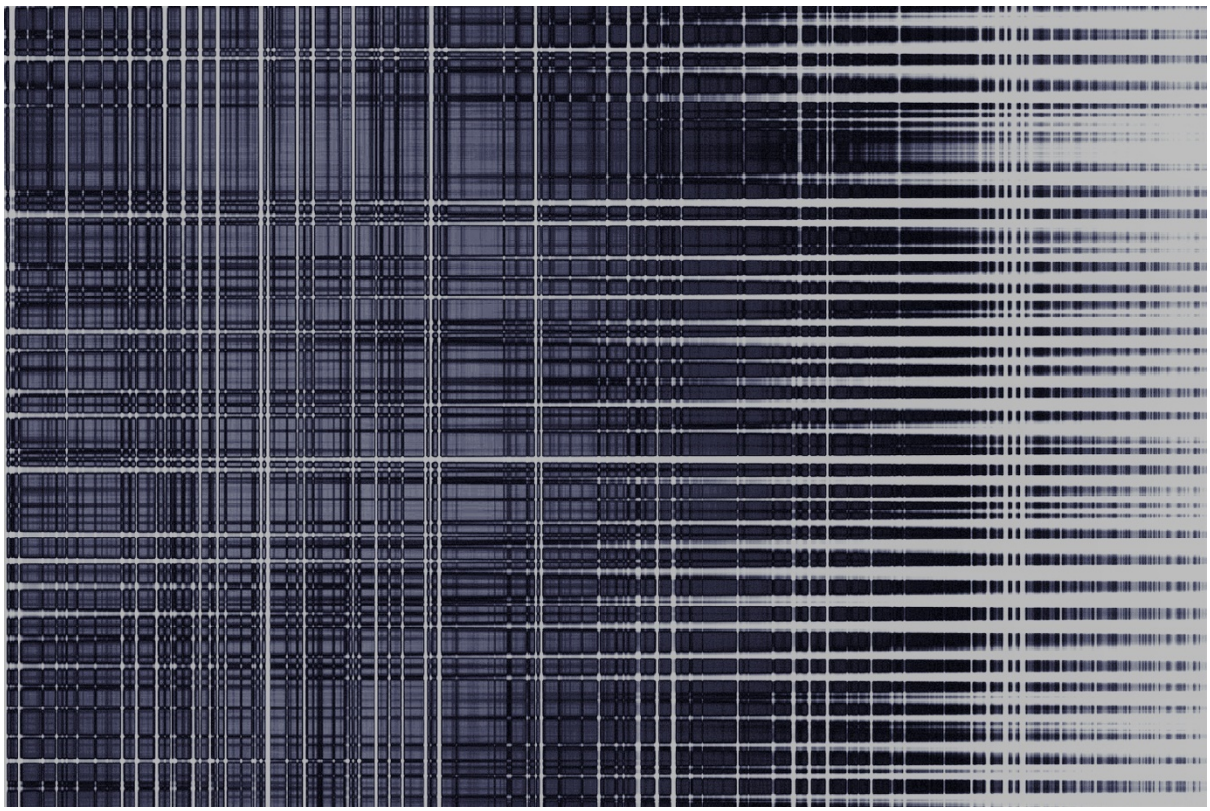
Australian National University

CENTRE FOR SOCIAL RESEARCH & METHODS

Social Research Centre

A SUBSIDIARY OF THE AUSTRALIAN NATIONAL UNIVERSITY



The Effects of Mode on Answers in Probability-Based Mixed-Mode Online Panel Research: Evidence and Matching Methods for Controlling Self-Selection Effect in a Quasi-Experimental Design

S Kocar, N Biddle, and B Phillips

CSRM & SRC METHODS PAPER
NO. 1/2021

The Effects of Mode on Answers in Probability-Based Mixed-Mode Online Panel Research: Evidence and Matching Methods for Controlling Self-Selection Effect in a Quasi-Experimental Design

S Kocar, N Biddle, and B Phillips

Sebastian Kocar is a PhD candidate at the ANU Centre for Social Research & Methods.

