results from the 2022 Australian Comparative Study of Survey Methods

D. Pennay, B. Phillips, D. Neiger, A. Ward, S. Slamowicz, A. Lethborg

August 2023

Social
Research
Centre

About the Social Research Centre

The Social Research Centre provides research and evaluation services to Australia's social science research community in order create new knowledge, inform decision-making and advance understanding of society.

The Social Research Centre is owned by and has deep links with the Australian National University (ANU) and is co-founder the ANU Centre for Social Research and Methods.

Authors

Darren Pennay is the Founder and former CEO of the Social Research Centre Pty Ltd. He is an Honourary Professor in the Practice of Survey Methodology at the ANU Centre for Social Research & Methods.

Dr Benjamin Phillips is Chief Survey Methodologist and Director, Life in Australia™ Operations at the Social Research Centre and Campus Visitor at the ANU Centre for Social Research & Methods.

Dr Dina Neiger is Chief Statistician at the Social Research Centre and Campus Visitor at the ANU Centre for Social Research & Methods.

Andrew Ward is Principal Statistician at the Social Research Centre.

Sam Slamowicz is a Survey Statistician at the Social Research Centre.

Anna Lethborg is a Research Director, Quantitative Research Consulting, at the Social Research Centre.

Abstract

Many studies comparing the accuracy of survey estimates generated from probability-samples and non-probability samples have been undertaken over the last fifteen years. This study (the Australian Comparative Study of Survey Methods – ACSSM) is one of only a few to build upon a previous study, thereby enabling not only point-in-time comparisons of the relative accuracy of estimates generated from probability and non-probability sample surveys, but also the relativity of these comparisons over time.

The ACSSM compares the results from eight parallel surveys of the residential Australian adult population. The survey methods used are (1) computer-assisted telephone interviewing (CATI) with persons contactable via randomly generated mobile phone numbers, (2) mixed-mode (computer-assisted web interviewing [CAWI] and CATI) interviews via a probability-based research panel, (3) video-assisted live interviewing (VALI) via a probability-based research panel, (4) using SMS push-to-web to obtain completed CAWI questionnaires from respondents via a random sample of mobile phone numbers and (5–8) four samples provided by four non-probability online panels.

We find that non-probability online panel surveys are cheaper, quicker, and generally less accurate, but sometimes only slightly so, than the probability-based alternatives. Within the limitations of this comparative analysis, there is also evidence that the accuracy gap in favour of probability-based surveys over non-probability online panel surveys may have narrowed in recent years. While both types of surveys produce more accurate measures of the same set of items in 2022 than in 2015, the non-probability surveys show greatest improvement.

The results generated from probability-based surveys are less variable than those obtained when the same questionnaire is administered to members of non-probability online panels. This lower variability, along with the increased methodological disclosure generally associated with probability-based surveys, provides survey researchers with grounds to be more confident in the results generated from probability-based surveys than those generated from non-probability online panels. We also find, although more equivocally than previous studies, that weighting is more effective in reducing bias for probability-based surveys than surveys conducted on non-probability online panels, for which weighting sometimes increases bias.

A pertinent issue remains for those choosing to fund non-probability sample surveys in that, for any given survey, or any given items within a survey, researchers have a less solid basis from which to affirm the accuracy and generalisability of their results than if the same questionnaire is administered to a probability-based sample. Nor can they be as confident as to whether they should use weighted or unweighted data.

It still does seem to be the case that if one wishes to generalise from a sample to the inferential population, that probability-based surveys of the general population allow one to do so with more confidence than do non-probability online panel surveys. The cost one is prepared to pay for this increased accuracy and increased confidence is the dilemma, with survey researchers—including academic survey researchers—turning increasingly to the cheaper non-probability online panels.

We conclude with a plea for transparency, especially about the recruiting and sampling practices used by non-probability panel providers. Greater transparency can only enhance the credibility of non-probability panels overall and may lead to new methodological insights which further improve the accuracy of the estimates generated from such panels. If this occurs, discerning survey researchers will have more reason for confidence in the survey estimates generated from non-probability online panels.

Acknowledgments

The authors gratefully acknowledge the contribution of the entire study team as listed below, comprised of staff from the Social Research Centre and the Australian Bureau of Statistics (ABS), as well as external experts from Australia, Germany, and the U.S.

The study was primarily funded by the Social Research Centre and the ABS also made a substantial financial contribution.

Thanks also to Dr Paul J Lavrakas and Dr Carina Cornesse for their helpful comments when reviewing this paper, to the Life in Australia™ panellists for their ongoing participation in the panel and to the interviewers who collected the CATI and VALI data.

This report makes extensive use of the ACSSM Technical Report (Phillips et al., 2023), which is available upon request.

Study Team

Role	Team Members	Affiliations
Study Leads	Dr Dina Neiger, AStat Dr Benjamin Phillips	Social Research Centre, ANU
Research	Simran Kothiya Anna Lethborg Dale VanderGert Joel Watt	Social Research Centre
Statistics and Methods	Jack Barton	Social Research Centre
	Kinto Behr	Social Research Centre
	Phil Carmo	ABS
	Kirsten Gerlach	ABS
	Sandra Ropero	Social Research Centre
	Sam Slamowicz	Social Research Centre
	Andrew Ward, AStat	Social Research Centre
Data Science	Wendy Guo Storm Logan Dinah Lope Ryan Tian	Social Research Centre
Operations Team	Clea Chiller Grant Lester Sam Luddon Meagan Jones Julie Olivine The interviewing team	Social Research Centre
Advisory Group	Dr Kylie Brosnan	Social Research Centre
	Dr Carina Cornesse	DIW Berlin, University of Bremen, University of Mannheim
	Emma Farrell	ABS
	Diane Herz	Social Research Centre
	Dr Paul J Lavrakas	Social Research Centre, University of Illinois Chicago, NORC at the University of Chicago
	Dr Paul Myers	Social Research Centre, ANU
	Darren Pennay	Social Research Centre, ANU

Declaration of interests

This study was conducted and primarily funded by the Social Research Centre. The probability-based online sample evaluated in this study—Life in Australia™—is owned and operated by the Social Research Centre. This report expresses the views of the authors and not necessarily the views of the Social Research Centre.

Acronyms

AAPOR	American Association for Public Opinion Research
ABS	Australian Bureau of Statistics
A-BS	Address-Based Sampling
ACSSM	Australian Comparative Study of Survey Methods
CATI	Computer-assisted telephone interviewing
CAWI	Computer-assisted web interviewing
COMR	Completion Rate
CONR	Consent Rate
CUMRR2	Cumulative Response Rate 2
DFRDD	Dual-frame random digit dialling
G-NAF	Geo-coded National Address File
IPND	Integrated Public Number Database
IVR	Interactive Voice Response
MRP	Multi-level Regression with Poststratification
OPBS	Online Panels Benchmarking Study
OPBS+	Online Panels Benchmarking Study plus Life in Australia™
PROR	Profile Rate
RDD	Random digit dialling
RECR	Recruitment Rate
RETR	Retention Rate
RR	Response Rate
RR3	AAPOR Response Rate 3
SMS	Short message service (aka text message)
VALI	Video-assisted live interviewing

Contents

1 Introduction

In 2010, an American Association for Public Opinion Research (AAPOR) Task Force Report (Baker et al., 2010) comparing the accuracy and validity of survey findings from probability-based surveys with those from non-probability (opt-in) online panels reached the following conclusions:[1]

- ‘Researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values.
- The few studies that have disentangled mode of administration from sample source indicate that nonprobability samples are generally less accurate than probability samples.
- There are times when a nonprobability online panel is an appropriate choice. Not all research is intended to produce precise estimates of population values, and so there may be survey purposes and topics where the generally lower cost and unique properties of Web data collection are an acceptable alternative to traditional probability-based methods’ (Baker et al., 2010, 714).

While the 2010 AAPOR Task Force Report refers to there being ‘few studies’ in this area, since 2010 there have been many studies comparing the accuracy of survey findings generated from probability-based sampling methods with those from non-probability online panels.

A comprehensive recent review of 25 comparative studies by Cornesse and colleagues (Cornesse et al., 2020) found that the higher accuracy of probability sample surveys has persisted and been demonstrated across various topics, such as voting behaviour, sexual behaviour and attitudes and socio-demographics. The higher accuracy of probability samples has also been reported in several countries including Australia, France, Germany, the Netherlands, Sweden, the United Kingdom, and the United States.

Furthermore, Cornesse and her co-authors found that the higher accuracy of probability-based surveys has been shown over time, with the first study demonstrating this undertaken in 2007 (Malhotra & Krosnick, 2007) to the most recent ones (Blom et al., 2018; Legleye et al., 2018; MacInnis et al., 2018; Sturgis et al., 2018). Cornesse et al. (2020, 15) conclude that ‘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’.

The initial Australian contribution to this field, the Online Benchmarking Study (OPBS), was undertaken in 2015 (Pennay et al., 2018; Kaczmirek et al., 2019; Lavrakas et al., 2022). This current study is known as the Australian Comparative Study of Survey Methods (ACSSM).

So why undertake another study comparing the relative accuracy of survey findings from probability-based samples with those generated from non-probability online panels? The reasons, in brief, are as follows: 1) coming seven years after the first study enables us to compare the current versus historical accuracy of the probability-based surveys and surveys conducted on non-probability online panels, 2) the ACSSM incorporates new and emerging probability-based sample survey designs not included in the 2015 study, 3) the new study has access to a wider range of benchmarks than the 2015 study thereby enabling more robust comparisons, 4) the methods used to analyse and weight the data generated from probability-based and non-probability sample surveys have continued to evolve, providing the opportunity for these new methods to be evaluated, 5) the context for this study, and survey research generally, has changed considerably since 2015 as a result of the impact of the COVID-19 pandemic lockdowns on survey response dynamics, and 6) the use of online research continues to grow, having increased from 24 per cent to 32 per cent of global market research industry revenue between 2013 and 2021 (ESOMAR, 2014 and 2021). An additional challenge impacting the survey environment in Australia in 2022 was the occurrence of several unrelated large-scale data privacy breaches, with the potential to negatively affect participation in both probability-based surveys and non-probability panels.[2]

An understanding of the timeline of our previous Australian research into this topic helps provide additional context for the current study. The initial study, the 2015 OPBS, as reported by Pennay and colleagues (2018) compared the findings from three probability-based surveys (two administered using computer-assisted telephone interviewing [CATI] using a dual-frame random digit dialling [DFRDD] sample and one administered to an address-based sample [A-BS] of households using online, mail-back and telephone modes of data collection, with five surveys administered to samples from non-probability online panels. The OPBS provided a basis for the establishment of the first, and still only, probability-based online panel in Australia, Life in Australia™, in November 2016.

The same questionnaire as used in the OPBS was administered to members of the newly established Life in Australia™ panel in January–February 2017, thereby enabling these results to be added to the original OPBS comparisons (see Kaczmirek et al., 2019). This means that the comparative accuracy of the estimates generated from Life in Australia™ relative to benchmarks and the other modes of sampling and data collection were undertaken when Life in Australia™ was just established. A key point of interest for the ACSSM is how Life in Australia™ estimates perform relative to other contemporary and emerging survey options now that the panel is in its eighth year.[3]

2 Previous Australian research

An overview of the survey methods and findings from the previous OPBS and OPBS+ studies provide important context for the current study.

The original 2015 OPBS comprised of eight surveys:

- A standalone DFRDD CATI survey fielded in November–December 2015, with 50 per cent of interviews completed via the landline frame and 50 per cent via the mobile frame (*n*=601).
- A mail survey fielded in November–December 2015 (*n*=538). The sampling frame used for this survey was the Geo-coded National Address File (G-NAF)[4] with questionnaires being mailed to households. To accommodate situations in which more than one person in a household was in-scope, the printed instructions on the questionnaire asked for the person aged 18 years or over with either the next birthday or the most recent birthday (alternating) to complete the questionnaire. Questionnaires could be completed in three ways: by mailing back the completed questionnaire in the envelope provided, online by following the instructions provided with the survey covering letter or by telephone when responding to a reminder call.
- The October 2015 ANUpoll (*n*=560). A DFRDD survey with a 60:40 split between landline and mobile phone interviews. Respondents who completed the ANUpoll, in October 2015 were invited to take part in a follow up survey, the OPBS, which was introduced to respondents as a 'survey about health and wellbeing.' Those who agreed to participate in the follow up survey provided their contact details. Out of 1,200 respondents who completed the ANUpoll, 693 agreed to be re-contacted. Depending on their preferences, these individuals were either emailed a link to complete the OPBS questionnaire online or mailed a questionnaire to return via the mail. Telephone reminder action and telephone-based data collection were also undertaken.
- Eight non-probability panel providers were invited to quote to undertake a 'nationally representative' survey of 600 respondents from their respective panels, to be fielded in November and December 2015. Instructions on how this task should be carried out were not provided. Five quotes were received and four panels selected based on the amount of paradata they could provide. Price was not part of the selection criteria.

Table 1 (next page) shows that, after weighting, the probability and non-probability sample surveys generally performed similarly with respect to the measurement of secondary demographics (i.e., those demographic variables not used for weighting). The average absolute bias (AAB) ranging from 4.3 per centage points (pp) for Panel 2 to 6.3pp for Panel 4, with Life in Australia™ the most accurate of the probability-based surveys (AAB – 5.0pp).

- With respect to the substantive measures, the standalone DFRDD CATI survey was the least biased (3.6pp) followed by Life in Australia™ (4.0pp). The probability-based surveys were all more accurate than the non-probability online panels.
- Overall, when substantive and secondary measures were combined in the OPBS+, Life in Australia™ was the least biased of the nine surveys compared in 2015. These results were consistent with the expectations of superior accuracy of the probability-based sample survey estimates compared with the non-probability online panel surveys.
- The findings from the 2015 Australian research as reported in Lavrakas et.al. (2022), Kaczmirek et al. (2019), and Pennay et al. (2018) accord with those of Yeager et al. (2011) and with the vast

majority of the subsequent studies in finding that:

- (non-probability) surveys done via the internet were less accurate, on average, than probability-based surveys regardless of mode of administration
- there was considerable variation in accuracy among the findings of non-probability samples, and much more so than among probability samples, and
- post-stratification with primary demographics sometimes improved the accuracy of non-probability sample surveys and sometimes reduced their accuracy.

- Yeager et al. (2011, 709) concluded that their results are consistent with the 'conclusion that non-probability samples yield data that are neither as accurate as nor more accurate than data obtained from probability samples'.

Table 1 Summary of average absolute bias from the 2015 OPBS+

Variable	Average absolute bias, probability surveys				Average absolute bias, non-probability surveys				
	DF-RDD	A-BS	ANU Poll	Life in Australia™	Panel 1	Panel 2	Panel 3	Panel 4	Panel 5
Secondary demographics	5.9	5.7	5.8	5.0	5.5	4.3	5.4	5.6	6.3
Substantive variables	3.6	4.0	4.0	4.6	10.5	10.9	7.2	7.8	6.8
Combined	5.1	5.2	5.2	4.8	7.2	6.5	6.0	6.3	6.5
Rank (lowest has least error)	2	3	4	1	9	8	5	6	7

Source: Kaczmirek et al. (2019, 25).

3 Study objectives

To our knowledge, only two previous studies have compared the accuracy of probability and non-probability sample surveys over time (MacInnis et al., 2016; Yeager et al., 2011). Both studies showed the ongoing superiority of estimates generated from probability-based surveys compared with those produced by non-probability online panels. The ACSSM is the first comparative study of this kind since the lifting of most COVID-19 pandemic restrictions and, as such, provides a contemporary view of the relative performance of probability-based and non-probability sampling and survey methods and how this may have changed over time. This study is also timely given the increased use of non-probability sampling methods by academics and practitioners across the social science disciplines (Rivera, 2019, 1) and the increase, in the U.S. at least and likely elsewhere, in the use of probability-based online panels for election polling and other public opinion research and the continuing decline of CATI (Kennedy, Popky & Keeter, 2023).

Given this framing, the ACSSM has two overarching objectives and several secondary aims, which will be explored and further developed over time.

The two overarching objectives are:

1) Evaluating contemporary and emerging practices for general population surveys, and

2) Improving contemporary and emerging practices for general population surveys.

In the context of these objectives, as well as the changing survey research landscape in Australia and around the world, the research aims of the study include, but are not limited to, the following:

- Comparing contemporary estimates from surveys administered on probability and non-probability sampling frames against each other and against external benchmarks.
- Understanding how the accuracy of the data generated by probability and non-probability sample surveys have changed over time, including variation between and within different surveys.
- Identifying differences in sample profiles between probability and non-probability panels to inform blending, weighting, and fit-for-purpose sampling solutions.
- Comparing the impact of various weighting methods on the accuracy of survey estimates produced from probability and non-probability samples.
- Exploring the differences in the multivariate relationships within and across sampling frames.
- Gaining insight into the motivations of survey respondents recruited through different modes and via different sampling frames, and
- Analysing response quality using available response metrics such as speeding, straight-lining, satisficing, use of non-substantive response options and the use of *non sequiturs* in verbatim responses.

4 Survey design and performance

4.1 Study overview

The study involved undertaking eight parallel surveys of the residential Australian adult population, (i.e., persons aged 18 years and over), using the sample frames, recruitment and data collection modes outlined in Table 2. The geographic coverage is residents of the six Australian states, the Northern Territory and Australian Capital Territory. Residents of the Jervis Bay Territory and Australian External Territories are excluded.

The survey components are briefly described below. Three different sampling frames are used for the eight surveys:

1) Life in Australia™ is the sampling frame for the (i) Video Assisted Live Interviewing (VALI) survey[5] and (ii) the standard mixed mode Life in Australia™ survey
2) Mobile phone numbers generated using random digit dialling (RDD) provide the sampling frame for both (iii) the CATI survey and (iv) the short messaging service (SMS) push-to-web survey.[6]
3) Four non-probability online panels (v—viii) provide the sampling frames for the non-probability surveys, all of which used an online mode of data collection.

