methods, data, analyses | Vol. 17(1), 2023, pp. 93-120 DOI: 10.12758/mda.2022.11

Do We Have to Mix Modes in Probability-Based Online Panel Research to Obtain More Accurate Results?

Sebastian Kocar[1] & *Nicholas Biddle*[2]

[1] *Institute for Social Change, University of Tasmania*

[2] *ANU Centre for Social Research and Methods, The Australian National University*

Abstract

Online probability-based panels often apply two or more data collection modes to cover both the online and offline populations with the aim of obtaining results that are more representative of the population of interest. This study used such a panel to investigate how necessary it is, from the coverage error standpoint, to include the offline population by mixing modes in online panel survey research. This study evaluated the problem from three different perspectives: undercoverage bias, bias related to survey item topics and variable characteristics, and accuracy of online-only samples relative to nationally representative benchmarks. The results indicated that attitudinal, behavioral, and factual differences between the online and offline populations in Australia are, on average, minor. This means that, considering that survey research commonly includes a relatively low proportion of the offline population, survey estimates would not be significantly affected if probability-based panels did not mix modes and instead were online only, for the majority of topics. The benchmarking analysis showed that mixing the online mode with the offline mode did not improve the average accuracy of estimates relative to nationally representative benchmarks. Based on these findings, it is argued that other online panels should study this issue from different perspectives using the approaches proposed in this paper. There might also be an argument for (temporarily) excluding the offline population in probability-based online panel research in particular country contexts as this might have practical implications.

Keywords: online panels, online and offline populations, mixed-mode data collection, representation errors, benchmarking

Mixed-mode survey research is becoming increasingly common, and the use of web surveys offers a range of opportunities for mixing modes of data collection (Bryman 2016). There are many reasons for employing mixed modes, but the following three are especially common: to reduce costs, to maximize responses, and to save money in longitudinal surveys (Groves et al., 2009, p. 175). In addition to these benefits, probability-based online panels[1] often apply two or more data collection modes to cover both the online and offline populations (Baker et al., 2010). While some of them collect data online only (e.g., Norwegian Citizen Panel), including by providing hardware with internet access (mixed-device, e.g., American Trends Panel, ELIPSS or LISS), others combine the online mode with telephone (e.g., Life in Australia™), mail (e.g., GESIS Panel), and face-to-face (e.g., KAMOS) data collection as the offline modes (Kaczmirek et al., 2019, pp. 4-5).

Generally speaking, mixing modes in probability-based online panel or web-push research might be necessary since internet-only samples may not be representative of the general adult population. This is due to significant differences in demographic and other characteristics between the online and offline populations (Baker et al., 2010), which still exist despite an increase of internet penetration over time (Mohorko et al., 2013; Sterrett et al., 2017). For example, in the United States in 2015, it was reported that 11% of adults did not self-identify as internet users and there were differences between the online and the offline populations (so-called onliners and offliners) in terms of age, race, marital status, education, and income (Keeter et al., 2015; Sterrett et al., 2017). In Europe, there were substantial differences in internet access between countries, as well as differences between the online-offline populations in age, gender, and education (Mohorko et al., 2013). In Australia, it was estimated that about 14% of Australian households did not have home internet access (Australian Bureau of Statistics, 2018), and there were notable differences between people with or without access to the internet in terms of age, location (urban-rural), employment status, qualifications, gender, household

[1] More often than not, probability-based online panels collect data from the offline population using an alternative offline mode, such as telephone and mail (Kocar & Kaczmirek, 2021). This makes most probability-based (predominantly) online panels, active as of 2021, mixed-mode panels. In this study, we use the term "probability-based online panels", which is consistent with terminology from Callegaro et al. (2014) and Baker et al. (2010).

Acknowledgements

We would like to acknowledge the contributions of the Social Research Centre for collecting and providing access to their survey data (Life in Australia™) and of the Department of Education, Skills and Employment for PhD program funding.

Direct correspondence to

Sebastian Kocar, Institute for Social Change, University of Tasmania
E-mail: sebastian.kocar@utas.edu.au

income, and country of birth (De Vaus, 2013, pp. 76-77). In addition, not every person with an internet connection has the skills or inclination to participate online (Pennay et al., 2016), which further decreases the share of the online population (Keeter et al., 2015), and the evidence suggest that those panellists should ideally be offered an offline mode to achieve better representation instead of providing them with technology (Cornesse & Schaurer, 2021). For all those reasons, an offline survey mode should be included or at least considered in probability-based panel research (Pennay et al., 2016).

To represent the general population, online panels have to find a way to include people without computer or internet access while balancing measurement equivalence and coverage (Blom et al., 2016). Besides to not introduce socio-demographic coverage bias, data are collected from the offline population in mixed-mode survey research to reduce potential non-demographic coverage bias. While socio-demographic bias can be mitigated with calibration, the same approach is less effective in reducing non-demographic coverage bias in probability online panels (see Rookey et al., 2008, p. 965). There has been limited research on the effect of undercoverage bias in online panels on the accuracy of derived non-demographic estimates, especially in the case of complete exclusion of the offline population (e.g., Eckman, 2016) and relative to nationally representative benchmarks. Furthermore, because internet access and willingness to complete surveys online is changing so rapidly and varies across different country contexts, studies that have been undertaken may need to be updated with more recent data and/or in different geographic/cultural contexts. As Kaczmirek et al. (2019, p. 3) raised a question if the offline population should even be included in probability-based online panel research to balance different types of errors and practical considerations (e.g., time, cost, questionnaire design), this research addresses the problem of undercoverage bias[2] and its effect on the accuracy/consistency by using data from six Australian probability-based online panel surveys. By comparing the estimates from online and offline (telephone) samples, the study aims to address the following research questions:

- *RQ1: How much undercoverage bias would there be if the offline population was completely excluded from probability-based online panel research?*
- *RQ2: What question and variable characteristics, such as question topic, represent the biggest differences between onliners and offliners?*

[2] 'Undercoverage bias' investigated in this paper is a hypothetical undercoverage bias which would be the result of completely excluding the offline population. Undercoverage bias is, in practice, measured as attitudinal, behavioral, knowledge, and factual differences between the populations (online vs offline), as well as the effect of those differences on the estimates in case of exclusion of the offline population. As of 2021, the probability-based online panel investigated in our study is a mixed-mode panel (online and telephone modes).

- *RQ3: Does calibration (raking) reduce the non-demographic differences between onliners and offliners?*
- *RQ4: Does including the offline population improve the accuracy of estimates relative to the nationally representative benchmarks?*

Before addressing these research questions, we will present the contemporary research on this highly relevant topic for the online panel research practice and build the study on the existing evidence on undercoverage bias in probability-based online panels.