Nicholas Biddle is Professor of Economics and Public Policy, Associate Director of the ANU Centre for Social Research & Methods, and Fellow of the Tax and Transfer Policy Institute.

Benjamin Phillips is Chief Survey Methodologist at the Social Research Centre and Campus Visitor at the ANU Centre for Social Research & Methods.

Abstract

Online probability-based panels often apply two or more data collection modes to cover both online and offline populations. They do so with the aim of obtaining results that are more representative of the population of interest, in most cases the general population, than Web mode only. This study investigates mode effects in two different surveys – a probability-based quasi-experimental web-push survey and a probability-based online panel study. For both surveys the same questionnaire including items with nationally representative benchmarks was used. The aim of this study is to identify differences in answers in three different modes, the online and two offline (mail and telephone) modes, to establish the degree of measurement errors due to mixing modes and to present evidence about the most suitable combination of online-offline modes in online panel research from a measurement error perspective.

In this paper, we provide evidence that mail mode is most associated with satisficing – it generates slightly more item nonresponse and non-differentiation, and limited primacy after reducing self-selection bias with matching. We also identified some recency and extreme category responding in telephone surveys, potential social desirability associated with interviewer-administered telephone mode, and indication of the presence of question format effect. After controlling for self-selection to mode using different matching solutions, there were fewer differences which we initially assigned to self-selection, but we could not reduce all bias. With exact and coarsened exact matching, we could reduce slightly more bias, and also identify mode effects which were initially revoked by that self-selection of mode. Propensity scores matching proved to decrease self-selection bias, but it also decreased the ability to identify mode effects. Mahalanobis distance matching was not as successful in reducing bias, but it also did not negatively affect post-matching measurement effect assessment.

Due to the potential presence of self-selection effects in a quasi-experimental design, we also tested five different approaches to control for the absence of random assignment of respondents to modes: using socio-demographic controls in regression models (no matching), propensity score matching, Mahalanobis distance matching, exact matching, and coarsened exact matching. We compared the results on mode-effects after controlling for self-selection with those techniques.

It would appear that mode self-selection effects or sample composition effect were a more significant source of differences in the distributions of response variables than measurement mode effects such as satisficing and social desirability or question format effects. That is why we encourage all researchers studying or adjusting for mode effects in a quasi-experimental design and without access to similar single-mode data to use a matching method, preferably (coarsened) exact matching, to reduce self-selection bias.

Keywords: probability-based online panels, mixed-mode data collection, mode effect, mode self-selection effect, matching methods, propensity score matching, Mahalanobis distance matching, exact matching, coarsened exact matching, survey design

Acronyms and abbreviations

ABS	Address-Based Sampling
ANU	Australian National University
CATI	Computer-assisted telephone interviewing
CAWI	Computer-assisted Web interviewing
CEM	Coarsened exact matching
CSRM	Centre for Social Research and Methods (ANU)
D-AEMR	Dynamic Almost-Exact Matching with Replacement
DFRDD	dual-frame random digit dialling
EM	Exact matching
FDR	false discovery rates
G-NAF	Geocoded National Address File
LISS	Longitudinal Internet studies for the Social Sciences
MDM	Mahalanobis distance matching
MM	mixed mode
OLS	ordinary least squares (models)
OPBS	Online Panels Benchmarking Study 2015
PAPI	paper self-administered mode
PSM	Propensity score matching
SEM	structural equation modelling
SMD	standardised mean difference
SRC	Social Research Centre

1 Introduction

1.1 Mixing modes in online panel research and mode effects

Mixed-mode survey research is becoming increasingly common, and the use of email and in particular web-based surveys offers a range of opportunities for mixing modes of data collection (Bryman 2016, 232). There are many reasons for employing mixed modes, such as to maximise response and reduce costs in both cross-sectional and longitudinal studies (Groves et al. 2009, 175). In probability-based online panel research, panel organisations often apply two or more data collection modes to cover both online and offline populations. However, any differences in modes may result in measurement errors or item nonresponse as types of mode effects (Lavrakas 2008). That is one of the reasons why certain panel organisations use a uni-mode approach (e.g., providing tablets) or do not cover the offline population to achieve maximum measurement equivalence (for an overview of different practices see Kaczmirek et al. 2019, 4–5).