The initial ACSSM plan was to undertake nine surveys including both a high effort and low effort CATI survey.[7] The high effort CATI survey used autodialler technology to dial numbers when requested to do so by an interviewer and adopted what would be described as fairly rigorous, but not atypical, contact and response maximisation protocols. The low effort CATI survey, which was commenced and abandoned, used predictive autodialler technology (which dials numbers in the background in anticipation of an interviewer being available – and can often result in an annoying delay when the call is answered before an interviewer comes on the line) and protocols designed to try and obtain interviews as quickly and cheaply as possible without efforts to boost household contact rates and response rates.[8]

The original reason for conducting high effort and low effort CATI surveys was to enable a survey cost versus survey accuracy comparison between the two approaches. Unfortunately, the low-effort CATI survey had to be abandoned part-way through fieldwork due to a combination of technical and configuration issues impacting the predictive autodialler, thus rendering the survey paradata unsuitable for our comparative purposes. However, tests of association between high and low effort CATI surveys showed only minor differences across the demographic and substantive variables between the two executions. On this basis, we concluded that predictive autodialler settings did not impact on the data collected, making it possible to include all the completed interviews in our analysis.

4.2 Methodology

A description of the methodology for each survey follows. The methodological detail provided in this paper is thought to be sufficient to enable readers to understand the differences between each of the ACSSM surveys and how these differences might contribute to the differences in the resultant estimates. For those interested in the complete methodological description, the survey technical report (Phillips et al., 2023) is available upon request.

4.2.1 Sampling frames

Life in Australia™ provides the sample frame for the VALI and Life in Australia™ surveys. Life in Australia™ is a probability-based research panel which includes people with and without internet access by virtue of using a mixed mode of data collection. The vast majority (>95%) of panellists' complete questionnaires online with the offline population included via CATI. Given the very small proportion of surveys completed via CATI, henceforth in this paper Life in Australia™ is referred to as a probability-based online panel.

Table 2 Summary of ACSSM surveys

Sampling Method	#	Survey	Sampling frame(s)	Recruitment mode(s)	Invitation mode(s)	Interview mode(s)	Sample Sizes Initiated	Sample Sizes Achieved	Incentives	Field dates
Probability-based surveys	1	VALI	Life in Australia™ (panellists recruited from the following frames: DFRDD, mobile RDD and A-BS using the G-NAF)	CATI, interactive voice response (IVR), mail push-to-web, SMS push-to-web	Email and SMS invitations; email, SMS, and telephone reminders; online booking system for VALI appointments	VALI	1,399	600	$10 voucher / donation	23 Nov – 20 Dec 2022
	2	Life in Australia™	As above	As above	Email and SMS (online only), telephone	Online, CATI	796	582	$10 voucher / donation	5 – 19 Dec 2022
	3	CATI high effort	Mobile RDD	CATI	CATI, pre-notification SMS	CATI	8,958	498	None	5 – 18 Dec 2022
	4	CATI low effort*	Mobile RDD	CATI	CATI, pre-notification SMS	CATI	23,040	305	None	5 – 13 Dec 2022
	5	SMS push-to-web	Mobile RDD	SMS	SMS	Online	20,000	599	$10 voucher	5 – 17 Dec 2022
Non-probability online panels	6	Non-probability Panel 1	Opt-in panel, nationally representative quotas	Open enrolment, email, affiliates (e.g., loyalty programs), online and physical ads, social media influencers	Panel portal	Online	Unknown	850	Points- or miles-based rewards	5 – 14 Dec 2022
	7	Non-probability Panel 2	Opt-in panel, nationally representative quotas	Mail, affiliates, online and physical ads, social media, personal invitations	Email	Online	8,952	852	Points-based rewards	5 – 13 Dec 2022
	8	Non-probability Panel 3	Opt-in panel, nationally representative quotas	Mail, telephone, online and physical ads, social media	Email	Online	11,070	891	Points-based rewards	7 – 16 Dec 2022

Sampling Method	#	Survey	Sampling frame(s)	Recruitment mode(s)	Invitation mode(s)	Interview mode(s)	Sample Sizes Initiated	Sample Sizes Achieved	Incentives	Field dates
	9	Non-probability Panel 4	Opt-in panel, nationally representative quotas	Open enrolment, online and physical ads, social media, member referral	Panel portal	Online	Unknown	853	Dollar-based rewards	5 – 16 Dec 2022

Notes: VALI sample initiated refers to panellists invited to set a VALI appointment. See Final Outcomes and Dispositions for further details. * The low-effort CATI arm using a predictive dialler was abandoned part-way through the experiment due to a combination of technical and configuration issues impacting one of the diallers thus rendering the results in relation to costs, call cycle and productivity invalid and unusable for comparison. Analysis of completes from the two arms confirmed that dialling issues did not impact on the collection of data, making it possible to combine all CATI interviews as a single arm for the purpose of analysis.

The small amount of research into the use of VALI indicates that recruiting for VALI is most effective when there is an established relationship between the research agency/sponsor and the potential survey respondents (McGonagle and Sastry, 2021). Given that Life in Australia™ is owned by the Social Research Centre, with panellists invited to complete a questionnaire every month, there is an established history of survey participation with the Social Research Centre. By virtue of this pre-existing relationship, it was felt that Life in Australia™ would be well-suited to use as a platform from which to recruit VALI participants. A further benefit of using Life in Australia™ as the VALI sample source is that having two surveys conducted on samples drawn from Life in Australia™, the standard online survey and the VALI survey, would enable direct mode comparisons, while controlling for the sampling frame.

To ascertain the feasibility of using Life in Australia™ as the VALI sample source, in July 2022, a subset of panellists were asked to indicate their willingness to participate in a VALI survey later in the year. Of the 3,441 panellists who were asked, 1,447 (42%) indicated an in-principle willing to participate, 1,553 (45%) were unwilling and 441 (13%) were unsure.

Of the various surveys implemented as part of the ACSSM, the experimental VALI survey is the most novel. The VALI experiment is co-funded by the Australian Bureau of Statistics (ABS), driven by their curiosity to see how the results from VALI compare with those obtained from other survey modes, in particular CATI. If VALI is to evolve into a mainstream data collection mode, on the back of the COVID pandemic-inspired upsurge in interest, it is right to include the survey estimates generated from VALI in this comparative study, and we have chosen to do so. However, given the already large scope of this paper, we decided that this is not the place to document all the design decisions, and all the development and testing and lessons learnt from undertaking this novel VALI survey. To try and fit such a discussion into an already large report would not do it justice. For this reason, while the VALI comparisons are included and brief methodological details provided, the full documentation of the VALI experiment, and the subsequent evaluative analysis will be provided in a separate paper.

4.2.2 Field methods

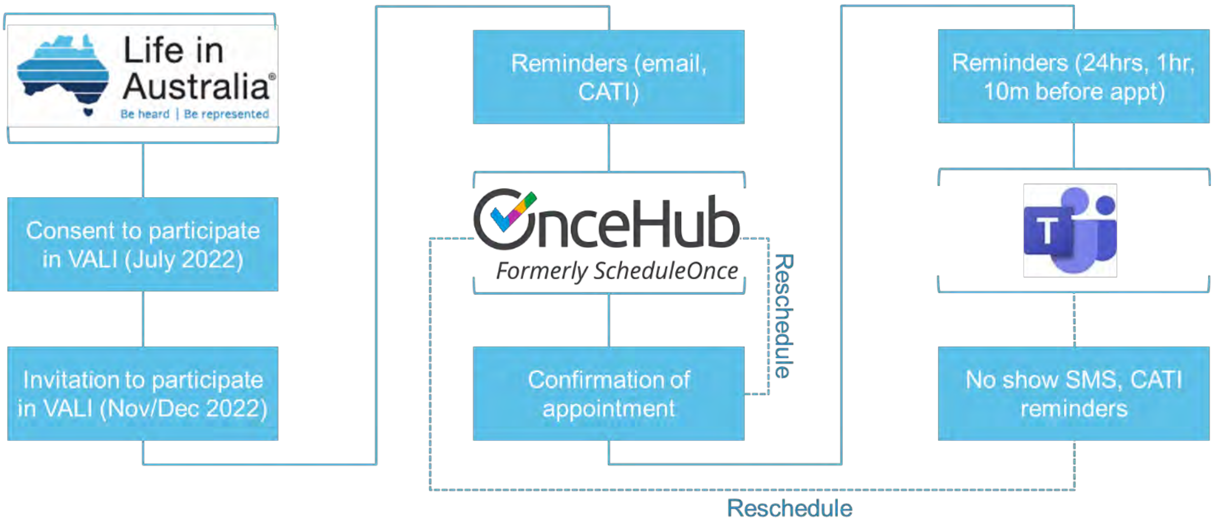
VALI survey

A two-stage approach was used to recruit Life in Australia™ panellists to the VALI survey. Following the initial screening exercise in July 2022 in November/December a representative probability sample of consenting panellists were invited to participate in the VALI survey. The lag between the seeking of consent and the follow up survey invitation is explained by a delay in fielding the overall ACSSM study, with fieldwork dates pushed back from October to November/December for logistical reasons. The VALI workflow is shown in Figure 1 (next page).

Schober et al. (2020) and Hanson (2021) informed our VALI design considerations. Skirmish interviews were also conducted, initially within the Social Research Centre, then within the ABS and, finally, with friends and family. These interviews, in conjunction with the previous research, informed the final VALI set-up.

The sample for VALI was released in replicates so that the specially trained six-person interviewing team could maintain a reasonable workflow of appointments/interviews. A total of 1,399 invitations were issued. Response rates are reported in Section 4.4.

Figure 1 VALI workflow



Life in Australia
Be heard | Be represented
Consent to participate in VALI (July 2022)
Invitation to participate in VALI (Nov/Dec 2022)
Reminders (email, CATI)
OnceHub
Formerly ScheduleOnce
Confirmation of appointment
Reschedule
Reminders (24hrs, 1hr, 10m before appt)
No show SMS, CATI reminders
Reschedule

Invitations and reminders were sent via email, with use of CATI as a final reminder for one replicate. To support respondents to whom video interviewing was likely to be a new concept, a microsite was created on the Social Research Centre website to explain what was being asked of panellists. The site included an explanatory video.

The invitation to panellists included a request to make an appointment for an available interviewing timeslot via the scheduling portal (OnceHub). Appointment-setting was necessary for cost control purpose to reduce the amount of idle time for interviewers. Reminders were sent from OnceHub 24 hours, 1 hour, and 10 minutes prior to the appointment time and reminders were also sent when appointments were not kept. The portal proved to be intuitive and easy to use. It offered a dashboard, could launch SMS reminders, offered integration with Outlook, API access and customisation of look-and-feel (e.g., brand colours, logo), and personalised URLs. Microsoft Teams was used for the video-conferencing platform.

Standard Life in Australia™ $10 incentives were provided to VALI panellists who completed a questionnaire. As is normal practice, respondents had the option of receiving the incentive themselves (either as a Coles e-gift voucher or via PayPal credit) or donating it to charity from a selected list of charities which is periodically changed by the Social Research Centre.

Life in Australia™ survey

Panellists were invited to complete the ACSSM questionnaire following usual Life in Australia™ protocols. For the online mode of data collection this involved sending email invitations and reminders, followed by a reminder call, with these activities spread over a two-week period. For the CATI mode of data collection, an interviewer briefing session was held and practice interviews undertaken ahead of commencing outbound telephone calls with data collection occurring over a two-week period. A total of 582 questionnaires were completed with 554 being completed online and 28 by telephone. Response rates are reported in section 4.4. The standard Life in Australia™ $10 incentive for interviews of this length was offered.

Mobile RDD survey

One of the main methodological changes between the OPBS and the ACSSM is the near total demise in the use of DFRDD sampling frames for general community CATI surveys. These have been replaced by single frame mobile RDD (see, e.g., Hughes, 2020).

Dual-frame RDD surveys involve randomly generating lists of both landline and mobile phone numbers into a composite sampling frame and then ensuring that a fixed proportion of interviews are obtained from each sample source. Weighting then corrects for any disproportionality. This approach, introduced into Australia in 2010 (Pennay, 2010), was originally designed to ensure that the mobile-only population (i.e., those with a mobile phone but not a landline) were included in general community telephone surveys. Over time, as mobile phone saturation became near universal and the use of landlines rapidly diminished, the method morphed into becoming a means of ensuring that the landline only population (i.e., persons who only had a landline and did not have a mobile phone) were included in general community telephone surveys. Mobile phones have become so ubiquitous nowadays that for most general community telephone surveys using a mobile RDD sampling frame is regarded as giving sufficiently good coverage of the adult population for most survey research purposes (Hughes, 2020). The Social Research Centre made this transition gradually from 2020.

SMS push to web survey

The SMS push-to-web survey also uses a mobile RDD sampling frame and involved sending SMS pre-notification messages to mobile phone numbers generated via RDD, followed by another SMS acting as an invitation to complete the survey questionnaire online via the short hyperlink provided. For a random subset of non-respondents an additional reminder SMS was sent to boost response.

Non-probability online panel surveys

The selection of the four non-probability panels to participate in this study considered the following factors:

- Cost
- Indicia of quality
    - Answers to ESOMAR 28/37 Questions
    - Industry body membership: Australian Data and Insights Association (ADIA), ESOMAR, The Research Society
    - Accreditation: ADIA Trust Mark, ISO 20252, 26362 and/or 27001,
    - Participation in our previous study, and
    - Methodological information and availability of paradata.
- In addition, the panel needed to respond to the RFQ (some approached did not) and, in one case, the panel asked whether the RFQ was for a comparative study (agreement could not be reached with this panel).

The final selection was holistic. It included three panels that participated in the OPBS and one that did not. The cost of the most expensive panel included in the study was more than double that of the least expensive panel included in the study.

Non-probability panel providers use various methods to recruit and refresh their panels. The ACSSM panels provided general information on the recruitment strategies they use. The information provided was of a high-level description nature typical of boilerplate for proposals or marketing material. The terminology used differed between panels and it was necessary to make some educated guesses as to what was meant. All panels mentioned marketing both online and offline (e.g., billboards, direct mail). Social media in some form was also mentioned by all panels; it was not always clear whether use of social media was in the form of advertisements, posts, or a combination. Uniquely, Panel 1 mentioned using social media influencers. Panels 1 and 4 allowed open enrolment; we were not able to determine whether Panels 2 and 3 also allowed direct sign-up. Panel 1 mentioned working with affiliates, such as loyalty programs, and use of email (the source of lists of email addresses was not mentioned). Panel 2 allowed personal invitations; it was not clear if this referred to member referral programs (which were used by Panel 4). Panel 3 also recruited via telephone.

It is common for non-probability panels to share sample where necessary to meet quotas. In this instance, all panels indicated that they were able to meet the sampling requirements using only their own panellists. This likely reflects the small sample size requested and the use of soft quotas (vs hard quotas with potentially hard-to-fill quota cells).

The non-probability panel providers approached for this study were asked to conduct a 'nationally representative' survey of 600 respondents. No instructions were provided as to how this task should be carried out.

Descriptions of sample selection and quotas used by each panel are provided below:

- Panel 1: non-interlocking quotas (quota variables not provided)[9]
- Panel 2: soft quotas only (quota variables not provided)[10]
- Panel 3: soft quotas on age, gender, and location
- Panel 4: non-interlocking quotas on age, gender, and location.

It was clear from the quotations that hard quotas would attract higher costs than soft quotas.

4.3 Questionnaire

The ACSSM questionnaire was designed to enable comparative analysis of the relative performance of the different survey methods across as many topic areas as possible. Decisions about the inclusion of specific items were initially based on the availability of high-quality benchmarks, their suitability for use in calibration models, their usefulness in enabling post hoc assessments of data quality, overlap with the OPBS and suitability for the VALI mode of data collection. These considerations were balanced with our desire to keep the questionnaire duration to no more than 15 minutes on average (for cost, data quality and response burden reasons) and provide a coherent experience for respondents. The questionnaire was presented to sample members as the 2022 Health and Wellbeing Survey.

A summary of the questionnaire items is included in Table 3 and a copy of the questionnaire is provided as Appendix 1 and the relevant benchmarks in Appendix 2.

Table 3 ACSSM questionnaire Items

Demographics	Gender, age, state, postcode, suburb Education, country of birth, speaks a language other than English at home Number of adults in household, number of children in household, marital status
Society & politics	Main problem facing Australia* Attitudes to euthanasia Political interest, vote preference Cultural tolerance, discrimination
Survey participation	Online survey panel membership
Health & disability	Requires support with everyday activities General health, life satisfaction, Kessler 6 measure of psychological distress Long-term health conditions†
Lifestyle	Smoking, exercise Alcohol consumption, age of first drink‡ Internet and social media use, TV consumption Time management, support networks, generalised trust
Employment & financial	Job status, home ownership Income† Caregiver status** Receipt of government payments

Notes: * Verbatim item – for mode effect and data quality analysis as well as VALI evaluation. † Long response frame – for mode effect analysis and VALI evaluation. ‡ Complex recall for first drink – for mode effect analysis and VALI evaluation. ** Unable to compare to benchmarks due to change in Census 2021 reporting

Comparing the time taken to complete the questionnaire is complicated by the different number of questions included in some survey modes (e.g., additional questions were asked in VALI) and the variable length of the introduction (e.g., longer introductions are needed for the RDD CATI surveys). Two interview lengths are shown below (see Table 4). The first is the total interview length per survey mode and the second is the interview length for the questionnaire modules common to all surveys. The latter provides a better indication of relative interview length. The median time to complete the questionnaire is shown, rather than the mean, as it is more resistant to outliers.

The median interview length ranged from 7.2 to 21.1 minutes for all content and from 7.1 to 16.5 minutes for the common content. The reason the CATI survey took longer to administer than the other modes is likely a result of the interviewer-assisted mode of data collection and the need for interviewers to read out all response options to respondents before they could answer. By comparison, the VALI survey, which was also interviewer-administered, used showcards to display response options. The median time taken by non-probability online panellists to complete the common modules was 7.1 minutes, on average, compared with 9.3 minutes for Life in Australia™ online completers.