Literature Review

Socio-demographic Undercoverage Bias in Online Panels

Including both online and offline populations in probability-based online panel research generally reduces undercoverage bias and results in better socio-demographic coverage. For example, the complete GIP (Germany) and LISS (the Netherlands) panels, which include both online and offline respondents, were found to be closer to the general populations than the population consisting of online respondents only (Blom et al., 2017; Leenheer & Scherpenzeel, 2013). Previous research has shown that online and offline populations in probability-based online panels differ in various socio-demographic characteristics, which are often consistent across online panels[3] from different countries. Some of those characteristics are *age* (Blom et al., 2015; Blom et al., 2017; Bosnjak et al., 2013; Hoogendoorn & Daalmans, 2009; Keeter et al., 2015; Leenheer & Scherpenzeel, 2013; Toepoel & Hendriks, 2016), *gender* (Blom et al., 2015; Blom et al., 2017), *education* (Bosnjak et al., 2013; Cornesse & Schaurer, 2021; Keeter et al., 2015; Revilla et al., 2016; Toepoel & Hendriks, 2016), *household size/structure/couple status* (Blom et al., 2017; Keeter et al., 2015; Leenheer & Scherpenzeel, 2013; Revilla et al., 2016; Toepoel & Hendriks, 2016), *ethnical background* (Blom et al., 2017; Keeter et al., 2015; Leenheer & Scherpenzeel, 2013; Toepoel & Hendriks, 2016)*, urbanization level* (Blom et al., 2017; Keeter et al., 2015; Leenheer & Scherpenzeel, 2013), *religion* (Keeter et al., 2015; Toepoel & Hendriks, 2016), *sexual orientation* (Zhang et al., 2009) and *income* (Bosnjak et al., 2013; Hoogendoorn & Daalmans, 2009;

[3] Differences in those characteristics have been reported for CentERdata (Hoogendoorn & Daalmans, 2009), LISS (Leenheer & Scherpenzeel, 2013; Toepoel & Hendriks, 2016), German Internet Panel (Blom et al., 2015; Blom et al., 2017), GESIS Panel (Bosnjak et al., 2013), ELIPSS (Revilla et al., 2016), and American Trends Panel (Keeter et al., 2015).

Keeter et al., 2015; Toepoel & Hendriks, 2016). Most of those characteristics are not included as covariates in typical post-stratification weighting.

Furthermore, non-internet households have lower response rates and higher attrition rates (Blom et al., 2017; Leenheer & Scherpenzeel, 2013; Revilla et al., 2016), which would ideally be accounted for in post-survey adjustment and panel recruitment/refreshment. However, including offliners results in more representative samples in comparison to weighting adjustments (Blom et al., 2017). Also, and more importantly, the main issue is that an exclusion of the offline population from probability-based online panel research does not only result in socio-demographic representation bias, but in potentially biased estimates for many survey topics (Kaczmirek et al., 2019). A few different studies have already looked at fundamental non-demographic differences between onliners and offliners for which no adequate benchmarks were available.

Attitudinal, Behavioral, and Other Factual Differences Between the Online and Offline Populations

The evidence suggests there are notable non-demographic differences between online and offline populations in probability-based online panel research, with or without statistically significant undercoverage bias and its effect on the final survey estimates. The differences between the populations are best captured in topics strongly related to internet access (Eckman, 2016), and internet and technology (Keeter et al., 2015). They can also be observed for various attitudes, behaviors, beliefs and other concepts such as: *political attitudes, knowledge, voting and civic actions* (Blom et al., 2017; Keeter et al., 2015; Pforr & Dannwolf, 2017; Toepoel & Hendriks, 2016; Zhang et al., 2009), *personality traits* (Bosnjak et al., 2013; Schaurer & Weiß, 2020; Toepoel & Hendriks, 2016), *health* (Toepoel & Hendriks, 2016), *purchasing power* (Blom et al., 2015), *financial circumstance* (Keeter et al., 2015; Toepoel & Hendriks, 2016), *housing (*Toepoel & Hendriks, 2016), *media consumption* (Pforr & Dannwolf, 2017), and *compliance with COVID-19 safety measures* (Schaurer & Weiß, 2020).

It has been reported that online and offline respondents differ in between one-third (Keeter et al., 2015; Rookey et al., 2008) and two-fifths (Eckman, 2016, p. 47) of attitudinal and behavioral questions (with statistically significant differences), and there seem be no trends in the direction, questionnaire section, or question type (Rookey et al., 2008). While the differences between the populations often tend to be relatively modest (Keeter et al., 2015), and univariate difference often do not translate into statistically significant differences at the multivariate level in countries with high internet penetration (Eckman, 2016), certain target groups are with much greater differences between the online and offline populations, such as those 65 years of age and older (Keeter et al., 2015).

Socio-demographic bias in data (if observable) can be reduced with different post-survey methods, such as post-stratification weighting which adjusts the sample totals to the population totals using nationally representative benchmarks (Kalton & Flores-Cervantes, 2003). On the other hand, weighting adjustment using socio-demographic covariates (including with regression models like GREG) does not sufficiently reduce non-demographic differences between onliners and offliners in probability-based online panel research (e.g., Pforr & Dannwolf, 2017; Rookey et al., 2008; Zhang et al., 2009). This suggests that excluding the offline population cannot be sufficiently adjusted with calibration or other post-survey adjustment methodology.

Estimation of Survey Accuracy with Benchmarking

There are at least two ways of estimating the effect of undercoverage bias on the accuracy of estimates. One way is by comparing survey results including the offline population with those excluding this population (see Eckman, 2016; Keeter et al., 2015; Rookey et al., 2008). The other approach is to compare the results obtained with and without the offline population with the estimates derived from a representative external data source – usually an expensive and sufficiently large government survey with great attention to data quality and accuracy of survey estimates (Bialik, 2018).

The practice of benchmarking is often used to study the accuracy of nonprobability-based online panels in comparison to probability-based ones (e.g., Kaczmirek et al., 2019; MacInnis et al., 2018; Pennay et al., 2018; Yeager et al., 2011), to perform mode effect analyses (Vannieuwenhuyze & Loosveldt, 2013), and to check the accuracy of findings in surveys and determine how to improve survey quality (Bialik, 2018). Benchmarking analysis can represent added value in coverage error research because the differences in distributions, which could be attributed to measurement mode effects in mixed-mode online panels, can add a net effect on undercoverage bias. Another advantage of high-quality government survey benchmarks is that they are often carried out with single-mode data collection (Vannieuwenhuyze & Loosveldt, 2013). On the other hand, the disadvantage of benchmarking analysis is that the required national representative data for non-factual and knowledge items are often not available, and in some cases, there is less trust in the validity of benchmarks[4] (Singh, 2011).

In this study, we use both approaches to estimation of undercoverage bias. The added value of this research is an ability to compare attitudinal, behavioral and

[4] This appears to be a less of an issue in certain countries (including in Australia, where this study was undertaken) where official statistical agencies are able to compel potential respondents to complete their surveys with the use of financial sanctions for those that do not comply.

other factual estimates to nationally representative non-demographic benchmarks due to a well-planned questionnaire design in one of the analyzed surveys.