While questions were presented in as consistent a manner as possible, there were some minor differences in presentation to accommodate the various data collection modes.

Table 4 Median interview length by survey mode

Mode and survey	Total (minutes)	Common sections (minutes)
VALI – Life in Australia™	21.1	10.9
Online – Life in Australia™*	9.7	9.3
Online – SMS push-to web	11.6	11.2
CATI – RDD	18.3	16.5
Online – Panel 1	6.9	6.6
Online – Panel 2	7.7	7.5
Online – Panel 3	8.0	7.4
Online – Panel 4	7.2	6.9
Online – Panels 1–4 combined	7.4	7.1

* Excludes the 28 Life in Australia™ interviews undertaken by CATI which had a median interview length of 16.5 minutes.

4.4 Final call dispositions and response rates

The response rates for each survey are provided in Table 5 (see next page). AAPOR definitions and response rates have been used wherever possible. The detailed workings, including full call outcomes and disposition codes are provided as Appendix 3.

For all surveys for which a response rate could be calculated, the response rates are less than 10 per cent, as are the completion rates for those online panels for which it was calculable. The completion rates for the two surveys administered to Life in Australia™ panellists varied considerably (42.9% for VALI compared with 73.1% for the standard execution).

Table 5 Completion and Response Rates by survey frame and mode

Survey mode	Completion rate (%)	Response rate (%)
VALI – Life in Australia™	42.9	1.0[a]
Online – Life in Australia™	73.1	5.6[a]
CATI – RDD (high effort mode)	n.a	7.7[b]
Online – SMS push-to web	n.a	4.0[b]
Online – Panel 1	N/A	n.a
Online – Panel 2	9.5	n.a
Online – Panel 3	8.0	n.a
Online – Panel 4	N/A	n.a

Notes: A meaningful response rate for the abandoned low effort CATI survey could not be calculated. [a] Cumulative Response Rate 2 (Callegaro & DiSogra, 2008). [b] AAPOR (2016) RR3. N/A – Not Available. n.a – Not Applicable.

5 Weighting

Sample surveys are subject to many forms of bias, notably coverage and non-response bias. Survey weighting is commonly undertaken to try to reduce these biases. Traditionally, weighting methods rely on known probabilities of selection to calculate design weights with further post-stratification adjustments for age, gender and geographic distributions applied to account for non-response (Särndal et al., 1992). However, these methods rely on assumptions that many statisticians deem no longer defensible, other than when applied to the relatively high response rate surveys carried out by official statistical agencies.

In a probability-based survey context, single digit response rates with non-ignorable self-selection violate assumptions of random selection thereby undermining the theory on which the design-based approach to weighting is founded. In a non-probability-based survey context, such as opt-in online panels, random selection is not attempted when recruiting the panel, resulting in unquantifiable coverage biases and unknowable chances of selection in relation to the general population of interest.

Superpopulation weights, described in more detail in the next section, are derived via a model-based approach that does not rely on the assumption of known probabilities of selection (Valliant et al., 2000).

Superpopulation weighting can be used for low response probability-based surveys and opt-in non-probability-based online panels. By adopting the same model across all the ACSSM surveys, we can make comparisons of the resulting estimates of means and proportions in relation to population benchmarks without having to account for differences in the weighting schemes. Note that optimising the weighting scheme for each survey to arrive at the most robust estimate from each one will be the subject of future research.

5.1 Superpopulation weighting

Superpopulation weighting involves calibrating the sample using superpopulation weights so that it aligns with population distributions for a broad range of socio-demographic characteristics over and above the usual staples of age, gender, and location.

Superpopulation weights (see, e.g., Dorfman & Valliant, 2005) posit a probability model (the ‘superpopulation model’) that characterises relations among variables that pertain to the units of the population. Such a model makes inferences about population characteristics using sample measurements and auxiliary information in the form of high-quality benchmarks. The model covers the unobserved processes behind a non-probability sample. This approach uses as broad an array of variables as possible for which high-quality benchmarks are available. Generalised regression (GREG) calibration is typically used for calculating superpopulation weights. GREG calibration is the approach used by many official statistics offices around the world, including the Australian Bureau of Statistics, and is implemented in the *survey* package (Lumley, 2020) in R (R Core Team, 2022).

For this study the choice of benchmarks used in the super-population model is based on an assessment of the items that are most different from the population benchmarks across both the probability-based and non-probability samples.

As noted by Valliant (2020), it is expedient to identify a superpopulation model that produces good results for many different outcome (dependent) variables and thus adjusts adequately for imbalances between sampled and non-sampled cases. To test this, we have applied the same set of covariates to predict each of the outcome variables (excluding those used to derive the covariates themselves) and then calculated fit statistics for each model. The statistics were McFadden's pseudo-R^2 (McFadden, 1987) and the area under the receiver operating characteristic curve (ROC; refer to Hosmer and Lemeshow, 2000, for applications to logistic regression). A summary showing the minimum, median, mean, and maximum of the fit statistics for each survey is provided in Table 6. According to the guidelines given by Hosmer and Lemeshow (2000), the average area under the curve values are in the 'acceptable' range for model fit, so we can conclude that the chosen set of covariates may be used for weighting and estimation across the available outcome variables.

Table 6 Summary of superpopulation model fit statistics

	McFadden's pseudo-R^2				Area under the ROC curve			
Survey	**Min**	**Median**	**Mean**	**Max**	**Min**	**Median**	**Mean**	**Max**
VALI	0.02	0.13	0.16	0.52	0.59	0.76	0.76	0.95
Life in Australia™	0.01	0.12	0.16	0.59	0.58	0.74	0.75	0.96
CATI	0.02	0.10	0.15	0.53	0.59	0.73	0.74	0.95
SMS push-to-web	0.02	0.11	0.15	0.59	0.59	0.77	0.75	0.96
Panel 1	0.01	0.09	0.12	0.49	0.58	0.71	0.72	0.94
Panel 2	0.01	0.08	0.13	0.46	0.57	0.72	0.73	0.92
Panel 3	0.01	0.09	0.14	0.56	0.58	0.73	0.73	0.94
Panel 4	0.01	0.09	0.13	0.54	0.57	0.71	0.73	0.97

Large differences in weights may lead to large variances in survey estimates, and so limiting these variations by weight trimming can improve the precision of estimates. The use of constraints in GREG aims to reduce the extent of extreme weights at the same time as ensuring the weights still satisfy the benchmark targets. The method applied is incorporated directly in the calibration process by setting the bounds as constraints. The same bounds were applied for all surveys, namely that extreme weights were trimmed to be no more than a factor of 6 from the mean weight for each survey (no less than one-sixth the mean and nor more than six times the mean).

5.2 Treatment of missing values

The superpopulation model weighting approach requires that there are no missing values present for calibration variables used in the model. Like most surveys, however, some respondents did not provide answers to all questions used for weighting.

A statistical model (Templ et al., 2011) was applied to each item with missing values to impute the most likely value for a respondent, conditional upon their other responses. Given the very low prevalence of missing values overall (generally much less than 5 per cent for any item), the imputation process is expected to have a negligible impact on weighted estimates made from the dataset.

Imputed values are not used outside of the weight construction process.

6 Methods

6.1 Variance estimation

Valliant et al. (2000) describe several methods for deriving the variance of estimators from a model-based approach to weighting. Assuming that the sampling fraction is negligible, as is the case for all the ACSSM surveys, linearisation (also known as the Taylor series method) is a good approximation (Valliant et al., 2018; Valliant, 2020). Alternatively, and the approach adopted here, is the use of re-sampling methods. These create a series of random sub-samples of the data, estimate the desired parameters for each sub-sample (that is, proportions, means or totals), and then summarise the variance across these values.

The method implemented in R (Lumley, 2020) is that by Rao & Wu (1988) which uses re-sampling with replacement from strata, defined here by geographic location. The full-sample weight for sampled cases is adjusted to account for the stratum size and the number of times cases are sampled. For each re-sample, the desired estimates are derived using the adjusted weight. Cases that are not included in a given re-sample receive a weight of 0. The estimate itself is derived from the full-sample weights, but the final variance is an average across the different re-samples, of which there were 500.

Weighting efficiency (Kish, 1992) is a commonly used measure of variance introduced into the estimates as a result of using the weights, it is estimated as follows:

$$weff = 100 \times \frac{(\sum_i^n w_i)^2/n}{\sum_i^n {w_i}^2}$$

where n is the number of respondents and w_i is the weight for the ith respondent. Lower weighting efficiency translates into a lower effective sample size, which is the sample size of an equivalent simple random sample that would be used to determine statistical power in hypothesis testing, these are shown in Table 7 (next page). Effective sample size is defined as:

$$n_{eff} = n \times weff$$

The surveys conducted on Life in Australia™ have relatively low weighting efficiency given that the panel was recruited on a probability proportional to size geographic basis and no data collection quotas were imposed. The weighting efficiencies for the two surveys using a mobile RDD sample frame ranges from 71 per cent to 74 per cent. The four non-probability online panels, which imposed various quota controls (see p. 14) had weighting efficiencies ranging from 62.9 per cent to 89.5 per cent.

Variance is also used in the calculation of the root mean square error ($RMSE$) defined as

$$RMSE_k = \sqrt{{B_k}^2 + {SE_k}^2}$$

where SE_k is the standard error of the kth estimate and B_k is the bias of the kth estimate. Calculation of bias is described next.

6.2 Testing for statistical significance

Unless otherwise stated, where results have been tested for statistical significance, this has been done using a bootstrap approach, due to the non-parametric nature of bias measures. The bootstrap approach is one of a number of methods that resample with replacement from the observed data to derive standard errors of estimates (see Davidson et al., 1997, for more details). To assess the significance of results, 5,000 samples of the same size as each survey were drawn, with replacement, and the measure derived for each resample. The resulting distribution for each measure yielded the relative frequency with which measures more extreme than the observed value occurred, and this served as the p-value for significance testing a re-sampling method, founded on the principle of re-sampling from observed data in order to simulate multiple iterations of the observed experiment. To assess significance of results, 5,000 re-samples of the responding population are used. p-values therefore represent the estimated probability of the observed result happening by chance. More details can be found in Davison et al. (1997).

6.3 Bias assessment

To compare the relative accuracy of the various ACSSM surveys, we look at the difference (or bias)[11] between estimates from each survey and the high-quality external benchmarks.

All variables included in the bias assessment were categorised as either demographic (characteristics that describe survey respondents) or substantive (measures of interest in a social research survey context).

Table 8 shows the final list of variables. Variables included in the survey questionnaire but excluded from the bias assessment are documented in Appendix 6.

Table 7 Weighting efficiency and effective sample size

Survey			
VALI	600	40.4	242
Life in Australia™	582	58.9	343
CATI	803	74.0	594
SMS push-to-web	599	71.0	425
Panel 1	850	80.5	684
Panel 2	852	62.9	536
Panel 3	891	70.6	629
Panel 4	853	77.1	657

Table 8 Questions used in bias comparison

Secondary demographics	Substantive outcomes
Age pension (b_agepension)	Moderate or intense physical activity (b_activity)
Country of birth (b_birthplace)	Daily smoker (b_dailysmoke)
Number of children living in the household (b_children)	Have experienced discrimination (b_discrim)
Labour force status (b_lfs)	Consumed alcohol in last 12 months (b_drinkfreq)
Marital status (b_marital)	Most people can be trusted (b_gentrust)
Person’s income (b_income)	General health status (b_health)
	Psychological Distress (b_k6)
	Life satisfaction (b_lifesatisfaction)
	Multiculturalism is good for a society (b_multicult)
	No long-term health condition (b_nohealthcondition)
	Feel rushed or pressed for time (b_rushed)
	Provide unpaid care in last two weeks (b_unpaidcare)
	First preference for the party vote on Saturday 21 May 2022 (b_votemajor)

Note: dataset variable names shown in brackets.

6.4 Overall measure

Average absolute bias (AAB) is a measure of the difference between a sample estimate and the corresponding benchmark for a characteristic or outcome of interest. To the best of our knowledge, this was first used as a metric in comparative studies of probability and non-probability surveys by Vonk et al. (2006) in the Dutch online panel comparison (NOPOVO) project. The closer this measure is to zero, the better the sample aligns with the population on the benchmark characteristics. The average absolute bias is calculated as follows:

$$AAB = \frac{\sum_{k}^{p} B_k}{p}$$

where:

p = number of variables used in the bias assessment and B_k is determined by

$$B_k \; = \frac{\sum_{j}^{c_k} \left| E(x_{jk}) - \hat{x}_{jk} \right|}{c_k}$$

$E(x_{jk})$ denotes the benchmark value of the jth value of the kth variable;

$\hat{x}_{jk}$ denotes the estimate of the jth value of the kth variable; and

c_k = the number of different values (i.e., categories) for the kth variable.

This calculation of bias is known as a modified Duncan Index (Bottoni & Fitzgerald, 2021) and provides a summary measure by combining bias measures across multiple variables.

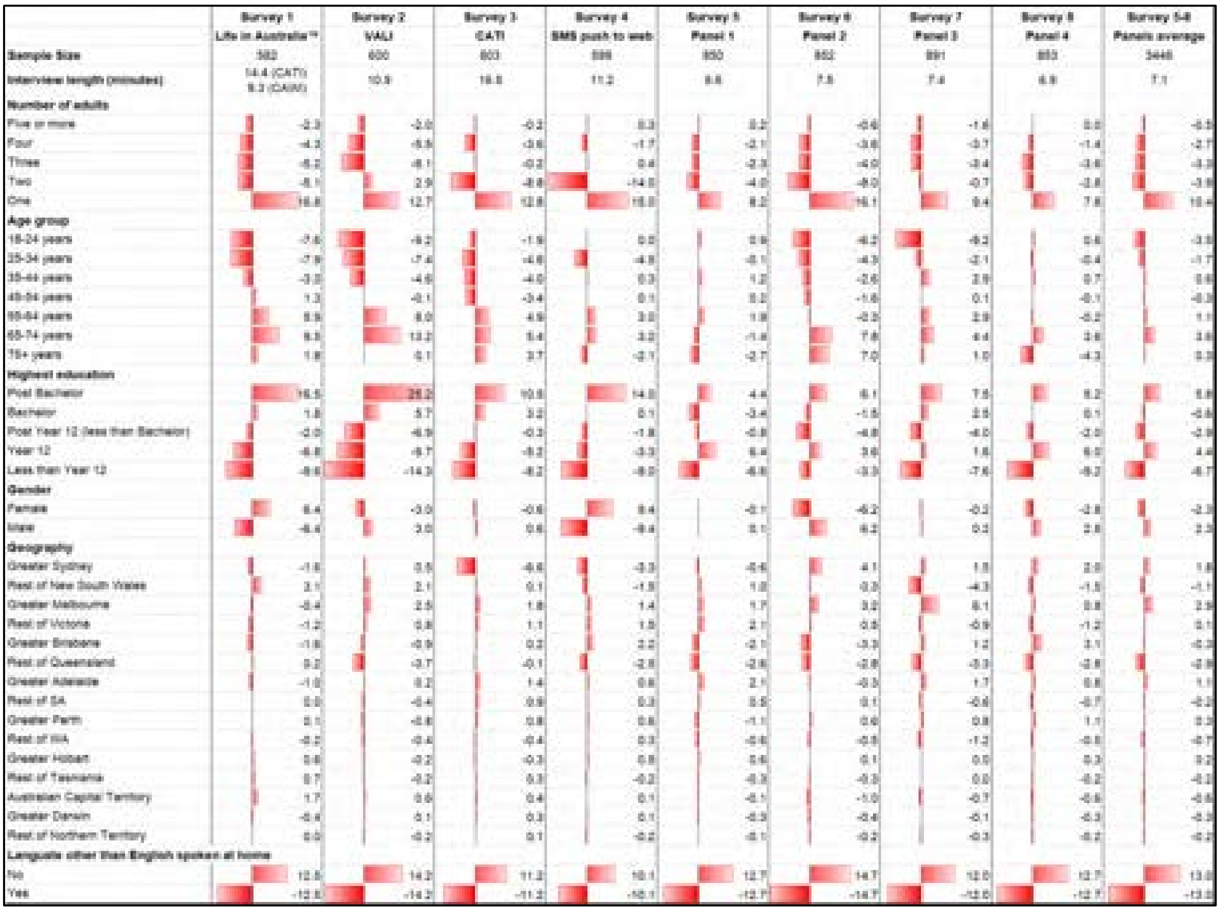
A summary measure for each variable type is calculated by averaging AAB and combining it with variance calculations in a single measure, RMSE, as defined in the previous section.

7 Results

7.1 Unweighted comparisons of bias for weighting variables

The characteristics used as weighting variables for all surveys are: the number of adults in the household, age group, highest level of educational attainment, gender, geography (15 strata formed by the Greater Capital City Statistical Areas), and whether a language other than English is spoken at home. The non-probability online panels used various quota controls (see p. 14). The unweighted bias comparisons for these weighting variables are provided below.