Methods

Data

We analyzed data from the Life in Australia™ surveys. Specifically, six out of the first 16 waves before the first panel refreshment in June 2018 were used in this study. Life in Australia™ is the only probability-based online panel in Australia. It was established and is managed by the Social Research Centre. The panel has been used to collect data on important topics for different clients, from academic to government and non-governmental organizations (see the list of studies in Kaczmirek et al., 2019, p. 20). However, as those research projects were funded by different clients, the current study only had access to the data collected for the Australian National University (ANU) as the largest Life in Australia™ client (waves 1, 2, 3, 7, 10, and 14). We used all available data to increase the range of survey topics and the number of survey items, required for greater statistical power to address RQ2. More information about the surveys is provided in Table 1 below.

While all six data files were analyzed to address research questions RQ1-3, only one out of the six data sources could be used in the benchmarking part of the study[5] (RQ4) due to the unavailability of high-quality nationally representative benchmarks for the majority of the Life in Australia™ substantive survey items. The Health, Wellbeing, and Technology Survey 2017 (also known as Life in Australia™ Wave 2, Pennay & Neiger, 2020) was analyzed to study the accuracy of estimates relative to nationally representative estimates. The questionnaire was designed based on the availability of high-quality benchmarks for Australia (see Table 5 in the Appendix) to study the accuracy of a probability-based online panel. Life in Australia™ Wave 2 data files can as well be used to establish the accuracy of online-only samples in comparison to mixed-mode samples.

[5] While there was a very small number of national level estimates included in the other five Life in Australia™ waves, including from the Household, Income, and Labour Dynamics in Australia (HILDA) Survey, we considered benchmark uncertainty from this source too large due to sample attrition and the HILDA panel not being refreshed since 2011.

Table 1 Life in Australia™ survey data collected for the ANU

Title of Life in Australia™ survey	Month and year	Wave	Final sample size	Completion rate (COMR)	Data DOI
Australian Personas Survey, 2016	December 2016	1	n=2,603	78.8%	10.26193/JFWRPI
Health, Wellbeing and Technology Survey 2017	January 2017	2	n=2,580	78.6%	10.26193/YF8AF1
ANU Poll 2017: Housing	March 2017	3	n=2,513	77.7%	10.26193/EL5WHN
ANU Omnibus Survey 2017	July 2017	7	n=2,290	74.3%	/
ANU Poll 2017: Job Security	October 2017	10	n=2,270	74.6%	10.26193/7OP0TI
World Values Survey, 2018	April 2018	14	n=2,106	71.4%	10.26193/ZXF0SQ

Population, Samples, and Data Collection Modes

In Life in Australia™, the panellists are defined as "residents of Australia aged 18 years or older (English speaking)" and were recruited in the second half of the year 2016 (n=3,322). The response rate at the establishment of the panel, calculated as the product of the recruitment rate and the profile rate, was 15.5% (AAPOR RR3 (The American Association for Public Opinion Research 2016)). To undertake recruitment, a dual-frame Random Digit Dialing (RDD) sample design was employed, with a 60:40 (pilot) and 70:30 (the main recruitment effort) split between mobile phone and landline sample frames[6]. The last birthday method was used to select potential panel members in landline frames and the phone answerers were selected for the mobile sample; only one person per household was invited to join the panel. Out of all panellists who were recruited, joined the panel, and were later invited to monthly surveys on different topics, about 87% can be defined as online (onliners) and about 13% as offline panellists (offliners). The online self-completion mode (CAWI) was used to collect data from the online panellists and the telephone mode (CATI) was used to cover the offline population. Data were collected at approximately monthly intervals. An incentives scheme was used for recruitment and monthly data collection – conditional incentives $10 per wave, with pan-

[6] Baffour et al. (2016) reported that 95% of Australians own a mobile phone and 80% of Australians have a landline, using single frames would lead to significant differences in estimates of populations' characteristics, and better coverage is provided in dual-frame telephone surveys.

ellists either receiving a supermarket coupon or donating to charity (Kaczmirek et al., 2019). As can be seen in Table 1, the Life in Australia™ survey sample size decreased with each survey, which is a result of an increasing proportion of nonrespondents over time, as well as accumulating voluntary panel attrition.

Data Processing and Analysis

There are three main components of this study: (1) undercoverage bias – extent of univariate bias (RQ1), (2) undercoverage bias – survey item characteristics (RQ2 and RQ3), and (3) benchmarking analysis (RQ4). We will present analytical approaches for each of these components separately. All data processing and analyses, except for multiple linear regression analyses in the second component (Stata), were carried using R software. The following packages were used for functions not directly provided by R's base or stats packages: *Hmisc, missforest, anesrake, survey, sjstats,* and *questionr.*

Undercoverage bias – extent of univariate bias. To estimate undercoverage bias at the univariate level in all six surveys and present evidence to answer RQ1, the following adapted Equation 1 from Eckman (2016) for absolute relative bias was used:

$$absolute\ relative\ bias\ (\bar{Y}_{web}) = \left|\frac{\bar{y}_{web} - \bar{y}_{combined}}{\bar{y}_{combined}}\right| \tag{1}$$

where $\bar{y}_{web}$ is the mean from the online population (excluding offliners) and $\bar{y}_{combined}$ is the mean from the full sample (onliners and offliners). Because the variables were measured in different units, absolute relative bias was estimated and averaged across all items (reporting median). The statistical significance of undercoverage bias was tested with different tests/models, with a significant regression coefficient indicating bias (consistent with Eckman, 2016). In addition to Chi-Square testing with nominal variables, linear (continuous variables), binary logistic (dichotomous variables), and ordinal regression models (ordinal variables) were analyzed with a substantive survey item as the response variable and the population as the predictor (0=online, 1=telephone).

Since the majority of survey items were categorical (nominal and ordinal), dummy variables were also created for those variables (e.g., an ordinal variable with five levels generated five dichotomous variables) and their absolute relative bias was compared. As different statistical tests must be used to test for significant differences in categorical variables, relative distance had to be calculated alternatively, like with sets of dummies. In practice, such results are often reported for one variable category only, e.g., the percentage of people strongly agreeing with a particular statement, which further justifies the undercoverage bias calculation with dummies.

Undercoverage bias – survey item characteristics. To extend the bias estimation findings and present evidence to answer RQ2, multiple linear regression models were created (see Equation 2):

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon \qquad (2)$$

where Y is the effect size, X_1 - X_n are the survey item characteristics such as item topic and question content, and ε is the error.

Comparison of the distributions of onliners and offliners was carried out by calculating effect sizes (Y from Equation 2) as measures of association between pairs of variables. A total of 368 effects sizes were calculated for associations between each of 368 substantive items from six Life in Australia™ surveys (listed in Table 1) and mode of completion (0=online, 1=telephone). The calculated differences between both populations were based on Cramer's V and Rank-Biserial Correlation (R-BS) measures as effect sizes; a higher coefficient value represented a greater difference between the studied populations in the concept measured. Two different effect size measures had to be used to calculate effect sizes for different variable types (nominal, ordinal, continuous), and they were calculated with both unweighted and weighted data. By raking survey data, the sample totals were adjusted to the selected population totals for both onliners and offliners separately. It was assumed that weighting would decrease some of the undercoverage bias.