Figure 2 Unweighted comparison of the variables used in weighting (difference from benchmarks, per centage points)



	Survey 1 Life in Australia™	Survey 2 VALI	Survey 3 CATI	Survey 4 SMS push to web	Survey 5 Panel 1	Survey 6 Panel 2	Survey 7 Panel 3	Survey 8 Panel 4	Survey 5-8 Panels average
Sample Size	582	600	803	899	850	852	891	853	3446
Interview length (minutes)	14.4 (CATI) 9.3 (CAWI)	10.9	16.8	11.2	8.8	7.8	7.4	6.9	7.1
Number of adults									
Five or more	-2.3	-2.0	-0.2	0.3	0.2	-0.6	-1.6	0.0	-0.5
Four	-4.3	-6.5	-3.6	-1.7	-2.1	-3.6	-3.7	-1.4	-2.7
Three	-6.2	-8.1	-0.2	0.4	-2.3	-4.0	-3.4	-3.6	-3.3
Two	-5.1	2.9	-8.8	-14.0	-4.0	-8.0	-0.7	-2.8	-3.9
One	16.8	12.7	12.8	15.0	8.2	16.1	9.4	7.8	10.4
Age group									
18-24 years	-7.6	-9.2	-1.9	0.0	0.6	-6.2	-8.2	0.6	-3.5
25-34 years	-7.9	-7.4	-4.6	-4.8	-0.1	-4.3	-2.1	-0.4	-1.7
35-44 years	-3.3	-4.6	-4.0	0.3	1.2	-2.6	2.9	0.7	0.6
45-54 years	1.3	-0.1	-3.4	0.1	0.2	-1.6	0.1	-0.1	-0.3
55-64 years	5.9	8.0	4.9	3.0	1.9	-0.3	2.9	-0.2	1.1
65-74 years	8.5	13.2	5.4	3.2	-1.4	7.8	4.4	3.6	3.6
75+ years	1.8	0.1	3.7	-2.1	-2.7	7.0	1.0	-4.3	0.3
Highest education									
Post Bachelor	16.5	28.2	10.8	14.0	4.4	6.1	7.5	6.2	5.8
Bachelor	1.8	5.7	3.2	0.1	-3.4	-1.5	2.5	0.1	-0.6
Post Year 12 (less than Bachelor)	-2.0	-6.9	-0.3	-1.8	-0.8	-4.8	-4.0	-2.0	-2.9
Year 12	-6.8	-9.7	-5.2	-3.3	6.4	3.6	1.6	6.0	4.4
Less than Year 12	-9.6	-14.3	-8.2	-9.0	-6.6	-3.3	-7.6	-9.2	-6.7
Gender									
Female	6.4	-3.0	-0.6	9.4	-0.1	-6.2	-0.2	-2.8	-2.3
Male	-6.4	3.0	0.6	-9.4	0.1	6.2	0.2	2.8	2.3
Geography									
Greater Sydney	-1.6	0.5	-6.6	-3.3	-0.6	4.1	1.5	2.0	1.8
Rest of New South Wales	2.1	2.1	0.1	-1.5	1.0	0.3	-4.3	-1.5	-1.1
Greater Melbourne	-0.4	2.5	1.8	1.4	1.7	3.2	6.1	0.8	2.9
Rest of Victoria	-1.2	0.8	1.1	1.5	2.1	0.8	-0.9	-1.2	0.1
Greater Brisbane	-1.6	-0.9	0.2	2.2	-2.1	-3.3	1.2	3.1	-0.3
Rest of Queensland	0.2	-3.7	-0.1	-2.8	-2.6	-2.8	-0.3	-2.8	-0.9
Greater Adelaide	-1.0	0.2	1.4	0.6	2.1	-0.5	1.7	0.8	1.1
Rest of SA	0.0	-0.4	0.9	0.3	0.5	0.1	-0.6	-0.7	-0.2
Greater Perth	0.1	-0.6	0.8	0.6	-1.1	0.6	0.8	1.1	0.2
Rest of WA	-0.2	-0.4	-0.4	0.3	-0.6	-0.5	-1.2	-0.5	-0.7
Greater Hobart	0.6	-0.2	-0.3	0.8	0.6	0.1	0.0	0.3	0.2
Rest of Tasmania	0.7	-0.2	0.3	-0.2	-0.3	-0.3	0.0	-0.2	-0.2
Australian Capital Territory	1.7	0.6	0.4	0.1	-0.1	-1.0	-0.7	-0.6	-0.6
Greater Darwin	-0.4	0.1	0.3	0.1	-0.3	-0.4	-0.1	-0.3	-0.3
Rest of Northern Territory	0.0	-0.2	0.1	-0.2	-0.1	-0.2	-0.3	-0.2	-0.2
Language other than English spoken at home									
No	12.5	14.2	11.2	10.1	12.7	14.7	12.0	12.7	13.0
Yes	-12.5	-14.2	-11.2	-10.1	-12.7	-14.7	-12.0	-12.7	-13.0

The non-probability panels perform well relative to benchmarks and relative to the probability-based surveys – none of which imposed quota controls.

The average distance of the non-probability panels from the gender benchmark is 2.3pp. The gender error range for the probability-based surveys is from 0.6pp for CATI to 9.4pp for SMS push-to-web.

Looking at the typically under-represented 18 to 24 year-old age group, the non-probability online panels, on average, under-represent this group relative to benchmarks by 3.5pp. The probability-based surveys under-represent 18 to 24 year-olds as follows; Life in Australia™ (7.6pp), VALI (9.2pp) and CATI (1.9pp). SMS push-to-web did not under-represent 18 to 24 year-olds (0.0pp).

Another common bias in survey research is the over-representation of people with a university qualification and under-representation of those without a Year 12 level of education. The bias in the unadjusted measures of having a university qualification for the probability-based surveys ranges from 10.5pp for CATI to 25.2pp for VALI. The same error range for the non-probability online panels is from 4.4pp for Panel 1 to 7.5pp for Panel 3.

One person households are over-represented in all the surveys, ranging from 16.8pp for Life in Australia™ to less than half that amount of error on Panel 4 (7.8pp).

All the surveys under-represent persons from households were a language other than English is spoken at home to a similar extent ranging from 10.1pp for SMS push-to-web to 14.7pp for Panel 2.

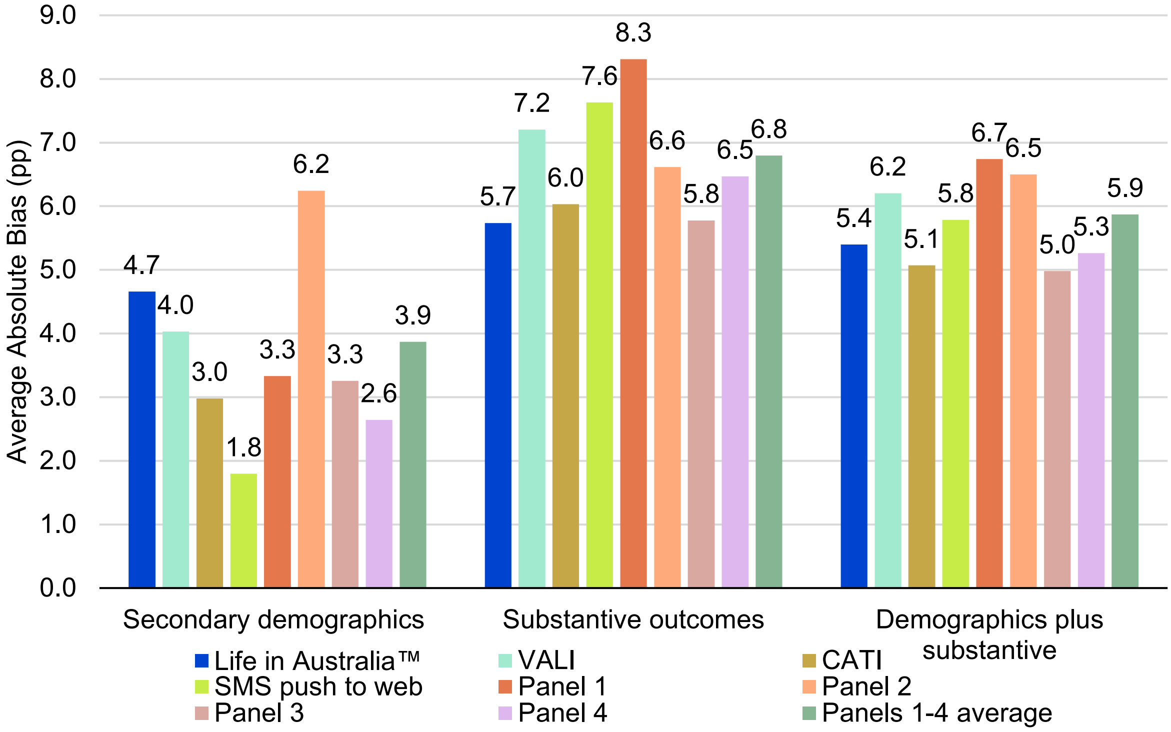
All surveys performed similarly with respect to the geographic dispersion of their samples relative to benchmarks.

7.2 Unweighted comparisons of bias for secondary demographics and substantive variables

The average absolute bias across the six secondary demographic variables for the non-probability online panels (3.9pp) is broadly similar to the error observed in the probability-based surveys: Life in Australia™ (4.7pp), VALI (4.0pp), and CATI (3.0pp), with the exception of SMS push-to-web (1.8pp) (see Figure 3). The error range for the probability-based surveys is 2.9pp and 3.6pp for the non-probability online panels. Although the non-probability online panels show greater variability in terms of the amount of bias occurring in their unadjusted estimates of secondary demographic characteristics, the amount of bias for these variables is quite similar for both probability-based surveys and non-probability online panels. This is consistent with findings reported in previous Australian and international studies (see, for example, Kennedy et al., 2016, 26, Lavrakas et al., 2022, 249, and Yeager et al., 2011, 719).

The third cluster of columns in Figure 3 shows the average absolute bias for all 19 variables. Panel 3 is still the best performed with an average absolute bias of 5.0pp, followed by CATI (5.1pp), Panel 4 (5.3pp), Life in Australia™ (5.4pp), SMS push-to-web (5.8pp), VALI (6.2pp), Panel 2 (6.5pp), and Panel 1 (6.7pp). Again, the variability of the unadjusted estimates produced from the probability-based surveys (1.1pp) is similar to that of the non-probability online panels (1.8pp), and the unadjusted estimates produced by the probability-based surveys are only marginally less biased.

Figure 3 **Average absolute bias by variable category and survey: Unweighted estimates**



7.3 Weighted comparisons of bias for secondary demographics and substantive variables

7.3.1 Secondary demographics

Once the data are weighted, Table 9 and Figure 4 show that the CATI and SMS push-to-web surveys have the lowest average absolute bias across the secondary demographic variables (1.7pp) followed in ascending order by Panel 3 (2.0pp), Life in Australia™ (2.2pp), VALI (2.4pp), Panel 4 (2.6pp), Panel 2 (2.9pp), and Panel 1 (3.6pp). As seen with the unweighted measures, the average absolute bias range is narrower for the probability-based surveys (1.7–2.4pp) than it is for the non-probability online panel surveys (2.0–3.6pp).

For Life in Australia™ and VALI the secondary demographic measure which has the most bias is personal income (3.6pp and 4.2pp, respectively). For CATI, SMS push-to-web and Panels 1,2 and 4 the most biased secondary demographic measure is labour force status. Panel 3’s most biased estimates is marital status (3.5pp). In general, Labour force status is one of the differentiators between probability-based surveys and non-probability online panels. The probability surveys tend to moderately overestimate the number of employed people and underestimate the number not in the labour force. The non-probability online panels overestimate unemployed and not in labour force, some quite significantly. For example (data not shown), Life in Australia™ overestimates the proportion of employed persons relative to benchmarks by 2.4pp (66.3% compared with 63.9%) and underestimates the proportion of persons who are unemployed and looking for work by a similar margin. In contrast, the average across the four non-probability panels is to underestimate the proportion of employed persons by 6.6pp and overestimate the proportion of unemployed persons and persons not in the labour force by 4.2pp and 2.4pp, respectively.

The maximum absolute error recorded for probability-based sample survey is 4.2pp (the VALI estimate of personal income), compared with 7.4pp (the Panel 1 estimate of labour force status) for the non-probability online panels (Panel 1).

Three of the four probability-based surveys (Life in Australia™, CATI, and SMS push-to-web) produce estimates that differ statistically significantly from benchmark values for two of the six secondary demographic items. VALI produces three estimates that differ significantly from benchmarks, whereas Panels 1 and 4 are significantly different from benchmark values for four out of six items and Panels 2 and 3 for five items.

In terms of the AAB across all items, only Panel 1 is significantly different from Life in Australia™, which indicates that the bias in Panel 1 is significantly higher than in Life in Australia™.

Table 9 Bias for secondary demographics (weighted)

Secondary demographics	Life in Australia™	VALI	CATI	SMS push-to-web	Panel 1	Panel 2	Panel 3	Panel 4
Receiving the aged pension	1.8	1.2	0.6	0.6	4.1	4.7	2.7	2.3
Birthplace	1.8	1.4	2.0	1.8	5.1	2.2	0.4	3.8
Number of children in the household	0.6	0.9	0.6	1.4	0.7	1.2	1.2	1.2
Personal income	3.6	4.2	2.9	1.7	2.0	1.2	2.1	2.3
Labour force status	1.6	3.2	3.0	3.2	7.4	5.2	2.2	3.9
Marital status	3.8	3.3	1.0	1.4	2.1	3.0	3.5	2.3
Total	**2.2**	**2.4**	**1.7**	**1.7**	**3.6**†	**2.9**	**2.0**	**2.6**
Ranking	4	5	1	1	8	7	3	6
Number of variables significantly different from benchmark	2	3	2	2	4	5	5	4
Largest average absolute bias	3.8	4.2	3.0	3.2	7.4	5.2	3.5	3.9

Note: Full descriptions of the benchmark variables are provided in Appendix 2. † Fewer than 1% of bootstrap resamples had a bias as different from Life in Australia™ as the observed difference, assuming that the true difference is 0. Refer to Davison & Hinkley (1997), especially Ch 4.

7.3.2 Substantive and overall outcomes

Generally, outside of official statistics, the role of survey research is less about profiling the population in terms of demographic characteristics and more about measuring substantive attitudes and behaviours. On this basis, the most important comparative assessments are how well the respective ACSSM surveys measure the substantive variables of interest once the data have been weighted.

Table 10 and Figure 4 show that Life in Australia™ (5.6pp) and CATI (5.8pp) produce the least biased weighted estimates of the substantive outcome measures, followed by Panel 3 (6.3pp), Panel 2 (6.4pp), Panel 4 (6.6pp), VALI (6.9pp), SMS push-to-web (7.1pp) and Panel 1 (8.1pp). The probability-based surveys (with an error range of 1.5pp) are, again, less variable than the non-probability online panels (1.8pp) and again, on the whole, more accurate.

The most biased weighted estimate of a substantive outcome produced by a probability-based sample survey is the reported level of having experienced discrimination in the last 12 months (SMS push-to-web - 20.5pp), compared with a highest bias for a non-probability sample of 14.5pp for Panel 1's estimate of the same item.

The rank order of the surveys in terms of the average accuracy of their weighted substantive measures shows that Life in Australia™ ranks first, followed by CATI and Panel 3. When looking at the weighted estimates for the demographic and substantive variables combined the rank order for the three least biased surveys remains the same followed by, Panels 2 and 4, SMS push-to-web, VALI, and Panel 1. As previously noted, Panel 3 the most accurate of the non-probability online panels, was the only panel that reported using outbound telephone calls as part of their panel recruiting strategy.

Of the 13 substantive measures estimated by each survey, the number of variables with a bias of less than 5pp (chosen as a heuristic value for reasonable accuracy) for each survey is: Life in Australia™ (7), VALI (7), CATI (8), SMS push-to-web (5), Panel 1 (3), Panel 2 (6), Panel 3 (6), and Panel 4 (4). On this measure, the probability-based surveys fare better than the non-probability online panel surveys.

Of the 13 substantive variables measured by each survey, 10 of the estimates produced from Life in Australia™ contain a statistically significant amount of error. The corresponding figures for the other surveys are CATI (11), SMS push-to-web and Panel 3 (12) and 13 for each of VALI and the rest of the non-probability online panels.

The average absolute bias of the estimates produced by three of the four non-probability online panel surveys (Panels 1, 2, and 4) are significantly higher than that of Life in Australia™. Only Panel 3 is statistically indistinguishable from Life in Australia™.

The other finding to emerge from these comparisons, consistent with previous research, is that having a relatively inaccurate unweighted demographic profile is not a good predictor that the substantive weighted results will be relatively inaccurate. The case in point is Life in Australia™, which ranks seventh in terms of the accuracy of its unweighted demographic profile, but a first in terms of weighted substantive variables. The opposite is true for SMS push-to-web, which ranks first in terms of the accuracy of its unweighted demographic profile but drops to seventh and sixth in terms of substantive measures and overall estimates, respectively. Three of the four non-probability panels (Panels 2, 3, and 4) produced less biased results than the probability-based VALI and SMS push-to-web surveys.

When all accuracy measures are considered, with the exception of Panel 1, the difference in the average amount of bias between the probability-based surveys and the non-probability online panels is relatively small.

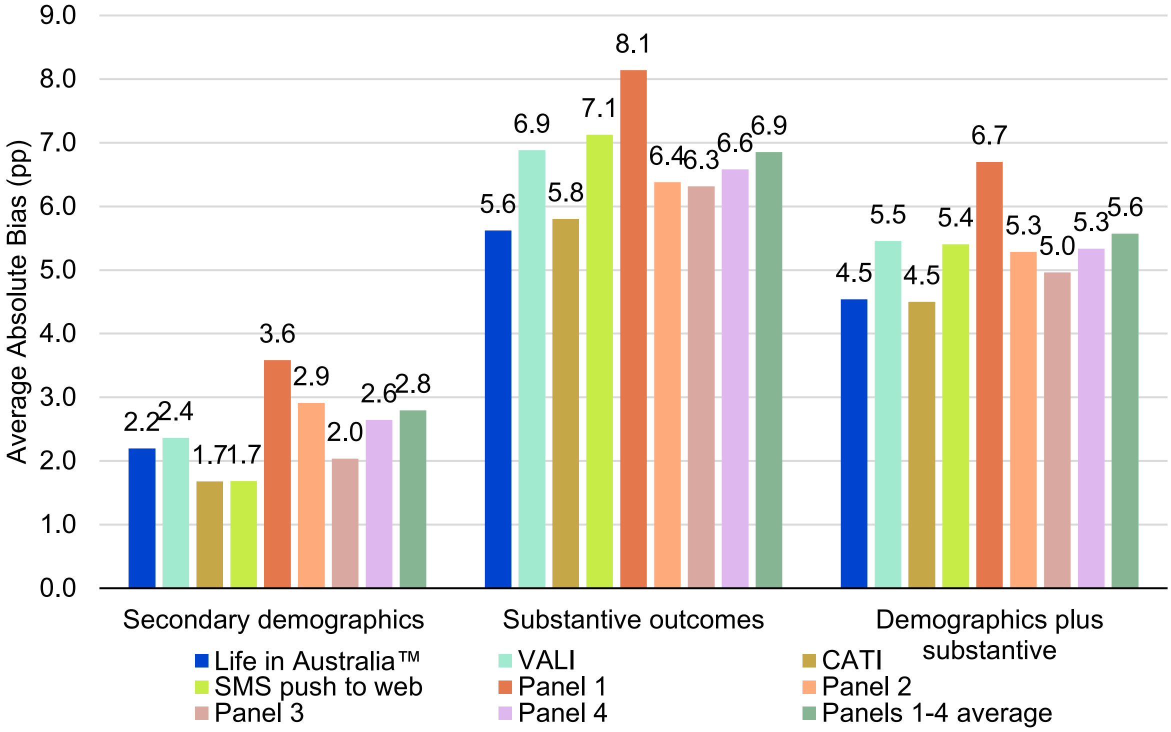
The relatively strong performance of non-probability online panels is not without precedent. The Pew Research Center's 2016 comparative study of U.S. panels and their probability based American Trends Panel, showed that Pew's American Trends Panel 'does not stand out in this study as consistently more accurate than the nonprobability samples' (Kennedy et al., 2016, 5). The authors of the study also concluded that online panels are not monolithic and choice of panel matters (Kennedy et al., 2016, 3).