The variable information (X_1 - X_n from Equation 2) was coded for all 368 variables from six Life in Australia™ waves. Using the European Language Social Science Thesaurus (ELSST) (UK Data Service, n.d.), survey item topics were identified and combined into 20 distinctive broad topics – the most common was *values and social capital* (12.8%), followed by *housing* and *finance* (both at 7.3%). To code the question content by type, the classification by Dillman (1978) was used; out of the four types, the combined attitudes and beliefs category was the most common type (65.5%), followed by behaviors (19.0%). The following variable types were used in the models: binary, nominal with 3+ categories, ordinal, and continuous (combining interval and ratio variable types). The most common variable type was ordinal (50.5%), followed by binary (33.7%). The modal categories were used as reference categories in the regression models presented in the Results section (e.g., *values and social capital* for broad topic).

For better statistical power, the Life in Australia™ ordinal variables were included in all models, both the ones for categorical variables (with *Cramer's V value* as the dependent variable) and for models with non-parametric effect sizes as dependent variables (with *Rank-Biserial Correlation coefficient*). Since R-BS coefficient values range from −1 to 1, and we were only interested in the magnitude of effect sizes and not the direction, an absolute version of *R-BS coefficient* with positive values only was used.

As the effect sizes were derived from the data collected from the same respondents in the same wave and partially matching respondents in different waves (due to unit nonresponse and voluntary attrition), we had to identify a way of dealing with dependencies in the data so as not to violate any assumptions of ordinary least squares regression. The literature suggests approaches such as panel data analysis, bootstrapping regression models, and regression with clustering. Here, it was decided to carry out a combination of bootstrapping and clustering. Bootstrapping was carried out to mitigate the problem of dependencies and calculate standard errors more accurately (Fox, 2015). Clustering was carried out to deal with regression model errors potentially being independent across clusters but correlated within clusters, i.e., waves with a unique sample composition (Cameron & Miller, 2015). This was performed using Stata 13.

For more detailed technical information about the selection of statistical tests and effect size measures, the selection of substantive survey items, data processing, coding, raking, and statistical modeling, please see the Appendix.

Benchmarking analysis. In this part of the research, the results from Life in Australia™ Wave 2 survey were compared with the nationally representative benchmarks listed in Table 5 (see Appendix). All substantive measures from the study from Kaczmirek et al. (2019) were selected for use in our analyses, which partially replicated the approach of the original benchmarking study. To measure bias, the average absolute error (AAE) measure proposed by Yeager et al. (2011) was used (see Equation 3), which was computed across three categories, that is secondary demographics, substantive items, and combined secondary demographics and substantive items:

$$AAE = \sum_{j=1}^{k} \frac{|\widehat{y_j} - y_j|}{k} \tag{3}$$

where $\widehat{y_j}$ is the j-th estimate from Life in Australia™ Wave 2 survey and y_j is the value for a corresponding benchmark. To estimate the accuracy of the online-only samples, the AAE values were compared between the online-only and online-offline samples. Bootstrapping was used to test for statistical significance of differences[7]. The absolute relative bias measure from the undercoverage bias estimation (see Equation 1) was also used in this part of the article.

Weighted estimates for the selected items and for all analyzed samples, in addition to the unweighted estimates, were calculated to assess the effect of calibra-

[7] Following Pennay et al. (2018, pp. 14-15) and Yeager et al. (2011), we used bootstrapping (n=1000 replications, each drawn sample was reweighted/raked to match sociodemographic population benchmarks) to calculate standard errors and to carry out statistical testing.

tion on bias. It was decided to employ a consistent approach with no base weights derived. Raking weights were calculated for each sample separately, i.e., the online-offline and online-only samples, to balance the samples on key socio-demographics. The same raking covariates/primary demographics as in Kaczmirek et al. (2019) were used, while in contrast, the weighting benchmarks were taken from the Australian Census 2016 (Australian Bureau of Statistics, 2016). Raking was carried out to adjust the samples to the national distributions by gender, age by education, state by capital city in state, country of birth (Australia, English-speaking background, non-English-speaking background), and telephone status (mobile, landline, dual user). All larger weights were trimmed down to a value of 5. The random forest technique was used to impute missing values (Stekhoven & Buehlmann, 2012) for the listed weighting variables so as not to exclude any cases with valid values for substantive items.

Benchmarks from some of the largest government-funded national surveys in Australia were used in this study: the Australian Census 2016 (Australian Bureau of Statistics, 2016), National Health Survey 2014-15 (Australian Bureau of Statistics, 2015), the National Drug Strategy Household Survey 2013 (Jefferson, 2015) and General Social Survey 2014, as well as the Australian Electoral Commission (2015) administrative data (benchmarks from Kaczmirek et al., 2019). These surveys should be considered as the best quality social research data sources in Australia, and the validity of the benchmarks should be the highest. For more methodological details, see Table 5 in the Appendix.

Results

This section will present the results of all analyses. It is divided into the following subsections: undercoverage bias – extent of univariate bias, undercoverage bias – survey item characteristics, and accuracy of estimation – benchmarking.

Undercoverage Bias – Extent of Univariate Bias

This section addresses the first research question, RQ1. To do so, the analysis from Eckman (2016) was partially replicated. To showcase the magnitude of differences between the populations, data were not weighted in the following analyses studying bias[8].

[8] Since Eckman (2016, p. 46) did not use weights and we applied the same analytical strategy to address RQ1, weighting was not used here in the univariate undercoverage bias part of the analysis for comparability purposes. The effect of weighting on undercoverage bias (RQ3) is addressed in the 'survey item characteristics' and 'benchmarking' subsections of the Results.

Table 2 Undercoverage bias in six Life in Australia™ waves

Wave	% offline panellists	Variables with significant* undercoverage bias[a] (n)	Dummy and continuous variables with significant* undercoverage bias[b] (n)	Absolute relative bias[c] (ARB) Median (n)
1	12.9%	77.4% (106)	55.3% (512)	6.4% (512)
2	13.8%	69.1% (55)	63.8% (232)	5.9% (232)
3	13.5%	52.2% (46)	32.9% (228)	5.0% (228)
7	14.2%	80.0% (45)	57.2% (201)	6.3% (201)
10	14.1%	62.5% (48)	34.5% (229)	4.7% (229)
14	14.1%	72.1% (68)	58.6% (251)	5.3% (251)

[a] Each variable is tested for undercoverage bias, no matter the scale (total n=368), [b] Each categorical variable is recoded into a set of dummy variables and tested for undercoverage bias together with all continuous variables (total n=1,653), [c] absolute relative bias can be reported for all newly created dummies and continuous variables (total n=1,653), *p<0.05.

The results in Table 2 reveal a fairly significant bias at the univariate level. With between 12.9% and 14.1% of offliners participating in the Life in Australia™ surveys, the results indicated that between 52.2% (out of 46, Wave 3) and 80.0% (out of 45, Wave 7) of items exhibited significant undercoverage bias, as determined by significance testing with regression modeling and Chi-Square testing. Further, dummy variables were generated from all categorical variables to estimate the average absolute bias. In this study, the median absolute relative bias was between 4.7% (Wave 10) and 6.4% (Wave 1), which is substantially more than in the study by Eckman (2016). Absolute relative bias seemed to be associated with significant undercoverage bias as examined with dummy variables (and a limited number of interval/ratio variables) and was less severe than the bias observed with the original variables. As categorical variables were split into dichotomous variables with lower proportions, and onliners and offliners might not differ in every single dimension measured by the variable, undercoverage was significant for a smaller portion (between 34.5% (Wave 3) and 63.8% (Wave 2)) of variables/variable categories.

Undercoverage Bias – Survey Item Characteristics

To identify the differences between onliners and offliners, which may be more generalizable than only comparing the distributions of individual items (univariate bias) or their dummies, four multiple linear regression models were constructed to address the second and third research questions RQ2 and RQ3. We primarily attempt to identify survey topics with the largest differences between the online and the offline populations to add new evidence to the existing research in the field (see 'Attitudinal, behavioral, and other factual differences between the online and offline populations' subsection of the Literature review), while also presenting the magnitude of those differences.

The results in Table 3 reveal some non-negligible differences between onliners and offliners which can be observed for the vast majority of topics - given that the reference category for *values and social capital* was fairly average in terms of the mean effect size[9], the non-significant coefficient should be interpreted as no difference between that topic and *values and social capital.* To address RQ2, the most significant topical differences measured with Cramer's V were observed for *international relations,* followed by *internet*[10]. Out of the other topics, *public figures and health, media* and *finance* (the latter only after weighting) had average effect sizes and *household and family, science and technology,* and *government and policy* items had below-average effect sizes. *Household and family* stood out as a topic with very few average differences between the online and offline populations.

Table 3 Ordinary least squares regression models with predictors of differences between onliners and offliners (carried out with bootstrapping and clustering – clusters as Life in Australia™ waves)

	Cramer's V, weighted data		Cramer's V, unweighted data		R-BS coefficient, weighted data		R-BS coefficient, unweighted data	
Predictors	Beta coef.	p value	Beta coef.	p value	Beta coef.	p value	Beta coef.	p value
Broad topics								
Values and social capital	0		0		0		0	
Environment	0.032	0.244	0.032	0.258	0.100	0.062	0.084	0.000**
Finance	0.024	0.000**	-0.010	0.680	-0.049	0.000**	-0.024	0.000**
Gender equality	0.003	0.714	0.002	0.627	0.042	0.219	0.049	0.000**
Government and policy	-0.015	0.000**	-0.026	0.000**	-0.030	0.000**	-0.037	0.000**
Health	0.032	0.000**	0.016	0.000**	0.007	0.010*	0.002	0.508
Household and family	-0.063	0.000**	-0.069	0.000**	0.063	0.148	-0.155	0.007**
Housing	0.004	0.886	-0.003	0.844	0.023	0.632	0.031	0.348
Internet	0.114	0.000**	0.166	0.000**	0.328	0.000**	0.466	0.000**
Labor, employment, work	-0.004	0.610	-0.045	0.000**	0.019	0.026*	0.252	0.003**

[9] Constants equal to between 0.106 (R-BS coefficient, weighted data) and 0.135 (Cramer's V, unweighted data).

[10] This topic stood out even after several internet items with the highest effect size values were removed as part of outlier detection analysis and treatment. More procedural details about excluding outliers are provided in the Appendix.

	Cramer's V, weighted data		Cramer's V, unweighted data		R-BS coefficient, weighted data		R-BS coefficient, unweighted data	
Predictors	Beta coef.	p value	Beta coef.	p value	Beta coef.	p value	Beta coef.	p value
Lifestyle	0.006	0.522	-0.008	0.340	0.025	0.428	0.023	0.000**
Multiculturalism	0.009	0.611	-0.018	0.555	0.034	0.291	0.083	0.003**
Politics and elections	-0.017	0.038*	-0.015	0.023*	-0.038	0.000**	-0.024	0.231
Science and technology	-0.021	0.001**	-0.062	0.000**	0.020	0.435	-0.020	0.001**
Wellbeing	0.005	0.450	-0.032	0.005**	-0.024	0.164	-0.063	0.000**
Discrimination	-0.023	0.013*	-0.012	0.067				
International relations	0.160	0.000**	0.180	0.000**				
Media	0.029	0.001**	0.039	0.000**				
Public figures	0.069	0.000**	0.093	0.000**				
Other	-0.001	0.898	0.012	0.139	0.029	0.013*	0.063	0.000**
Type of question content								
Attitudes and beliefs	0		0		0		0	
Behaviors	-0.006	0.415	-0.006	0.608	-0.031	0.261	0.003	0.430
Attributes	0.008	0.608	0.048	0.001**	0.078	0.000**	0.277	0.000**
Knowledge	-0.024	0.064	-0.070	0.000**				
Variable type								
Ordinal	0		0		0		0	
Nominal	-0.002	0.515	-0.002	0.718				
Binary	-0.039	0.003**	-0.059	0.000**				
Interval/ratio					0.083	0.046*	0.088	0.025*
No. of variable values	0.003	0.000**	0.002	0.000**	-0.003	0.000**	-0.005	0.000**
Constant	0.108	0.000**	0.135	0.000**	0.106	0.000**	0.129	0.000**
N	342		342		194		194	
Adjusted R-Squared	0.349		0.286		0.416		0.563	
Root Mean Square Error	0.053		0.066		0.066		0.083	

*p<0.05, **p<0.01

The R-BS models showed that the differences between onliners and offliners were captured the most prominently in *internet*, but also in *labor, employment, work,* and partially in *health, environment,* and *multiculturalism*. The topics with below-average differences were *finance* (in contrast to the Cramer's V model), *politics and elections,* and *government and policy*. Except for the *internet* and *government and policy* topics (and to some extent *international relations*), there were no observable trends – in some cases, weighting decreased bias in others it had no effect; effect sizes differed substantially between Cramer’s V and R-BS models for the same topics; topics with above and below-average effect sizes could not be grouped further into broader homogenous topics with more or less undercoverage bias.

To address RQ3, both weighted and unweighted estimates of the differences between onliners and offliners are presented. The results show that raking reduced some of the differences between onliners and offliners. After weighting, both the Cramer’s V coefficients for topics and mean Rank Biserial coefficients for topics were decreased (see constants and coefficients), but most of the magnitude of the effect size remained. Nevertheless, on average, the differences between onliners and offliners were small (see the interpretation of effect sizes in Cohen, 1988, pp. 79-81). Moreover, the effect of weighting on the decreased magnitude of differences can be observed for *attributes* as a type of question content. This should come as no surprise since *attributes* are, generally speaking, other “non-weighting” socio-demographic or factual information about respondents and are associated with primary socio-demographics used in calibration. As no other type of question content category stood out as a predictor of differences in the weighted models, it can be concluded that the differences between onliners and offliners, when controlling for primary demographics, are fairly stable across question content.

On the other hand, the differences measured with *binary variables* were smaller than those measured with *ordinal variables* (the reference category) in the Cramer’s V models, and the differences measured with *continuous variables* were greater than those measured with *ordinal variables* in the R-BS Coefficient models. Moreover, the *number of variable values* had a statistically significant effect in all four models. These results indicate that regression modeling and controlling for variable characteristics, in contrast to analyses such as ANOVA, can provide more robust results.