Table 10 Bias for substantive variables (weighted)

	Life in Australia™	VALI	CATI	SMS push-to-web	Panel 1	Panel 2	Panel 3	Panel 4
Amount of daily physical activity	1.4	4.6	5.5	1.9	2.4	2.8	2.6	1.7
Daily smoking	2.5	4.3	0.8	0.5	11.8	4.9	1.5	6.9
Experienced discrimination in the last 12 months	10.1	10.1	13.0	20.5	14.5	10.6	9.2	10.6
Frequency of drinking alcohol in the last 12 months	2.5	2.6	2.7	2.1	1.5	1.6	1.8	1.9
Generalised trust in most people	4.1	3.7	5.0	7.9	5.6	7.4	4.4	6.7
Self-assessed health status	8.0	4.5	4.3	6.8	7.8	9.0	10.2	7.6
Kessler 6 measure of psychological distress	2.8	3.4	0.2	6.4	11.6	4.5	6.9	7.6
Overall life satisfaction	5.4	7.5	4.1	4.1	5.8	5.6	6.0	5.1
Level of agreement that multiculturalism is good for society	8.6	5.7	5.3	7.7	11.4	13.9	14.1	10.7
Have no long-term health conditions	13.3	18.7	18.3	17.6	12.0	8.2	12.5	11.9
How often rushed or pressed for time	3.6	2.9	2.3	3.8	2.7	4.6	3.6	2.7
Unpaid care provider	3.6	13.2	10.8	8.0	10.1	3.3	3.0	3.9
Vote choice at the previous election	7.2	8.3	3.3	5.4	8.7	6.6	6.5	8.4
Total	**5.6**	**6.9**	**5.8**	**7.1**	**8.1‡**	**6.4†**	**6.3**	**6.6†**
Ranking	1	6	2	7	8	4	3	5
Number of variables significantly different from benchmark	10	13	11	12	13	13	12	13
Largest average absolute bias	13.3	18.7	18.3	20.5	14.5	13.9	14.1	11.9

Note: Full descriptions of the benchmark variables are provided in Appendix 2. ‡ and † indicate respectively that fewer than 1% and 5% of bootstrap resamples had a bias as different from Life in Australia™ as the observed difference, assuming that the true difference is 0.

Figure 4 Average absolute bias by variable category and survey: Weighted estimates



7.4 The impact of weighting on the survey estimates

The Cornesse et al. (2020, 20–21) review of comparative studies found that the application of standard weighting procedures generally resulted in a considerable bias reduction for the probability-based sample survey estimates but did not consistently reduce the bias in the non-probability online panel estimates to the same extent. In some studies (e.g., Lavrakas et al., 2022; MacInnis et al., 2018; Yeager et al., 2011), weighting resulted in an increase in overall bias for some of the non-probability online panel surveys.

The impact of applying the weighting procedures as outlined in Section 5 are now considered.

7.4.1 Secondary demographics

Table 11 shows the variation in the impact of the weights on individual secondary demographic items. Weighting reduces the bias for 5 out of the 6 secondary demographic items for Life in Australia™, 4 out of 6 for VALI, CATI, Panel 2, and Panel 3, 2 out of 6 for SMS push-to-web and 1 out of 6 for Panel 1 and Panel 4.

The impact of weighting on the individual survey estimates for receipt of the aged pension is the most wide-ranging, from a 0.9pp reduction in bias for the VALI survey to an 8.2pp reduction for Panel 2.

The average reduction in bias across these 6 items ranges from a reduction of 3.3pp for Panel 2 to an increase of 0.3pp for Panel 1. Weighting had virtually no impact on the secondary demographic estimates generated by Panel 4 or the SMS push-to-web survey.

7.5 Bias and variance

By combining bias and variance to produce a measure of RMSE, as described in Section 1.1, we can compare the surveys in terms of their total error (i.e., bias and variance). On this basis, the most accurate survey in terms of secondary demographics is CATI (2.3pp), followed by SMS push-to-web (2.5pp), and Panel 3 (2.6pp).

The rank order of the surveys in terms of having the least amount of RMSE error for the substantive measures of interest is Life in Australia™ and CATI (both 6.2pp), Panel 3 (6.6pp), Panel 2 (6.7pp), Panel 4 (6.9pp), VALI (7.4pp), SMS push-to-web (7.6pp), and Panel 1 (8.4pp).

When the secondary demographic variables are combined with the substantive variables, the CATI survey has the lowest total error (5.0pp), followed by Life in Australia™ (5.1pp) then Panel 3 (5.4pp).

Table 11 Percentage point change in bias due to weighting the secondary demographic items

Secondary demographics*	Life in Australia™	VALI	CATI	SMS push-to-web	Panel 1	Panel 2	Panel 3	Panel 4
Receiving the aged pension	-5.7	-0.9	-2.5	0.4	3.7	-8.2	-3.0	1.5
Birthplace	-3.7	-5.8	-3.7	-3.2	-3.7	-4.3	-3.2	-3.3
Number of children in the household	-1.6	-2.4	-1.9	0.6	0.2	-2.1	-0.2	0.4
Personal income	0.9	0.1	0.0	0.1	0.0	0.3	0.2	0.0
Labour force status	-3.1	0.8	1.1	1.6	1.0	-5.8	-1.2	0.7
Marital status	-1.6	-1.8	-0.9	-0.1	0.3	0.2	0.1	0.7
Overall	-2.5	-1.7	-1.3	-0.1	0.3	-3.3	-1.2	-0.0
Number of items with reduced bias	5	4	4	2	1	4	4	1

Note: * Full descriptions of the benchmark variables are provided in Appendix 2.

7.5.1 Substantive outcomes

Table 12 shows the impact of weighting on the amount of bias present in the substantive outcome measures. It is only really for the estimate of ‘no long-term health conditions’ that weighting results in a substantial (4–5pp) reduction in bias for most of the surveys. The generally greater level of bias reduction for this item, and for the item measuring receipt of the aged pension (see above) is likely due to the differential impact attributable to the down-weighting of older panellists to align with their prevalence in the population. Overall, for most of the substantive items, the reduction in bias is less than 1pp. That said, weighting still has a desirable effect on the majority of items for VALI (9 out of 13), Panel 1 (8 out of 13), Life in Australia™, SMS push-to-web, and Panel 2 (7 items) but not so for CATI and Panel 4 (6 items) or Panel 3 (3 items).

The average overall impact of the weights on bias (the bottom panel of Table 12) is uniformly small, meaning that these findings only partially support those of previous comparative studies which generally show that weighting was more effective in reducing bias for probability-based surveys than surveys conducted on non-probability online panels. Average overall bias for the ACSSM surveys across all 19 items varies very little, ranging from a very slight increase in bias for Panel 3 (0.5pp) and Panel 4 (0.1pp) to a 1.2pp decrease in bias for Panel 2. For the probability-based surveys the decrease in bias across all 19 variables ranged from 0.4pp for SMS push to web to 0.9pp for Life in Australia™.

However, in terms of the impact of standard weighting on individual items, we do see a differential impact across the probability-based surveys and the non-probability panels. Bias reduces for only 7 of the 19 variables for Panels 3 and 4 compared to 9 for SMS push-to-web and Panel 1, 10 for CATI, 11 for Panel 2, 12 for Life in Australia™, and 13 for VALI.

So, overall, although the amount of bias reduction attributable to weighting is small, the probability-based surveys tend to gain the most benefit.

Table 12 Percentage point change in bias due to weighting the substantive outcomes items

Substantive outcome	Life in Austr-alia™	VALI	CATI	SMS push-to-web	Panel 1	Panel 2	Panel 3	Panel 4
Amount of physical activity	-0.7	-0.9	0.1	-0.7	0.4	-0.6	0.7	0.4
Daily smoking	-0.8	-1.7	0.1	0.4	-0.6	0.6	0.9	-0.3
Experienced discrimination in the last 12 months	3.0	4.5	3.0	0.0	0.5	2.5	2.4	0.6
Frequency of drinking alcohol in the last 12 months	-0.5	-0.9	0.2	0.2	0.1	-0.1	0.5	0.2
Generalised trust in most people	0.8	-0.9	1.3	0.6	-0.8	0.7	0.4	0.1
Self-assessed health status	1.7	1.7	-0.1	0.9	-0.5	-0.6	0.4	-0.1
Kessler 6 measure of psychological distress	1.6	-1.1	-0.3	-0.3	-0.6	1.6	3.3	0.5
Overall life satisfaction	1.4	0.6	-0.4	-0.9	-0.6	0.2	0.4	-0.5
Level of agreement that multiculturalism is good for society	-0.8	1.9	0.1	0.7	-0.4	-0.2	0.4	-0.1
Have no long-term health conditions	-5.1	-4.2	-4.6	-4.3	-1.4	-5.3	-2.5	0.2
How often rushed or pressed for time	1.1	-0.5	0.0	-0.1	0.5	-2.7	-0.1	-0.2
Unpaid care provider	-1.7	-2.2	-1.1	-2.0	1.3	2.2	0.9	0.7
Vote choice at the previous election	-1.6	-0.5	-1.4	-1.2	-0.2	-1.5	-0.7	-0.2
Overall (+/- pp) substantive items	-0.1	-0.3	-0.2	-0.5	-0.2	-0.2	0.5	0.1
Number of items with reduced bias (out of 13)	7	9	6	7	8	7	3	6
Overall (+/- pp) demographic and substantive items (19 variables)	-0.9	-0.7	-0.6	-0.4	0.0	-1.2	0.0	0.1
Number of items with reduced bias (out of 19)	12	13	10	9	9	11	7	7

Table 13 Root mean squared error by survey (pp)

Weighted comparison RMSE	Life in Austr-alia™	VALI	CATI	SMS push-to-web	Panel 1	Panel 2	Panel 3	Panel 4	Panel aver-age
Secondary demographics	2.9	3.4	2.3	2.5	3.9	3.4	2.6	3.1	3.3
Substantive outcomes	6.2	7.4	6.2	7.6	8.4	6.7	6.6	6.9	7.2
Secondary plus substantive	5.1	6.2	5.0	6.0	7.0	5.7	5.4	5.7	5.9

8 Historical comparisons between the 2015 and 2022 studies

Three comparisons between the OPBS+ and ACSSM are provided in Table 14. All are based on a like-for-like comparison which uses a common approach to calculating bias, measured by the average absolute bias (AAB) for each study, and is limited to the non-weighting variables common to both studies (i.e. only seven variables).[12] We compare the AAB for each variable and overall and the largest AAB generated by each survey.

Following the method used throughout this report, the AAB calculations for the OPBS+ measures have been recalculated so that they reflect the average error for each response category relative to its benchmark value, not just the modal response category (which was the approach used in OPBS+).

These historical comparisons are limited to the comparable methodologies, that is, CATI, Life in Australia™, and the three non-probability online panel providers that provided sample in both 2015 and 2022.

All five of the surveys included in this historical comparison produced more accurate measurements of these survey items in 2022 than 2015. This comes as somewhat of a surprise in the case of CATI, given the steep decline in response rates between 2015 and 2022, but serves as a reminder that response rates are generally a poor predictor of survey accuracy (Kennedy & Hartig, 2019). We are also surprised that the Life in Australia™ estimates are more accurate in 2022 than 2015, given the cumulative effects of panel attrition. Based on the seven measures common to both studies, that is, excluding the Kessler 6 item, Table 14 shows that, on average, bias reduced from 3.9pp to 3.6pp (0.3pp) between 2015 and 2022 for the weighted estimates generated from the Life in Australia™. This compares with a 0.9pp reduction in bias for CATI and, respectively, 0.7pp, 1.7pp, and 1.8pp for the three non-probability online panels (an average bias reduction across the non-probability panels of 1.4pp).

All the surveys, except Panel 1, generated improved estimates for birthplace (Australian born, overseas born from an English-speaking background, overseas born from a non-English speaking background). The estimates of daily smoking rates were less accurate for Life in Australia™ and marginally so for CATI, and less accurate for the panels overall due to a 3pp increase in error for this estimate for Panel 1.

The measures of alcohol consumption improved slightly for each of the survey methods across the years matched by an across-the-board improvement in the accuracy of the personal income measure.

Labour force estimates were more accurate for all the surveys, excepting Life in Australia™, for which bias increased from 1.1 to 1.6pp. The three non-probability panels all produced a more accurate measure of life satisfaction, not so the probability-based surveys.

For four of the five 2015 surveys, the largest absolute error was recorded with respect to the Kessler 6 measure, with errors ranging from 12.4pp for the Life in Australia™ survey to 17.2pp for Panel 1. The exception to this was CATI, with error for the Kessler 6 of 5.4pp. In 2022, the largest errors across the surveys ranged from 4.3pp (the largest error for the CATI survey with respect to self-assessed general health) to 11.8 (the largest error for Panel 1's daily smoker estimate).

Panel 3, the only panel which reported including telephone as an offline recruitment method, is the least biased of the non-probability online panels in both 2015 and 2022.

These findings support an earlier observation that online panels are not monolithic and choice of panel matters (Kennedy et al., 2016, 3). Table 14 shows that when using the AAB for those variables common to the 2015 and 2022 studies (excluding Kessler 6) as our measure, of the non-probability panels, Panel 1 has the highest amount of AAB in both 2015 and 2022, and by a fair margin, (6.6pp in 2015 and 5.9pp in 2022) and Panel 3 the least (5.3pp in 2015 and 3.5pp in 2022).

The problem for those who commission surveys on non-probability online panels is that there is no way of knowing in advance whether they have commissioned a relatively accurate or relatively inaccurate panel with respect to their specific measures of interest. Adding measures to their questionnaire that are related to their items of interest, and for which high-quality benchmarks are available, provides a post hoc means for assessing the likely degree of bias in their measures of interest and, perhaps, some prospect of adjusting their data accordingly.

It is apparent from this analysis that changes in accuracy did not happen uniformly across all variables and, as such, if we were able to undertake a series of comparisons using another set of items, we might get a different result in terms of the specific and overall changes in bias over time. To help illustrate this point, the bottom row of Table 14 shows the amount of error in the ACSSM estimates for those measures not shared with the OPBS. Across these 12 items the average absolute bias is generally higher than it was for the 7 shared items.

This allows for the possibility that had a different set of comparative variables been available to us, we might have seen a different result, that is, non-probability online panel estimates having lower error than probability-based survey estimates. We feel, however, that, if such a result was to eventuate, it would be the exception to the rule. Our rationale for this assertion is based upon the results of the many previous comparative studies that demonstrated the superior accuracy of probability-based sample survey estimates for a wide array of variables. The review by Cornesse et al. (2020) documents the various topics covered by previous studies (see p. 1) and the findings from the large replication study undertaken by MacInnis et al. (2018) give us confidence that cautious generalisations can be made from our findings. MacInnis et al. (2018) replicated and extended Yeager et al. (2011), increasing the number of variables included in the probability/non-probability comparisons from 18 to 38 and covering non-demographic issues such as ‘characteristics of housing structures, consumption behavior, economic expenditures, health quality, health-related behaviors, and health care utilization’ (MacInnis et al., 2018, 712).

They found that despite the deterioration in response rates for probability-based surveys during the intervening years, ‘the probability samples interviewed by telephone or the internet were (still) the most accurate. Internet surveys of a probability sample combined with an opt-in sample were less accurate; least accurate (still) were internet surveys of opt-in panel samples. These results were not altered by implementing poststratification using demographics’ (MacInnis et al., 2018, 707).

Table 14 Comparisons between comparable OPBS+ and ACSSM: average absolute bias and largest absolute error

Outcome	Life in Australia™		CATI		Panel 1		Panel 2		Panel 3		Three Panel Average	
	OPBS+ 2015	ACSSM 2022	OPBS+ 2015	ACSSM 2022	OPBS+ 2015	ACSSM 2022	OPBS+ 2015	ACSSM 2022	OPBS+ 2015	ACSSM 2022	OPBS+ 2015	ACSSM 2022
Birthplace	7.5	1.8	2.3	2.0	2.3	5.1	2.3	2.2	2.3	0.4	2.3	2.6
Daily smoker	1.4	2.5	0.6	0.8	8.8	11.8	5.6	4.9	1.6	1.5	5.4	6.1
Frequency of drinking alcohol	2.7	2.5	5.0	2.7	1.6	1.5	4.4	1.6	3.3	1.8	3.1	1.6
General health	6.3	7.9	3.8	4.3	6.4	7.8	7.5	8.9	7.0	10.2	7.0	9.0
Personal income	5.0	3.6	5.4	2.9	5.8	2.0	6.1	1.2	5.8	2.1	5.9	1.8
Kessler 6	12.4	2.8	5.4	0.2	17.2	11.6	15.7	4.5	16.6	6.9	16.5	7.7
Labour force status	1.1	1.6	7.9	3.0	12.5	7.4	8.3	5.2	10.7	2.2	10.5	4.9
Life satisfaction	3.4	5.4	0.9	4.1	8.5	5.8	7.1	5.6	6.6	6.0	7.4	5.8
AAB (pp)	5.0	3.5	3.9	2.5	7.9	6.6	7.1	4.3	6.7	3.9	7.3	4.9
AAB excluding Kessler 6 (pp)	3.9	3.6	3.7	2.8	6.6	5.9	5.9	4.2	5.3	3.5	5.9	4.5
Largest AAB, excl. K6 (pp)#	7.5	7.9	7.9	4.3	12.5	11.8	8.3	8.9	10.7	10.2	10.5	9.0
AAB for the non-shared items		4.0		3.9		5.2		4.3		4.2		4.6

Note: # Estimates of AAB for each survey have been produced both with and without the Kessler 6 measure. For reasons we have been unable to explain the K6 estimates were very inaccurate in 2015. So as not to overstate the change over time we have also produced an AAB measure which excludes the K6 item.

Table 15 shows selected comparisons between some of the survey measures over time, that is, the differences in bias between, for example, the CATI survey and the most accurate non-probability panel in 2015 compared to the same gap in 2022. This gives an indication of the changing relativities between the surveys.

In 2015, Life in Australia™ had 1.4pp less bias than the most accurate non-probability panel. In 2022, Life in Australia™ had only 0.2pp less bias than the most accurate non-probability online panel. The gap between Life in Australia™ and the three-panel average was 2.0pp in 2015, down to 0.9pp in 2022.

Comparisons between CATI and the non-probability online panels reveal a similar narrowing of the gap. In 2015, CATI had 1.6pp less bias than the best performed non-probability panel survey but by 2022 CATI was only had 0.6pp less bias than the most accurate non-probability online panel and 1.7pp less bias than the three-panel average.

Within the limitations of this comparative analysis, we see an across the board decrease in the performance gap enjoyed by the probability-based surveys over the non-probability online panel surveys.

Table 15 Average bias: selected comparisons: OPBS+ and ACSSM

AAB gap between …	OPBS+ 2015	ACSSM 2022
Life in Australia™ and the least biased non-probability panel	-1.4	-0.2
Life in Australia™ and the three-panel average	-2.0	-0.9
CATI and the least biased non-probability panel	-1.6	-0.6
CATI and the three-panel average	-2.2	-1.7

9 Survey costs and survey quality

To decide whether a particular survey solution is going to meet their needs, the person or agency funding or undertaking the survey should consider the cost of a particular survey method relative to the survey quality. The ABS (2009) Data Quality Framework (DQF)[13] provides a useful way of framing this assessment. According to the ABS, data quality is comprised of the following seven elements:

- Institutional Environment: The institutional and organisational factors which may have a significant influence on the effectiveness and credibility of the agency producing the statistics. (We exclude the Institutional Environment from our review because it is not related to survey methods or sampling frames.)
- Relevance: How well the statistical product or release meets the needs of users in terms of the concept(s) measured, and the population(s) represented. (This is also excluded from consideration because the concepts measured are not related to the choice of survey methods or sampling frames. Coverage is addressed in section 10.1.)
- Timeliness: The delay between the reference period (to which the data pertain) and the date at which the data become available; and the delay between the advertised date and the date at which the data become available (i.e., the actual release date).
- Accuracy: The degree to which the data correctly describe the phenomenon they were designed to measure. This is an important component of quality as it relates to how well the data portray reality, which has clear implications for how useful and meaningful the data will be for interpretation or further analysis.
- Coherence: The internal consistency of a statistical collection, product, or release, as well as its comparability with other sources of information, within a broad analytical framework and over time.
- Interpretability: The availability of information to help provide insight into the data. Information available which could assist interpretation may include the variables used, the availability of metadata, including concepts, classifications, and measures of accuracy.
- Accessibility: the ease of access to data by users, including the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which information can be accessed. The **cost** of the information may also represent an aspect of accessibility for some users (ABS, 2009). For our purposes, the relevant dimension of accessibility is cost; we provide a comparative assessment of survey costs (presented as cost ratios) under this heading.

9.1 Accessibility (cost) and survey accuracy

The components that make-up the variable data collection cost for each of the OPBS+ and ACSSM survey are provided in Appendix 5. Actual dollar values are used to calculate a variable cost per unit for each survey (see Olson et al., 2021, 925 and 931-932). Actual dollar values are not provided in this paper to preserve proprietary information.

The fifth and sixth columns of Table 16 show the differences between the relative unadjusted and quality adjusted variable cost ratios for each ACSSM survey and the survey with the least total error (CATI, RMSE = 5.0) calculated as follows:

$$UaCR_i = \left[\frac{C_i}{n_i}\right] / \left[\frac{C_B}{n_B}\right]$$

$$QaCR_i = \left[\frac{C_i}{n_{eff,i}}\right] / \left[\frac{C_B}{n_{eff,B}}\right]$$

where

$UaCR_i$: Unadjusted cost per interview ratio

$QaCR_i$: Quality adjusted cost per interview ratio

C_i: Survey cost for survey i

n_i: Achieved sample size (n) for survey i

$n_{eff,i}$: Effective sample size for survey i

C_B: Survey cost for survey with the least RMSE (i.e., CATI)

n_B: Number of interviews completed for survey with the least RMSE

$n_{eff,B}$: Effective base for survey with the least RMSE.

The unadjusted variable cost per interview ratio for each survey (Table 16, column 5) is calculated by dividing the actual variable survey costs (A$) by the achieved sample size (n) to get variable cost per interview and showing this as a ratio of the CATI survey's variable cost per interview. To establish a link between survey costs and survey error (see Olson et al., 2021, 929) a quality-adjusted cost ratio is also provided (column 6). This is calculated in the same way as the unadjusted cost ratio, but the effective sample size (n_{eff}) replaces the achieved sample size as the denominator used to calculate the variable costs per unit.

In 2022, the variable cost per unit for the Life in Australia™ survey was about one quarter that of CATI (0.26) when using the unadjusted cost ratios and 0.32 times the cost of CATI when using quality-adjusted cost ratios. These same metrics are 1.00 and 1.84 for VALI, 0.35 and 0.36 for SMS push-to-web and 0.09 and 0.10 for the non-probability online panel surveys.

The three right-hand columns of the table show the sample size, AAB, and unadjusted cost ratios, relative to CATI, for 2015 OPBS+. The effective sample size and the RMSE could not be calculated given the weighting methods used in 2015 and, as such, nor could a quality adjusted cost ratio. Nonetheless, based on the comparative data we have at hand, we see that the unadjusted variable cost per unit for Life in Australia™ reduced from 0.40 times the cost of CATI in 2015 to 0.26 times the cost of CATI in 2022. This speaks to both efforts to reduce the cost of Life in Australia™ as well as the increasing cost of CATI. With respect to reducing the cost of Life in Australia™, there is a lower proportion of Life in Australia™ interviews completed by phone in 2022 relative to 2015, 4.8 per cent vs 7.3 per cent, as well as the use of SMS push-to-web for recruitment, which is considerably less expensive than other modes (Phillips et al., 2022). In terms of the increasing cost of CATI, the difficulty, and hence, cost, associated with conducting CATI surveys increased dramatically between 2015 and 2022. One indicator of this is the number of telephone records called per interview obtained. For the OPBS+ DFRDD survey, this ratio is 6.8 telephone numbers per interview, for the ACSSM mobile RDD, the equivalent ratio for the high-effort CATI survey is 16.3 records per interview.

Based on unadjusted cost ratios and considering that both RDD CATI and Life in Australia™ have an almost identical RMSE, it is evident that the value-for-money proposition for Life in Australia™ over CATI is stronger in 2022 than in 2015.

In terms of the difference between Life in Australia™ and the non-probability online panels, in 2015 the unadjusted variable per unit cost for Life in Australia™ was 3.1 times higher than the average cost of the non-probability online panels. The cost difference is virtually unchanged in 2022, with the Life in Australia™ survey being 2.9 times the average unadjusted variable per unit cost of the non-probability online panels.

In 2015, Life in Australia™ had, on average, 1.7pp less bias than the non-probability on-line panels. In 2022, the gap in error in favour of Life in Australia™ over the non-probability online panels had reduced to 1.1pp.[14]

To sum up, in 2022, Life in Australia™, at 0.26 times the unadjusted cost of CATI and with the same amount of error (4.5pp), is clearly the best value-for-money of the probability-based surveys covered in this study. The cost of Life in Australia™ relative to non-probability online panel surveys remains largely unchanged. The question those who are considering undertaking online panel surveys should be considering is, whether, given the current cost versus accuracy relativities, the higher direct cost for Life in Australia™ over non-probability online panels is worth the, on average, 1.1pp reduction in bias. A consideration of the other elements of the DQF may help resolve this issue.

Table 16 Direct costs and quality adjusted costs by ACSSM survey component

	ACSSM (2022)						OPBS+ (2015)		
Survey	n	n_{eff}	AAB	RMSE	$UaCR$	$QaCR$	n	AAB	$UaCR$
VALI	600	242	5.5	6.2	1.00	1.84	-	-	-
Life in Australia™	582	343	4.5	5.1	0.26	0.32	2,580	4.8	0.40
CATI	498	594	4.5	5.0	1.00	1.00	553*	5.1	1.00
SMS push-to-web	596	425	5.4	6.0	0.35	0.36	-	-	-
Panel 1/A	850	684	6.7	7.0	0.08	0.08	601	7.2	0.11
Panel 2/B	852	536	5.3	5.7	0.09	0.10	600	6.5	0.11
Panel 3/D	891	629	5.0	5.4	0.13	0.14	640	6.3	0.13
Panel C	-	-	-	-	-	-	636	6.0	0.13
Panel 4	853	657	5.3	5.7	0.07	0.07	-	-	-
Panel E	-	-	-	-	-	-	601	6.5	0.16
Panel average	-	-	5.6	6.2	0.09	0.10	-	6.5	0.13

Note: The same panel companies provided the samples for panel surveys 1/A, 2/B and 3/D for both studies. *Excludes refusal conversion interviews as these were not undertaken in 2022.

9.2 Timeliness

The non-probability sample surveys used in the OPBS and ACSSM studies required less time in field to complete the required number of questionnaires than the probability-based surveys. For the ACSSM surveys, the fieldwork durations in days are: VALI (28), Life in Australia™ (15), CATI (14), SMS push-to-web (13), Panel 1 (10), Panel 2 (9), Panel 3 (9), and Panel 4 (12).

The gap between the probability-based and non-probability sample surveys might not be as great as first thought. However, the relationship between probability and non-probability samples and timeliness is very dependent on mode of interview.

Due to their reliance on a finite resource—interviewers—CATI and VALI are most subject to decreasing timeliness as sample sizes increase or if there is a need to sample rare or hard-to-reach sub-populations.

Non-probability online panels will generally be faster, although time in field will still be driven to a degree by sample size or sub-populations, particularly if the vendor needs to work with partners to achieve the sample size or sub-population targets for hard quotas.

SMS push-to-web is not constrained by the need for interviewers and has a short field period. In theory, a very large number of records can be released in a short space of time to yield large samples very quickly, although a staged approach that required the release of sample in replicates was adopted for this study.

9.3 Coherence

The fact that survey estimates generated from probability-based surveys are, generally, both closer to benchmarks and less variable than those produced via non-probability online panels, means that they are more 'coherent' (i.e., comparable with other sources).

9.4 Interpretability

While there is no theoretical reason for there to be a distinction between probability-based surveys and non-probability online panels with respect to which their results can be presented in a fashion that is easy to interpret, the practical reality is somewhat different.

The reason for this is that, by and large, non-probability panel companies are not transparent about the methods used to recruit their samples but, instead, couch them as proprietary, and rely on generic descriptions of sampling processes.

Mercer et al. (2017, 219) note that the most common forms of recruitment (for non-probability online panels) are 'directly through a panel website, clicking on banner advertisements, or when corporations grant panel vendors access to members of their customer loyalty programs'.

The AAPOR Taskforce Report on Online Panels (Baker et al., 2010, 719) notes that 'there is no generally accepted best method for building a (non-probability) panel, and many companies protect the proprietary specifics of their methods with the belief that this gives them a competitive advantage'. The same report also notes that 'panel companies rarely disclose the success rates from their recruitment strategies' (Baker et al., 2010, 721).

Cornesse et al. (2020, 25) similarly note that the 'lack of information available from some online panel vendors can unfortunately make it impossible for researchers to comply with their own codes or certification' and the AAPOR task force report *Evaluating Quality in Today's Complex Environment,* notes that 'transparency in all phases of a study is essential if we are to fully assess survey quality' (Baker et al., 2016, 2) and this applies equally regardless of whether probability or non-probability sampling methods are being used.

The material provided from the panels included in this study matches these descriptions, giving only very broad descriptions of their recruitment techniques, without, for instance, detailing the balance of panellists recruited via online and offline means or reporting how each panellist is recruited.

Due to the relative paucity of methodological disclosure typically seen from non-probability panel companies, the resultant survey estimates are less accessible and more difficult to interpret.

9.5 Summary

Those commissioning survey research must decide which survey method is fit for their specific purpose, and they should be transparent in justifying their choice. The decision-making criteria can be broadly collapsed into making trade-offs between cost (accessibility), timeliness, and quality (accuracy) and, ultimately, the weight given to these somewhat competing demands will determine the optimal survey method. What this study has shown is that non-probability online sample surveys are much cheaper, somewhat quicker, and generally less accurate, but sometimes only slightly so, compared to probability-based alternatives. Within the limitations of this comparative analysis (see Section 11), this study also shows that the accuracy gap in favour of probability-based surveys over the non-probability online panel surveys has reduced. However, it is still the case that estimates produced by probability-based surveys are generally less variable that those produced by non-probability online panel surveys. This, along with the increased methodological disclosure generally associated with probability-based surveys, provides survey researchers with grounds to be more confident in the results generated from probability-based surveys than those generated from non-probability online panels.

An important problem persists for those choosing to fund non-probability sample surveys whereby, for any given survey, or any given items within a survey, researchers have a less firm basis from which to attest to the accuracy and generalisability of their results than if the same questionnaire had been administered to a probability-based sample (cf. Lavrakas et al., 2022). Nor will they have the same basis for confidence as to whether they should be using weighted or unweighted data.

10 Discussion

Lavrakas et al. (2022) used the Total Survey Error (TSE) framework (see Groves, 1989; Groves et al., 2009; Groves & Lyberg, 2010) to undertake a comparative assessment of the sources of survey error most likely to afflict probability-based surveys and non-probability online panels. The TSE paradigm provides a useful context to frame our discussion of the results from the current study. This section draws heavily on Lavrakas et al. (2022).

10.1 Coverage and coverage errors

The CATI and SMS push-to-web surveys used randomly generated mobile phone telephone numbers as sampling frames. Due to the availability of the relevant official statistics in Australia, the coverage gaps associated with the use of mobile RDD frames are knowable. The most recent official estimates relating to the use of mobile phones for voice calls in Australia is that 63 per cent of Australian adults are mobile-only for voice communications, 34 per cent have a landline and a mobile-phone and fewer than two per cent of adults rely solely on a landline (ACMA, 2022b).[15] As such, the gap in the coverage of the RDD mobile frame comprises the less than two per cent of adults only contactable via a landline plus the estimated two per cent of adults without a telephone (Phillips et al., 2019) and the one per cent error positive rate of working number look-ups. We consider the resulting 95 per cent coverage rate adequate for most research purposes.

Most of the Life in Australia™ panellists included in the VALI and probability-based online surveys were recruited via either A-BS push-to-web (64%) or RDD CATI (29%). These are the major methods of recruitment that have been used. Others include RDD SMS push-to-web, and RDD Interactive Voice Response (for further details, see Phillips et al, 2022, 2023). The sampling frames used to build Life in Australia™ at various times have covered the landline and mobile phone populations, the mobile phone population only, and all persons able to receive mail at their residential address. Given these overlapping sampling frames, the coverage properties of Life in Australia™ are likely slightly better than the mobile RDD frame and should, in our judgement, have adequate coverage for most research purposes.

The four surveys that were fielded on non-probability online panels used a variety of convenience frames to build their respective panels. The latest official estimates produced by ACMA (2023a, 3)[16] suggest that online panels have the *potential* to cover the Australian adults very well with '95 per cent (of Australian adults) having used a communication or social media website or app for personal purposes in the six months to June 2022'. However, the reality of the convenience-based sampling methods used by panel providers to recruit their panellists, is that only a small unrepresentative and non-random slice of web users will ever be approached to join a panel. This non-coverage is inherent in the design of these panels is undoubtedly very large and differential (non-random) in nature. It is differential because those who are exposed to an invitation to join non-probability panels are different in many non-ignorable ways from those not exposed to such invitations. 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 et al. (2015) identified significantly different responses between members of probability and non-probability 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. In this study we see quite large differences between the estimates generated from probability-based surveys and non-probability online panels with respect to daily smoking prevalence, the prevalence long-term health conditions, self-assessed health status, experiences of discrimination, and attitudes to multiculturalism.

In summary, uncorrected coverage error in the non-probability panels is a probable contributing factor to the level of bias and variance found in such surveys.

10.2 Sampling and sample errors

Increasingly lower (single figure) response rates for probability-based surveys undertaken outside of the official statistical agencies, have raised questions as to whether random sampling from a sampling frame with unbiased coverage of the population of interest is sufficient to calculate a known probability of selection, and therefore design weights, which along with further post-stratification adjustments, make it possible to calculate the level of precision of the sample estimates with a known degree of confidence. As illustrated by the discussion in *Survey Methodology*, vol. 48, no. 2, increasingly, model-based approaches that are not dependent on the strict assumptions of the frequentist design-based methodology are used instead, as we have done in this study.