Accuracy of Estimation – Benchmarking

Finally, benchmarking was performed to establish how the observed differences between onliners and offliners affected the accuracy of estimates relative to the nationally representative benchmarks (see Table 4). Our focus was on the comparison of the Life in Australia™ online-offline and online-only samples. With this benchmarking analysis, the aim was to address RQ4. By presenting weighted and unweighted results, we will provide additional evidence to address RQ3.

The primary focus of this analysis was on the comparison of the accuracy of estimates if the offline population was completely excluded. Firstly, the results indicated that the Life in Australia™ estimates for all 18 items with available nationally representative benchmarks would differ very little if no offliners were included. The absolute relative bias (median) was 2.6% for unweighted and 1.7%

for weighted data. For unweighted data, ARB was about half that of the median ARB for all items from all six Life in Australia™ surveys that were analyzed in the first part of this paper (see Table 2, far right column). Also, the difference in ARB between weighted and unweighted Life in Australia™ Wave 2 estimates indicates that weighting can slightly decrease undercoverage bias as the difference between onliners and offliners in practice. This is consistent with our previous results (see Table 3).

Despite observing differences in the average absolute errors between samples with or without offliners, with errors being consistently smaller in samples including offliners (e.g., combined AAE for online+offline, weighted data: 5.41, combined AAE for online only, weighted data: 5.74), none of those differences tested with bootstrapping were statistically significant at $p<0.05$. The evidence suggests that excluding the offline population would not deteriorate the quality of estimates in the Life in Australia™ for the studied concepts. This general finding applies to both calibrated and unweighted data. In the case of secondary demographics, the results showed that weighting (AAE 7.00 -> 5.75) was more efficient in reducing error than including the offline population (AAE 7.00 -> 6.65).

Table 4 Benchmarking results, accuracy relative to the benchmark with and without the offline population

		Life in Australia™ Wave 2			
		Weighted		Unweighted	
Survey item	Benchmark (%)	Online+offline, n=2,580 (Δ in %)	Online only, n=2,166 (Δ in %)	Online+offline, n=2,580 (Δ in %)	Online only, n=2,166 (Δ in %)
Australian citizen	87.12	0.53	0.41	4.47	3.69
Couple with dependent children	38.35	-11.16	-10.70	-14.71	-11.66
Currently employed	61.61	5.38	6.69	0.17	5.33
Enrolled to vote	78.47	7.21	7.10	11.72	10.68
Home ownership with a mortgage	28.82	2.22	2.68	1.30	3.68
Not Indigenous	97.73	-0.23	0.09	-0.06	0.28
Language other than English (speak only English)	76.50	8.62	8.69	8.73	8.17
Living at last address 5 years ago	56.85	1.50	0.50	6.25	4.18
Most disadvantaged quintile for area-based SES	20.00	-6.71	-7.77	-7.29	-8.50
Resident of a major city	66.80	4.15	5.08	2.89	4.99
Voluntary work (none)	79.39	-17.07	-17.08	-20.67	-21.17
Wage and salary income $1000–1249 per week	13.80	-1.64	-2.22	-1.55	-1.69
Consumed alcohol in last 12 months	81.87	3.62	5.03	3.09	5.11
Daily smoker	13.52	-1.97	-3.47	-3.40	-4.98
General health status (very good)	36.20	-2.96	-1.49	-2.60	-0.88
Life satisfaction (8 out of 10)	32.60	-1.24	-0.73	0.07	0.23
Has private health insurance	57.10	3.43	7.03	9.22	12.80
Psychological distress, Kessler 6 (low)	82.20	-17.76	-16.65	-13.63	-13.71

		Life in Australia™ Wave 2			
		Weighted		Unweighted	
Survey item	Benchmark (%)	Online+offline, n=2,580 (Δ in %)	Online only, n=2,166 (Δ in %)	Online+offline, n=2,580 (Δ in %)	Online only, n=2,166 (Δ in %)
Absolute relative bias (ARB), median* (online+offline and online only)		0.017		0.026	
Average absolute error (combined)		5.41	5.74	6.21	6.76
Average absolute error (secondary demographics)		5.53	5.75	6.65	7.00
Average absolute error (substantive items)		5.16	5.73	5.34	6.28

*directly comparing online+offline and online only estimates, Δ in % - difference in percentage points

Discussion and Recommendations

Mixed-mode surveys seem to be almost the standard in probability-based online panel research, but they do not come without a price tag. Increasing costs of interviewer-administered data collection, no threat of mode effects in single-mode surveys, a unified paradata system, and more convenient data collection and panel management are some of the reasons for not carrying out mixed-mode research. Based on the current findings, we share the opinion of Kaczmirek et al. (2019) and Revilla et al. (2016) who discussed the serious dilemma of whether researchers should include offliners (or to provide equipment) to balance different types of error, while not overlooking practical considerations such as time, cost, and questionnaire design.

Making a decision on (temporarily) excluding the offline population is a multi-dimensional problem. One could argue that the offline population should be included no matter the costs due to the offline population being fundamentally different to the online population; this has been supported by evidence from multiple studies (e.g., Eckman, 2016; Keeter et al., 2015; Rookey et al., 2008; Schaurer & Weiß, 2020). Similarly, the undercoverage bias analysis described here revealed statistically significant bias for more than half of all studied variables from all surveys. Yet, the magnitude of differences between the populations, as well as the size of the offline population, should be a factor in the decision making, as the effect of undercoverage is a function of these two dimensions. With statistically significant but relatively small differences, and with a small proportion of offline respondents in the general population (in countries with high internet penetration rates, high-level internet literacy, and low online privacy concerns), there might be a much less significant effect of undercoverage than one would expect. Based on the evidence presented in this study, exclusion of the offline population generally does not substantially affect the derived estimates, which is consistent with findings from Toepoel and Hendriks (2016). However, caution should be taken in the case of probability-based online panels with a larger proportion of offliners, such as the GESIS Panel (see Schaurer & Weiß, 2020).

The findings of this research are based on data from one country only (Australia) and country-specific effects cannot be ruled out. The results indicate that inclusion of the offline population in probability-based online panel research seems to be, to some extent, unnecessary from the coverage error and accuracy perspectives. This could potentially be generalized to other developed countries with high internet penetration rates, narrower socio-economic and demographic distributions, and consequently, relatively minor differences between those with and without internet connection. At the very least, offliners could be temporarily excluded for certain topics which the current study identified as lesser predictors of differences between the populations, such as *household and family, government and policy,* or partially, *finance*. On the other hand, it might be more prudent to think reversely - what items should never be included in probability-based online panel surveys if data are collected from an online sample only, e.g., *internet* or *international relations* items in Life in Australia™. However, overall, the current study observed differences across the majority of topics with no particular trends. This is in line with the findings of Rookey et al. (2008) and other authors who have reported differences for various topics (e.g., Blom et al., 2015; Bosnjak et al., 2013; Eckman, 2016; Keeter et al., 2015; Schaurer & Weiß, 2020; Zhang et al., 2009).