If these frequentist assumptions no longer apply to probability-based surveys, then it can no longer be claimed that probability-based surveys have inherently superior statistical properties than non-probability sample surveys. This is not to say, however, as we have already seen, that probability-based surveys do not have other desirable features not shared by non-probability sample surveys such as the very important attribute of random selection and, typically, much better coverage of the population.

10.3 Non-response and non-response errors

The degree of the non-response that occurs in probability-based surveys can be readily calculated. Even when a survey is of members of a probability-based panel, such calculations are relatively easy to make and are the product of the response rate that is achieved when building the panel, the retention rate within the panel, and the completion rate for the questionnaire for which panel members were sampled (Callegaro & DiSogra, 2008). For probability sample surveys, including those conducted within a probability panel, a number of approaches can be pursued to estimate the extent of non-response bias (Montaquila & Olsen, 2012). This also is a function of the nature of the non-response that occurred when building the panel, the non-response from panel attrition, and the nature of the non-response that occurred within the sample/panel for a particular questionnaire. The four probability-based surveys that were conducted as part of this study encountered a very high level of unit non-response with the CATI and Life in Australia™ surveys both having considerably higher rates of non-response in 2022 than was the case in 2015. The AAPOR response rates for the probability-based surveys in the OPBS+ study are DFRDD CATI (17.9% RR3) and Life in Australia™ (12.1% Cumulative Response Rate 2) as compared to 7.7 per cent and 5.6 per cent, respectively, in the ACSSM. The response rates for the other two ACSSM probability-based surveys are lower still (SMS push-to-web, 4.0 per cent; VALI, 1.0 per cent).

For the non-probability online panel surveys, it is 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 per cent of all persons who were exposed to invitations to join a non-probability panel end up joining (Tourangeau, Conrad & Couper, 2013, 42). Although a within-panel completion rate can sometimes be calculated for non-probability 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. As such, with opt-in non-probability panel surveys, there is no well-accepted scientific approach to account for the amount or nature of the non-response biases that may have occurred for a given survey.

On this basis, in addition to the large amount of non-coverage associated with non-probability online panel surveys, they also have an appreciably (non-ignorably) higher level of non-response than do probability sample surveys, even when allowing for the very large decline in response rates for probability-based surveys. The much greater amount of non-response for non-probability panel surveys, compared to surveys using probability samples, occurs at the stages when the panels are built, during the lifetime of the panel (i.e., panel attrition), and each time that panellists are invited to complete a questionnaire. As such, differential non-response bias is likely less of a contributing factor to bias in probability-based surveys than is the case for non-probability online panel surveys.

10.4 Weighting and adjustment errors

Identical weighting schemes were applied to each of the ACSSM surveys, so the use of different weighting schemes is not a contributing factor to the differences in the amount of bias observed. The fact that the non-probability sample surveys applied quotas (with varying degrees of enforcement) to control the distribution of their samples, should, however, mean that the bias reduction for the variables used in quotas should be less for the non-probability panel surveys than it is for the probability-based surveys. Evidence for this is provided in Figure 2 on page 26.

In line with the effects noted in most previous research in this field, weighting, on average, reduced the bias, albeit only marginally, for the substantive variables measured by probability-based surveys, whereas weighting had a negligible effect on the accuracy of the estimates produced by the non-probability online panel surveys and, in the case of Panel 3, increased the amount of bias for the substantive outcome variables by 0.5pp.

10.5 Measurement errors

Apart from some very slight adjustments to accommodate the different modes of data collection, the questions used in the eight ACSSM surveys were almost identical. As a result, there is little reason to expect any differential questionnaire-related measurement error across the eight surveys.

There is, however, the prospect of differential respondent, interviewer, and mode-related measurement errors across the surveys.

As was the case with this study, it is common for 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. Despite this, the two interviewer-administered surveys are likely affected by a combination of interviewer and respondent errors.

The interviewer-administered modes are known to be more likely to generate social desirability bias, especially when sensitive questions are asked, than are self-administered modes (Kreuter et al., 2008).

Respondents and interviewers also may contribute to measurement error in the form of recency effects, where response alternatives that are heard most recently by the respondent are more likely to be chosen than those heard earlier (Holbrook, 2008). However, in self-administered data collection, primacy effects are more of a problem, where 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 at the end of the list of choices (Scanlan, 2008).

The fact that Life in Australia™ uses a mixed mode of data collection – overwhelmingly CAWI but with a small component of CATI to enable the participation of offline panellists, almost certainly leads to a small amount of differential measurement error, not present in the other surveys, due to combining data from the different data collection modes. This is a disadvantage of mixed-mode data collection.

Another measurement error that may affect panel surveys, but not one-off surveys, is panel conditioning. The concern is that repeated interviewing over time may change panellists' attitudes and the way in which they respond to survey questions in a way that is detrimental to data quality. Research has found both harmful and beneficial data quality effects arising from panel conditioning (Amaya et al., 2021; Clinton, 2000; Pennay et al., 2023) and it is difficult to know whether probability-based online panels or non-probability online panels would be differentially affected. On the one hand, the higher retention rates achieved by probability-based online panels would result in a higher proportion of panellists being long-term panellists, increasing the potential impact of panel conditioning. On the other hand, members of non-probability online panels are generally interviewed more frequently than members of probability-based online panels, and are often members of multiple panels, increasing the potential for panel conditioning.

One final emerging measurement error and one that is particular concern to non-probability online panels given their opt in nature, is the threat posed by fraudulent survey data generated by survey bots. A recent U.S. study undertaken by the Pew Research Center found that the various measures they put in place to detect bogus responding from survey bots classified between 3 and 7 per cent of responses across the various opt in online panels as bogus compared with 1 per cent of responses for a survey conducted on an address-based sample (Kennedy et al., 2020).

Finally, previous research shows that members of general population non-probability online panels, as a group, are more likely to generate certain respondent-related measurement errors than are respondents to probability-based surveys (see, Baker et al., 2014; Greszki, Meyer & Schoen, 2014; Hillygus, Jackson & Young, 2014). To try and combat this, our non-probability panel providers exercised what have become standard practices for them and took steps to exclude 'poor quality' responses from the final data. These steps include removing 'straight-liners,' removing 'junk'/poor quality responses to open ended questions, and removing speeders (as variously defined by the panel providers). The effectiveness of these steps in improving overall data quality is not known.

10.6 The special case of the VALI survey

A separate evaluation report of the experimental VALI survey is to be prepared, so just a few summary comments are provided below.

The two-stage recruitment process used for the VALI survey, which involved seeking consent to being interviewed via video-conferencing prior to issuing a survey invitation, resulted in a very pronounced self-selection bias towards panellists with university (i.e., bachelor's degree and above) qualifications (see Figure 2). While post-stratification to educational attainment benchmarks re-aligned the VALI estimates on this characteristic to those of the population, this came at the cost of introducing more variance into the VALI estimates. This is reflected in the relatively high RMSE for VALI of 3.4pp, the highest of the probability-based survey methods and higher or on a par with all but one of the non-probability online panels.

As an interviewer-administered mode of data collection, there is scope for interviewer-related measurement error being present in the VALI data, as is the case for the CATI survey and for a small proportion of Life in Australia™ interviews. For VALI, the potential for interviewer and respondent-related measurement error may be greater than the other interviewer-administered modes in this study, given that, for VALI, the interviewer and respondent are visible to each other and because both parties were unfamiliar with the format.

Despite initially thinking that the Life in Australia™ panel would prove to be a good platform for VALI, given the relationship that exists between the panellists and the Social Research Centre, ultimately this turned out not to be the case. Although the findings from a round of interviewer and respondent de-briefing interviews showed that VALI generally works very well and is well-received, panellists' post-survey preference was still for CAWI.[17] Respondents see little added value in VALI and identify an increased respondent burden due to the need to set and keep appointments and to be 'seen' by the interviewer. It was felt that VALI interviews warrant a higher incentive payment to respondents.

This experiment showed VALI is a viable alternative data collection mode. Its best use is probably as an alternative to face-to-face data collection in situations where there is an established relationship with respondents, e.g., subsequent waves of a longitudinal survey program.

10.7 Overall assessment

A summary of what the ACSSM tells us about these comparative/complementary/competing survey methodologies is now provided.

10.7.1 Accuracy

Overall, the CATI and Life in Australia™ surveys produced the most accurate results, followed by Panel 3, SMS push-to-web, VALI, and Panels 1, 2, and 4, with Panel 1 generally showing the largest biases (as it did in 2015).

As previously noted, Panel 3 reportedly used outbound CATI as one of its recruitment methods, but whether this contributed to their superior accuracy is not known. The finding that non-probability online panels sometimes produce results that are more accurate than those produced by probability-based surveys, while not common, is consistent with findings reported by the Pew Research Center (Kennedy et al., 2016).

One dimension of accuracy that does consistently favour probability-based surveys over non-probability online panels is that probability-based surveys routinely produce more consistent (i.e., less variable) results than non-probability online panel surveys.

The historical comparisons presented in this paper are limited to CATI, Life in Australia™ and the three non-probability panel providers used in both studies and to a common set of variables. On average, there was a reduction in AAB for the three survey methods for the measures common to all surveys over time. In 2015, Life in Australia™ produced estimates for these variables that had, on average, 2.0pp less error than the equivalent non-probability online panel estimates, on average. This gap shrank to 0.9pp for the same comparison in 2022. The largest AAB across the surveys over time decreased from 7.9pp to 4.3pp for CATI and from 12.4pp to 7.9pp for Life in Australia™ and from 16.5pp to 9.0pp for the non-probability online panels, on average.

Based on the limited data available to us, we find that the gap between probability-based surveys and non-probability online panels has narrowed since 2015.

10.7.2 Survey costs and survey quality

The possibility that the relative accuracy of probability-based surveys and some surveys conducted on non-probability online panels may be converging, while the cost advantage that non-probability online surveys have over probability-based surveys is either the same (in the case of Life in Australia™) or increasing (in the case of CATI) means that it is harder to justify choosing probability surveys over non-probability online panels in 2022 than it was in 2015. Of the probability-based surveys tested, the probability-based online panel (Life in Australia™) emerges as best value for money for survey researchers placing a premium on generating the most accurate estimates.

Ultimately, those who commission or undertake surveys must decide which survey method is fit for their specific purpose. This study shows that non-probability online sample surveys are much cheaper and somewhat quicker than probability-based sample alternatives and that the accuracy advantage enjoyed by probability-based surveys over non-probability panel surveys may have narrowed.

On balance, bearing in mind all aspects of data quality (Section 9) and survey error (this section), it still does seem to be the case that if one wishes to generalise from a sample to an inferential population, that probability-based surveys, undertaken by a reputable provider committed to a high-level of transparency, allow one to do so with more confidence than do non-probability online panel surveys, on average. It is also true, however, that those commissioning survey research are continuing to find the price premium required to undertake probability-based surveys is too high.

11 Limitations of the study

11.1 Sample size

Due to the self-funded nature of the ACSSM and the desire to cover a range of methods, sample sizes are relatively small. This impacts sampling error for the probability samples. Although sampling error is not applicable to non-probability samples (see, e.g., Baker et al., 2013), similar concerns apply to our ability to generalise to the broader universe of non-probability online panels from the non-probability samples used in the ACSSM.[18]

A further point to be made regarding sample sizes is that the sample sizes for the surveys, while broadly similar in the main, do vary across surveys and across study years (2015 and 2022). We explored whether these differences in sample sizes made any meaningful difference to our comparative measures of bias by generating 1,000 replicate samples for each survey constrained to a Life in Australia™ equivalent sample size (n=582). The results of this analysis indicated very little difference in the mean bias measures based on the original sample size for each survey and the resultant measure from the 1,000 resampled replicates.

The different sample sizes would certainly differentially affect significance testing across survey years with the smaller sample sizes having wider confidence intervals. This is more of an issue for comparisons across survey years (i.e., 2015 cf. 2022), as the sample sizes are fairly similar within studies but not between studies. It is for this reason that no significance testing is undertaken when comparing bias across the 2015 and 2022 surveys. See Section 12 for a discussion of future work.

11.2 Generalisability

The ACSSM and similar comparative studies have a different focus to normal surveys. Estimands from a normal survey are intended to generalise to a specific population (e.g., Australian residents over 18 years) for the constructs measured in the survey; the difference between survey estimates and the true value of each construct measured for the population of interest is survey error. By contrast, estimands from a comparative study like the ACSSM are intended to generalise about the cost and error properties of a population of *surveys* that are, were or might be fielded. This has impacts on how we think of the limitations of the design.

11.2.1 To what sampling frames and modes does the ACSSM generalise to?

The ACSSM does not speak to all types of surveys. Methods not covered in the ACSSM that are in use in Australia include address-based sampling with push-to-web, face-to-face surveys (although these are becoming less common; see, e.g., increasing use of mixed-mode by the ABS) and IVR telephone surveys. CATI surveys of landline sample are likely to be very rare due to rapid declines in landline usage; thus, the omission of landline CATI from the ACSSM is unlikely to limit its utility.

Although we did not use the Integrated Public Number Database (IPND) as a sampling frame for CATI surveys, ACSSM results for Mobile RDD CATI are likely to apply to IPND CATI surveys as well, as, based on our experience, there are minimal differences (Phillips et al., 2022).[19] The primary advantage offered by the IPND is the ability to sample local areas.

Findings from the ACSSM CATI surveys cannot be generalised to CATI surveys using listed sample. Listed sample surveys of the general population are further from benchmark values than are RDD surveys but cost less.

Other more novel data collection approaches are also not addressed in the ACSSM. We did not trial the use of chat bots, use of sensors on mobile devices, or SMS surveys (where the mode of interview is the SMS), for instance.

11.2.2 How well does the ACSSM generalise to other implementations of the included methods?

VALI

VALI is an emerging mode of data collection, making it difficult to generalise about other implementations. Broadly speaking, the following points should be borne in mind when evaluating the generalisability of findings from the ACSSM to other implementations of VALI (see Schober et al., 2020 for a useful listing of design considerations):

- Is the sample cross-sectional or longitudinal? Early findings from other research indicates that VALI can struggle with cross-sectional sample and seems to be work better in a longitudinal context, like Life in Australia™, where there is a pre-existing relationship between the survey research organisation and the respondents.
- Is VALI the sole data collection mode or is part of a sequential multi-mode design? Due to the expense of VALI (see previous discussion of cost), it may be reserved for use after less expensive alternatives (e.g., push-to-web) have been exhausted. In the present case, VALI was the sole data collection mode.

Other potential limitations of generalisations from VALI are the fact that the ACSSM was the first time the Social Research Centre had conducted VALI. This lack of prior experience with VALI is, oddly enough, more likely to enhance than detract from generalisability to other contemporary implementations because no survey research organisation globally has extensive experience with VALI, due to it being a very recently developed mode of data collection for surveys. Over the longer term, the degree to which the ACSSM findings can be generalised is likely to be compromised by advances in the field, as organisations gain more experience with VALI and best practices are emerge. See, e.g., the development of norms of data collection from mobile phones (Lavrakas et al., 2010) and as interviewers gain experience in administering questionnaires using VALI.

Life in Australia™

Life in Australia™ is currently Australia's only probability-based online panel, beside developmental work conducted by the ANU Centre for Social Research & Methods (Hahn, 2022). Any future Australian probability-based online panels are likely to differ from Life in Australia™ with respect to some of the methods used for recruiting panellists and the many decisions that must be made about how the panel operates. The design of Life in Australia™ was broadly informed by other probability-based panels, most notably the Pew Research Center's American Trends Panel circa 2015. Elements of this include discrete monthly waves, incentives paid each wave rather than a points-based system, and the use of an alternative data collection mode to accommodate offline panellists.

Looking internationally to other probability-based online panels, Life in Australia™ is unusual in several aspects:

- Use of CATI for interviewing offline panellists. Generally, members of the offline population are either unable to join or are given a device with internet access to enable them to complete questionnaires. It should be noted, however, that the offline fraction of interviews in the ACSSM (4.8%) is small and therefore unlikely to have a large impact on results.
- Use of CATI for reminders. It is extremely rare for panels to use CATI for reminders. This is unlikely, however, to have much of an impact on results.
- Use of a wide variety of sampling frames and invitation modes for recruitment (RDD CATI, A-BS push-to-web and CATI, RDD IVR, RDD SMS push-to-web). Most panellists in the ACSSM (93%) were, however, recruited via either RDD CATI

(29%) or A-BS push-to-web (64%). Life in Australia™ mirrors U.S. panels' similar evolution from RDD CATI to A-BS push-to-web for recruiting panellists. The number of surveys completed by panellists recruited via IVR and SMS push-to-web is low and unlikely to harm the ability to generalise to other probability panels recruited via RDD CATI and A-BS push-to-web approaches.

Readers will need to draw their own conclusions about the generalisability of the results based on the degree to which the manner of operation of Life in Australia™ differs to other panels of interest.

CATI

The performance of CATI from a cost and potentially quality perspective is potentially affected by a host of decisions made as to whether a pre-notification SMS is sent, the call cycle (number of calls, intervals between calls, time of day of calls), use of an autodialler and autodialler settings and recruitment, training, retention and supervision of interviewers.[20] However, the similarity in the responses between the high- and low-effort arms (refer back to Section 0) suggests that findings should be generalisable across a reasonable range of these settings. Caution should, however, be exercised at generalising from the ACSSM to cross-sectional studies using a far higher number of call-backs, noting that any such survey would be extremely expensive to conduct; we are not aware of any such surveys being fielded in Australia nowadays.

SMS push-to-web

SMS push-to-web with RDD sample is in limited use in Australia (Hahn, 2022; Kocar, 2022), which makes it difficult to understand the degree to which the ACSSM may be generalisable to other implementations.[21] Due to the limitations inherent to SMS: messages must be short, both due to social expectations and the fact that SMS providers charge based on length.

Non-probability panels

The ACSSM's use of non-probability panels does not replicate all possible approaches used in non-probability panels. This potentially limits the generalisability of results, although care was taken to include multiple non-probability panels to be able to provide some evidence of the degree of variability between panels.