We have to note that the observed bias might well be a result of a combination of fundamental differences between the populations (potential undercoverage bias), differential nonresponse in panel studies over time, as well as measurement error, such as due to measurement mode effects. With our regression modelling, we observed that *variable type* and *number of variable categories* had a significant effect on the differences between onliners and offliners. This indicates that measures of the magnitude of effect size might be more dependent on the number of categories/ranges of continuous variables than theory suggests (see Cohen, 1988; Glass, 1965), and that measurement mode effects were present in our data. For example, in the case of binary variables, the difference between the populations might be smaller due to acquiescence, i.e., tendency to agree with the interviewer. In this study, we did not attempt to disentangle the effects of coverage from the effects of survey participation in different modes on the observed bias. That would require a proper experimental design.

Moreover, the evidence suggests that while differences between onliners and offliners are present in probability-based mixed-mode research in Australia, any negative impacts on data accuracy should be minimal for the majority of topics, question contents, and variable types, even relative to the nationally representative estimates. In this study, we had a privilege to analyze online panel data with corresponding non-weighting benchmarks, something that was not done in previous research on undercoverage bias. Using this approach, we confirmed that online-offline probability-based online panel samples produce slightly different estimates compared to online-only samples, but we could not confirm that those estimates were consistently more accurate. In the future, it would be worth exploring if undercoverage bias and its effect on survey estimates decrease at the bivariate or multivariate level, as previously reported by Eckman (2016) for probability-based online panels and by Biddle et al. (2018) for opt-in panels.

The current analyses were limited, to some extent, by the number of studied items and their characteristics. With a larger sample of items and variables with available benchmarks, possibly from questions related to different broad topics and with more continuous variables, future studies would have greater statistical power and better evidence for data-informed decision making. The current findings might have to be slightly adjusted in that case. This study presents a combined approach to studying undercoverage bias and its effects on data accuracy, and as this was examined in the Australian context only, future research should focus on online-offline population differences in other countries. This is particularly pertinent in regions with both lower internet penetration rates and wider socio-economic and demographic distributions. Such studies could help establish how necessary mixing modes and inclusion of the offline population are in a particular country's context.

References

The American Association for Public Opinion Research. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 9th edition. AAPOR.

Australian Bureau of Statistics. (2015). *National Health Survey 2014-15* [Data set]. Australian Bureau of Statistics.

Australian Bureau of Statistics. (2016). *2016 Census of Population and Housing* [Census TableBuilder], accessed 1 November, 2020.

Australian Bureau of Statistics. (2018, March 28). *Household Internet Access*. http://www.abs.gov.au/ausstats/abs@.nsf/mf/8146.0

Baffour, B., Haynes, M., Western, M., Pennay, D., Misson, S., & Martinez, A. (2016). Weighting strategies for combining data from dual-frame telephone surveys: emerging evidence from Australia. *Journal of Official Statistics*, *32*(3), 549.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., & Zahs, D. (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly, 74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Bialik, K. (2018). *How asking about your sleep, smoking or yoga habits can help pollsters verify their findings*. Pew Research Center.

Biddle, N., Sinibaldi, J., & Sheppard, J. (2018). The social determinants of health and subjective wellbeing: A comparison of probability and nonprobability online panels. *CSRM and SRC Methods Paper, 2018* (6).

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A. S., Das, M., Douhou, S., & Krieger, U. (2016). A comparison of four probability-based online and mixed-mode panels in Europe. *Social Science Computer Review*, *34*(1), 8-25.

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field methods*, *27*(4), 391-408.

Blom, A. G., Herzing, J. M., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2017). Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German Internet Panel. *Social Science Computer Review*, *35*(4), 498-520.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods*, *25*(4), 339-360.

Bryman, A. (2016). *Social research methods*. Oxford University Press.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, *50*(2), 317-372.

Cohen, J. (1988). Statistical Power Analysis *for the Behavioral* Sciences (2nd ed.). Lawrence Erlbaum Associates.

Cornesse, C., & Schaurer, I. (2021). The long-term impact of different offline population inclusion strategies in probability-based online panels: Evidence from the German Internet Panel and the GESIS Panel. Social Science Computer Review, 0894439320984131.

De Vaus, D. (2013). *Surveys in social research*. Routledge.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method* (Vol. 19). Wiley.

Eckman, S. (2016). Does the inclusion of non-internet households in a web panel reduce coverage bias?. *Social Science Computer Review*, *34*(1), 41-58.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Glass, G. V. (1965). A ranking variable analogue of biserial correlation: Implications for short-cut item analysis. *Journal of Educational Measurement*, *2*(1), 91-95.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons.

Hoogendoorn, A., & Daalmans, J. (2009). Nonresponse in the recruitment of an internet panel based on probability sampling. *Survey Research Methods, 3*(2), 59-72.

Jefferson, A. (2015). *National Drug Strategy Household Survey, 2013* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.4225/87/USGEQS

Kaczmirek, L., Phillips, B., Pennay, D. W., Lavrakas, P. J., & Neiger, D. (2019). Building a probability-based online panel: Life in Australia™. *CSRM and SRC Methods Paper, 2019* (2).

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81-97.

Keeter, S., McGeeney, K., Mercer, A., Hatley, N., Patten, E., & Perrin, A. (2015). *Coverage Error in Internet Surveys: Who Web-Only Surveys Miss and How That Affects Results*. Pew Research Center.

Kocar, S. (2018). A universal global measure of univariate and bivariate data utility for anonymised microdata. *CSRM and SRC Methods Paper, 2019* (2).

Kocar, S., & Kaczmirek, L. (2021). *A meta-analysis on worldwide recruitment rates in 23 probability-based online panels, between 2007–2019*. Manuscript submitted for publication.

Leenheer, J., & Scherpenzeel, A. C. (2013). Does it pay off to include non-internet households in an internet panel?. *International Journal of Internet Science*, *8*(1), 17–29.

Lehmann, E. L. (2004). *Elements of large-sample theory*. Springer Science & Business Media.

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. J. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly, 82*(4), 707-744.

Mohorko, A., Leeuw, E. D., & Hox, J. (2013). Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time. *Journal of Official Statistics*, *29*(4): 609-622.

Pennay, D., Borg, K., Neiger, D., Misson, S., Honey, N., & Lavrakas, P. (2016). *Online Panels Benchmarking Study (Technical Report)*. The Social Research Centre.

Pennay, D. W., Neiger, D., Lavrakas, P. J., & Borg, K. (2018). The Online Panels Benchmarking Study: a Total Survey Error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia. *CSRM and SRC Methods Paper, 2018* (2).

Pennay, D., & Neiger, D. (2020). *Health, Wellbeing and Technology Survey (OPBS replication), 2017* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.26193/YF8AF1

Pforr, K., & Dannwolf, T. (2017). What do we lose with online-only surveys? Estimating the bias in selected political variables due to online mode restriction. *Statistics, Politics and Policy*, 8(1), 105-120.

Revilla, M., Cornilleau, A., Cousteaux, A. S., Legleye, S., & de Pedraza, P. (2016). What is the gain in a probability-based online panel of providing internet access to sampling units who previously had no access?. *Social Science Computer Review*, *34*(4), 479-496.

Rookey, B. D., Hanway, S., & Dillman, D. A. (2008). Does a probability-based household panel benefit from assignment to postal response as an alternative to internet-only?. *Public Opinion Quarterly, 72(5)*, 962-984.

Schaurer, I., & Weiß, B. (2020). Investigating selection bias of online surveys on coronavirus-related behavioral outcomes. *Survey research methods*, 14 (2), 103-108.

Singh, L. (2011). Accuracy of web survey data: The state of research on factual questions in surveys. *Information Management and Business Review*, *3*(2), 48-56.

Stekhoven, D. J., & Buehlmann, P. (2012). MissForest - nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118. doi: 10.1093/bioinformatics/btr597

Sterrett, D., Malato, D., Benz, J., Tompson, T., & English, N. (2017). Assessing changes in coverage bias of web surveys in the United States. *Public Opinion Quarterly*, 81(S1), 338-356.

Toepoel, V., & Hendriks, Y. (2016). The impact of non-coverage in web surveys in a country with high internet penetration: Is it (still) useful to provide equipment to non-internet households in the Netherlands?. *International Journal of Internet Science*, *11*(1), 33-50.

UK Data Service. (n.d.). *ELSST – European Language Social Science Thesaurus*. Retrieved November 1, 2020, from https://elsst.ukdataservice.ac.uk/

Vannieuwenhuyze, J. T., & Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: three methods to disentangle selection and measurement effects. *Sociological Methods & Research*, *42*(1), 82-104.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709-747.

Zhang, C., Callegaro, M., Thomas, M., & DiSogra, C. (2009). Do We Hear Different Voices?: Investigating the Differences Between Internet and non-Internet Users On Attitudes and Behaviors. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 6063-6076.

Appendix

Selection of Statistical Tests and Effect Size Measures

In practice, various bivariate measures of association are used for pairs of variables of different types and distributions, such as epsilon squared, eta squared, Spearman's rho, or Pearson's r (see the bivariate effect size review from Kocar, 2018). Most of them were unsuitable for our analysis. For example, Bosnjak et al. (2013), who compared sample composition discrepancies in online panels, used Cohen's d (comparing means) and Hasselblad and Hedges's d (percentages).

However, our study had to use an effect size measure for nominal variables which would indicate the same magnitude of association regardless of the number of cells in the contingency table or the degrees of freedom. Since the minimum number of either rows or columns was always two (modes: online and telephone), Cramer's V coefficient could be used, whereby *min(r-1, c-1)=2* always equals Phi and Cohen's w values (see Cohen, 1988 for more information). This enabled comparability of coefficients, which would have been more challenging with larger contingency tables.

Secondly, due to the fairly low number of interval and ratio variables in the selected Life in Australia™ data (n=17), and as not all of them were normally distributed, non-parametric tests were used for ordinal and continuous substantive survey items and *survey mode* as a binary variable (0=online, 1=telephone). This was considered an acceptable adjustment since the Rank-Biserial Correlation measure is based on the Mann-Whitney U test, and the literature indicates that this test is only 5% less effective than a t-test even when the assumption of normality holds (Lehmann, 2004, p. 176).

Data Processing and Effect Size Analysis

The data processing and effect size analysis was performed according to the following steps:

- Selection of all substantive survey items in the Life in Australia™ data (six surveys), excluding: (1) those with less than 20% valid responses (to avoid statistical power issues with small samples of offliners), (2) primary socio-demographics which were not asked in each wave but added to the data from the Life in Australia™ profile dataset, (3) open-ended question items, (4) paradata variables. A total of 368 items were selected;
- Coding of variables, adding information on: broad item topic, type of question content, variable type, and number of variable categories as predictor variables;
- Calculation of raking weights for each of the six Life in Australia™ surveys, for onliners and offliners separately (to balance the samples on key socio-demo-

graphics) using the selected covariates – calibration was carried out to adjust the samples to match the 2016 Australian Census distribution by age, gender, education, state, country of birth (Australia, English-speaking background, non-English-speaking background), and telephone status (mobile, landline, dual user);

- Calculation of Cramer's V (Cohen, 1988) and Rank-Biserial Correlation coefficient (Glass, 1965) for each Life in Australia™ substantive survey item in a pair with *survey mode* (weighted data and unweighted data);
- Creation of a new data matrix with Life in Australia™ survey items as cases (rows), and effect size measures (dependent) and coded survey item information (predictors) as variables (columns);
- Construction of multiple linear regression models with *Cramer's V value* and *Rank-Biserial Correlation coefficient* (weighted and unweighted, a total of four models) as response variables, and *broad item topic*, *question content*, and *variable type* as regressors;
- Testing for all assumptions of ordinary least squares (OLS) regression and adjustment of the models according to the assumption test results (see outlier detection analysis below).

Outlier Detection Analysis

Outlier detection analysis identified a number of outliers affecting the normality of the residuals. Thus, a few units/cases (i.e., survey items) were removed based on the following criteria for outlier detection: standardized residuals (as discrepancy measures), leverage (as a distance measure), Cook's distance and DFBETA (as influence measures). We identified a limited number of survey items which stood out with extreme values for most of the outlier detection measures.

In the end, nine outliers out of 351 nominal or ordinal variables were removed from the Cramer's V models and nine outliers out of 202 ordinal or continuous were removed from R-BS coefficient models. It was observed that a number of outliers in the Cramer's V models were *internet* broad topic survey items and removing them decreased the clearly inflated Adjusted R-Squared coefficients from 0.445 to 0.349 (weighted) and 0.375 to 0.286 (unweighted), respectively. At the same time, the Root Mean Square Errors, as an absolute measure of fit, decreased significantly after removing outliers, which indicates a better absolute fit for both models.

While a number of *internet* topic survey items were identified as outliers and removed from the model, the remaining ones were intentionally left in the model to compare the magnitude of differences between *internet* and other topics. In the models with R-BS coefficient values as dependent variables, Adjusted R-Squared increased and Root Mean Square Errors decreased after removing outliers, which meant a better absolute and relative fit in those regression models.

Table 5 Benchmarking data sources and nationally representative benchmarks

Study	Data collection mode	Sample size	Benchmark
National Health Survey 2014-15	F2F	n=19,259 (18+ years old n=14,561)	Psychological distress (Kessler 6) General Health Private health insurance Wage and salary income
General Social Survey 2014	F2F	n= 12,932 (18+ years old n=12,348)	Life satisfaction
National Drug Strategy Household Survey 2013	self-administered paper based	n= 23,855 (18+ years old n=22,696)	Daily smoker Alcoholic drink of any kind in the past 12 months Household status (couple with dependent children)
Australian Census 2016	self-administered online, F2F	n= 23,401,892 people (18+ yrs old n= 18,193,864; private dwellings n=9,901,496)	Australian citizenship Employment status Home ownership with a mortgage Indigenous status Language other than English Living at last address 5 years ago Most disadvantaged quintile for area-based SES Resident of a major city Voluntary work
Australian Electoral Commission, 2015 data	administrative data	n= 16,405,465 Australians eligible to enrol	Enrolled to vote