The ACSSM instructed panels to use soft quotas. Clients may require hard quotas, forcing panels to supply completed surveys in proportion to the client's quota scheme. This will increase cost but may reduce bias, although supporting evidence for the efficacy of quotas is limited. A moderate degree of caution is required when generalising the ACSSM's findings to studies using hard quotas.

In many cases, panels will share sample. For large samples, repeated cross-sectional studies with re-contact restrictions, studies focused on low incidence or hard-to-reach populations or studies with hard quotas, panels may need to supplement their own panellists with those from other panels. This was not the case for the ACSSM, where panels were able to fulfil study requirements using only their own panellists. Given the poor performance of non-probability panels in comparative studies, there is little reason to believe that sharing sample will meaningfully reduce total survey error.

The selection criteria used for non-probability panels in the ACSSM (refer back to Section 0, p. 13), with a strong focus on ISO certification, membership in industry bodies and answering ESOMAR questions, means that the panels selected represent the middle to top tier of the market. If there is a bias from this focus, it would tend to overstate the accuracy of the broader population of non-probability panels.

Although individual non-probability panels claim unique features that distinguish them from their competitors, it is not clear to what extent these claims of uniqueness hold up to scrutiny and—to the extent that they do—that they reduce total survey error. We address this point because meaningful quality distinctions between panels would tend to lessen the ACSSM's generalisability; on the other hand, if panels are a fungible commodity, the ACSSM's findings should be more easily generalisable. The material received from non-probability panels in the course of the ACSSM is free of the kind of supporting methodological detail that we usually expect to see in survey research.[22] This is not a new observation. Callegaro et al. (2014, 6) note that 'Companies that created nonprobability panels tend to be secretive about the specifics of their recruiting methods, perhaps believing that their methods provide them a competitive advantage (Baker et al., 2010). For this reason, there are few published sources to rely on when describing recruitment methods.' The international comparative literature casts a harsh light on claims of uniqueness, as—although there is indeed panel-to-panel variation—whatever unique attributes panels have seem to fail to bring them to the same level as probability samples with respect to total survey error. Supporting the contention that non-probability panels are—to a large degree—fungible, is the nature of the market. As indicated by the very low cost of research on non-probability panels, it is highly cost-competitive and unlikely to support product differentiation. Moreover, the exchange of sample between panels indicate that in deeds—if not in words—panels themselves believe their samples are fungible.

One possible exception to the above is YouGov. The panel's Chief Scientist has articulated a principled approach to non-probability sample selection (Rivers, 2007) and the panel was an early user of multi-level regression with poststratification (MRP) in political polling (Bailey & Rivers, 2020). It also has had notable success in calling elections (YouGov, 2022) and in a Pew Research Center comparative study, where it was more accurate across a range of benchmarks than the Pew Research Center's own probability-based online panel (Kennedy et al., 2016; Rivers, 2016). The extent to which YouGov's unique approach is adopted by YouGov in Australia is unclear. There have been notable departures in Australia from its global norms, such as using IVR alongside non-probability sample in election polling (White, 2019 cited in Pennay et al., 2020). YouGov has made limited use of MRP in Australia, with most Australian YouGov polls not using this method (YouGov, 2023), despite the notable success when doing so, of correctly calling in advance the Treasurer's loss of his blue-ribbon Liberal seat (see, e.g., Maiden, 2022). This suggests that, although YouGov may offer superior performance in the U.K. and U.S., the same may not apply in Australia, outside of surveys using MRP. ACSSM findings may therefore generalise to non-MRP YouGov surveys fielded in Australia.

The ACSSM exclusively uses commercial non-probability panels. Different response dynamics are likely for volunteer panels that do not offer incentives, such as the ABC's (2021) Australia Talks survey and the University of Tasmania's (n.d.) Tasmania Project or cross-sectional volunteer samples like smartvote (ANU, n.d.). Results from the ACSSM cannot be generalised to such panels or cross-sectional samples.

11.2.3 How well does the ACSSM generalise internationally?

The findings of previous comparative studies of probability and non-probability samples across Australia, Canada, Europe, and the U.S. have broadly been consistent in indicating the inferiority of non-probability samples and the failure of weighting to remediate bias (Cornesse et al., 2020, Table 1), suggesting that findings from the ACSSM are likely to generalise to at least these societies and, likely, others of similar ilk where, to the best of our knowledge, no comparative studies have been conducted (e.g., Israel, New Zealand).

With that said, some elements of potential difference between Australia and other nations should be borne in mind:

- The legal environment regarding the use of SMS and autodiallers notably differs from the U.S., where these are restricted by the Telephone Consumer Protection Act (47 U.S.C. § 227) (Ballon et al., 2021). Sending SMS messages and the use of an autodialler for mobile sample without prior consent is legal in Australia, without the need to use workarounds (e.g., manually sending SMS). This impacts SMS push-to-web. Although the CATI surveys did send an advance SMS, there is no consistent evidence showing the impact of such an SMS on the characteristics of the achieved sample of respondents (Dal Grande et al., 2016, Pennay, Borg & Lavrakas, 2016).
- Unlike some European countries, Australia does not have population registries that are accessible for use in sampling.[23]
- Due to the lack of a single dominant non-English language in Australia (c.f. Spanish in the U.S.), all modes were fielded in English only.
- In general, use of face-to-face modes of interview is less common in Australia than the U.S. and Europe. This reflects Australia's low population density, which makes face-to-face interviewing outside of capital cities extremely expensive.

11.2.4 To what topics do the findings from the ACCSM generalise to?

A comparative study focused on benchmarks will necessarily be focused on the available benchmarks. The ACSSM is therefore focused on topics primarily found in ABS products. Although we attempted to include a broad range of topic areas, attitudinal questions are relatively under-represented in the questionnaire due to the focus of most ABS surveys on collecting information on behaviours and characteristics of individuals, families, households, and dwellings.

11.3 Comparisons between the ACSSM and OPBS+

The fact that both the OPBS and the ACSSM were designed to evaluate contemporary approaches to survey research is, necessarily, a factor that limits direct comparisons between the two studies.

As previously discussed, the foremost limitation of the historical comparative analysis is that it is limited to only seven directly comparable variables common to both studies. Clearly, this is too few from which to draw firm conclusions as to the general performance of the various survey methods over time. Again, as previously noted, it is possible, although unlikely given the range of variables that have been tested in the various similar comparative studies around the world, that another set of variables would yield different results. The findings of previous surveys summarised by Cornesse et al. (2020) and the large replication study by MacInnis et al. (2018) provide us with confidence that cautious generalisations can be made from our findings with respect to the relative performance of probability-based surveys and non-probability online panel surveys in 2015 and 2022.

Comparisons of the relative performance of the standalone CATI surveys included in the OPBS and ACSSM also require some caution. Both studies used methods contemporary to their time, which means that differences in their conduct need to be borne in mind. These differences include the transition from DFRDD in 2015 to the use of a mobile RDD frame in 2022. Approaches to survey weighting also evolved over this period. That said, while the approaches adopted for the CATI surveys are different, it is nonetheless possible to compare the bias and variance of these two approaches as examples of 'typical' CATI surveys for their time.

Some context is also needed when comparing the relative accuracy of the survey estimates generated by Life in Australia™ over time. The OPBS replication study (undertaken in January 2017) was just the second survey conducted on the then new Life in Australia™ panel. Recruitment was undertaken in November 2016 using a DFRDD sampling frame with a 30:70 landline to mobile phone split resulting in 3,203 panellists. The ACSSM was conducted in December 2022 drawn from a much larger pool of Life in Australia™ panellists (*n*=7,396) with the panel having been replenished using a variety of different methods and the proportion of offline panellist completing via the telephone having about halved.

When comparing the performance of the-non-probability online panels over time it is necessary to consider, but hard to know, weather the panel providers are using the same or different recruitment methods and avenues from which to source panellists. It is also important to note that only three of the four panels used in the OPBS were also used in ACSSM.

12 Concluding remarks and next steps

This study shows that although non-probability online sample surveys are much cheaper and quicker, they are generally less accurate, but sometimes only slightly so, than the probability-based alternatives. There is also evidence to suggest that the accuracy gap in favour of probability-based surveys over the non-probability online panel surveys may have narrowed.

Despite this narrowing of the accuracy gap in favour of probability sample surveys over non-probability online panel surveys, it is still the case that the estimates produced by probability-based surveys are generally less variable that those produced by non-probability online panel surveys. This, along with the greater methodological disclosure usually associated with probability-based surveys, provides survey researchers with grounds to be more confident in the results generated from probability-based surveys than those generated from non-probability online panels.

As noted previously, an important problem persists for those choosing to fund non-probability sample surveys in that, for any given survey, or any given items within a survey, researchers have a less firm basis from which they can confidently assert the accuracy and generalisability of their results than if the same questionnaire had been administered to a probability-based sample. Nor will they have the same degree of confidence as to whether they should be using weighted or unweighted data.

It still does seem to be the case that if one wishes to generalise from a sample to the inferential population, that probability-based surveys allow one to do so more accurately and with much more confidence than do non-probability online panel surveys. Increasingly, however, those commissioning survey research are deciding that such confidence comes at a price they are not prepared to pay, particularly if there is a chance that less expensive approaches may only be slightly less accurate (but could be considerably less accurate).

There are many issues arising from the ACSSM study left to explore. For this reason, and in the interests of transparency, the Technical Report from the study, the data file, and all explanatory documentation will be lodged with the Australian Data Archive. Once lodged, these will be accessible to researchers via an application process and subject to Australian laws governing privacy and confidentiality.

Among the issues left to explore include the following:

- Can survey-specific optimal weighting reduce the bias in the estimates generated from the probability and non-probability samples used in this study without unduly adding to the variance?
- Can blending and calibration improve the estimates generated from the non-probability online panel surveys?
- Is there a discernible difference in the amount of measurement error in responses provided by panellists on probability-based online panels compared to responses provided by panellists on non-probability online panels? The presence of differential measurement error could be indicated by measures of speeding, straight-lining, satisficing, use of non-substantive response options, and non sequiturs in verbatim responses.
- Are there differences in the multivariate relationships within and across sampling frames?
- Is there any notable difference in our bias measures when sample sizes are

held constant for each survey. Preliminary analysis based on 1,000 replicate samples for each survey, when the sample size for each survey is constrained to a Life in Australia™ equivalent sample size (n=582), show very little difference in the mean bias measures based on the original sample size for each survey and the resultant measure from the 1,000 resampled replicates. Still, we would like to pursue this line of inquiry.

We conclude with a plea for transparency, especially about the recruiting and sampling practices used by non-probability panel providers. Methodological disclosure can only enhance the credibility of the method overall and may lead to methodological insights that further improve the accuracy of the estimates generated from such panels. If this happens survey researchers may be able to use non-probability online panels with more confidence.

Notes

[1] Probability and non-probability surveys share the common objective of wanting to efficiently estimate the characteristics of a large population based on measurements of a subset of that population. Both therefore ideally require that (i) the sampled units are exchangeable with non-sampled units that share the same measured characteristics, (ii) no parts of the population are systematically excluded from the sample, and (iii) the composition of the sampled units with respect to observed characteristics either matches or can be adjusted to match the composition of the larger population. The crucial difference between the sampling methods is that probability samples require that each member of the population has a known non-zero chance of being included in the sample and as such a known probability of selection. This in turn enables a degree of confidence for each estimate to be calculated based on established mathematical principles. By contrast, non-probability approaches do not rely on each sampling unit having a known non-zero chance of selection and the consequent mathematical principles to assert a known degree of confidence about their estimates but instead rely on untested modelling assumption to attest to the accuracy of the resultant estimates (Cornesse et al. 2020, 7).

[2] Several large-scale data breaches occurred in Australia in the second half of 2022 attracting widespread publicity and heightening public concern about this topic. The Office of the Australian Information Commissioner (OIAC) reported an increase of 26% in notifiable breaches in the second half of 2022 (OAIC, 2022). This included the second largest data breach ever reported in Australia, the Optus data breach with potentially 9.8 million customers impacted (Turnbull, 2022). Other widely reported data breaches in 2022 included the Medibank data breach with over a quarter of a million records potentially compromised (Min, 2022) and VicRoads data breach where 942,000 Victorian motor vehicle licence holders had their details compromised (Cowie, 2022).

[3] Throughout this paper we distinguish between the two earlier studies by labelling the initial 2015 study the OPBS and the later study which was expanded to include the results from the same questionnaire being administered to members of Life in Australia™ as OPBS+. For the sake of convenience, although the first study was fielded in November 2015 and the second in January 2017, when it comes to time-series comparisons with the existing study, we label the first study as the 2015 study and the current study as the 2022 study.

[4] G-NAF is maintained by Geoscope Australia (formerly the Public Sector Mapping Authority) and is the authoritative national address index for Australia. The sample was selected from the G-NAF database using a stratified sample design in accordance with the distribution of the Australian residential population aged 18 years and above.

[5] Data collection via the use of video conferencing platforms such as Zoom, Webex, Teams, etc. goes by various names including Video Assisted Live Interviewing (VALI), Video Interviewing and Computer-Assisted Video Interviewing (Hanson, 2021; Schober et al., 2020). Whatever nomenclature is used, the concept is the same: data being collected by an interviewer from a respondent via a synchronous two-way video call with the interviewer entering the data into a programmed survey questionnaire. Within these basic parameters a great deal of variation in how

such interviews are administered is possible. For example, the decision to use prompt cards or not.

[6] SMS push-to-web is what would be called ‘text-to-web’ in an American context.

[7] The sample frames for the CATI and SMS push-to-web surveys were purchased from SamplePages, the only remaining Australian-based supplier of Australian RDD sample. SamplePages selects numbers randomly from the Australian Communication and Media Authority’s register of numbers, which shows all allocated blocks of mobile numbers (i.e., telephone number prefixes that are potentially in use). SamplePages does not use a list-assisted approach (Brick et al., 1995); a pure RDD sample is drawn. Before release to the survey company, sampled numbers undergo home location register look-up to check for active status (a process sometimes called ‘pulsing’ or ‘pinging’), with inactive numbers excluded. SamplePages reports a 1 per cent false negative rate for these checks for active status. When a person was reached for the ACSSM CATI surveys, the phone answerer / SMS recipient was the selected respondent, provided they were an adult aged 18 and above and resident in Australia. Coverage of the mobile RDD frame is estimated at 95 per cent of the Australian adult population (ACMA, 2022a).

[8] The protocols established for the high effort CATI survey comprised the use of autodialling technology in conjunction with the following; the sending of a pre-notification SMS 1 day prior to sending a survey invitation link via SMS, a maximum of 6 contact attempts or 4 consecutive not-contacts – whichever was reached first and leaving an automated message when a voicemail was first encountered. The low effort CATI survey comprised the use of predicative dialling technology in conjunction with the following call protocols; the sending of a pre-notification SMS 1 day prior to sending a survey invitation link via SMS and a maximum of 4 contact attempts or 2 consecutive not-contacts – whichever was reached first and leaving an automated message when a voicemail was first encountered.

[9] An example of a non-interlocking quota would be a requirement to fulfill a certain number of completed questionnaires by age group AND a certain number of completed questionnaires for a prescribed geographic segment whereas an interlocking quota would require achieving a prescribed number of questionnaires by age group *within* each geographic segment.

[10] Hard quotas set an exact number of questionnaires to be completed per sample segment or cross-classified sample cell whereas soft quotas require either a prescribed minima or working to a loose target.

[11] The terms ‘bias’ and ‘error’ are used interchangeably throughout this report.

[12] Estimates of AAB for each survey have been produced both with and without the Kessler 6 measure. For reasons we have been unable to establish, the Kessler 6 estimates were very inaccurate in 2015. So as not to overstate the change over time we have produced AAB measures both with and without the Kessler 6 item.

[13] The DQF is based on the Statistics Canada (2002) Quality Assurance Framework and European Statistics Code of Practice (Eurostat, 2023).

[14] The bias relativities used for these analyses differ from those presented in the previous section (Table 15).

[15] ACMA’s estimates of the use of telephone for voice calls are derived from a survey conducted on Life in Australia™.

[16] ACMA's estimate of the online population are derived from a survey conducted on Life in Australia™.

[17] The de-briefing interviews were conducted by Philip Carmo of the ABS.

[18] Comparative studies may be an exception to the rule that inferential statistics are not applicable to non-probability samples, as inference is to the population of non-probability samples rather than, e.g., Australian adults.

[19] The IPND is a sampling frame which provides postcodes for mobile numbers that is available for Commonwealth public policy, public health, and Federal, state, and local government electoral matters (ACMA, 2022a).

[20] Most firms conducting CATI interviews in Australia are ADIA members and pay interviewers the same rates, removing this as a potential variable.

[21] There is more non-RDD use of SMS for survey invitations. For example, the Victorian government surveyed recipients of the COVID-19 vaccine via SMS survey. As there was a very clear nexus between a specific event (vaccination) and survey, the context is very different from an RDD survey invitation that comes 'out of the blue' without warning.

[22] An example of meaningful supporting detail was Panel 3's description of their practice of only sending incentives by physical mail and the resulting benefits in reducing the likelihood of fraud. By contrast, most descriptions were very broad and lacking specific detail (e.g., vague references to 'affiliate networks' in recruitment).

[23] There are analogues to population registries in Australia: the electoral roll (under the control of the Australian Electoral Commission) and the Medicare database (under the control of Services Australia). While neither have full coverage of the population, they have still have very high coverage rates. Access for research is, however, limited. The electoral commission has increasingly scrutinised applications and the Medicare database generally requires consent of individuals before passing contact information to researchers, which likely increases non-response error.

Level 5, 350 Queen Street
Melbourne VIC 3000

Locked Bag 13800
Law Courts VIC 8010

Phone (61 3) 9236 8500
Fax (61 3) 9605 5422

info@srcentre.com.au
srcentre.com.au

Social
Research
Centre

Fully owned by

Australian
National
University