Differences in modes and mode effects are associated with different aspects of surveys, from sampling, coverage, unit nonresponse, item nonresponse and measurement error (Lavrakas 2008). Completely excluding the offline population (those who are unable to or unwilling to complete surveys online) may result in coverage error – and minimising survey error across various sources is key. For example, in Australia, approximately 14% of households were without access to the internet at home in 2016/2017 (Australian Bureau of Statistics 2018a[1]), but in 2019 there were reports of 91% of adult Australians having access to the internet on their mobile phones (Australian Communications and Media Authority 2020). Internet-only samples are not representative of the general population, since there are significant differences in demographic characteristics of the online compared to the offline population in Australia in terms of age, location and remoteness, gender, household income, employment status, highest qualification and country of birth (De Vaus & de Vaus 2013, 76–77). In addition, not every person with an internet connection at home will have the skills or inclination to participate online (Perrin & Bertoni 2017). Some authors therefore argue that an offline survey mode should be included or at least considered in probability-based panel research (Kaczmirek et al. 2019). To mitigate coverage error, i.e., to avoid undercoverage or completely excluding particular socio-demographic subgroups with a higher propensity to be offline, a mixed-mode approach should be carried out.

Modes of data collection as sources of survey errors at the level of the survey question differ in several ways: the medium in which questions are presented; who administers the questions and records answers; whether the questions (and supporting information) are presented aurally or visually; and the mode of responding (Tourangeau et al. 2000, Dillman et al. 2014). Also, differences in question format as a result of adjustments in different modes can add to the net effect of data collection mode (De Leeuw et al. 2011), as well as question or answer order effects. In the Total Survey Error Framework, measurement mode effects are measurement errors in the survey process (Groves et al. 2009), i.e., a departure of the measurement from the true value. When data are collected from different groups of respondents using different survey modes, mode effects and differential measurement error in particular may threaten the validity of results (De Leeuw et al. 2011).

[1] The Australian Bureau of Statistics has ceased undertaking the collection on internet activity in 2018 (Australian Bureau of Statistics 2018b).

Mode-related measurement errors are present in different ways, such as acquiescence response bias, social desirability and satisficing (Lavrakas 2008; Groves et al. 2009, 224). Social desirability refers to respondents providing answers to put themselves in good light with the interviewer whereas acquiescence response bias refers to a tendency to agree rather than disagree. Both are often associated with interviewer's presence (Dillman et al. 2014). Moreover, intrusive questions or the perceived risk of identification of the respondents can lead to unit or item nonresponse, especially in interviewer-administered modes (Tourangeau et al. 2000). Generally speaking, self-administration has a higher potential for satisficing than interviewer-administration, including in mixed-mode probability-based online panels (Baker et al. 2010). Satisficing as a source of measurement error is related to the cognitive effort required for generating respondents' answers to survey questions. The meaning of each question has to be carefully interpreted, respondents' memories extensively searched for information, that information integrated into judgements, and those judgements communicated clearly and precisely (Krosnick et al. 1996). However, some respondents are likely to make the task of responding to survey questions as easy as they can and this leads to taking shortcuts such as using ranges or rounding values (numeric answers), to making ratings following a few simple principles (scales) or bypassing serious consideration of questions (Tourangeau et al. 2000, 254). It can result in item nonresponse, non-differentiation (tendency to provide the same answer to all questions in a block), acquiescence response bias (tendency to agree with the interviewer), non-substantive responses (e.g., don't know and refusal to answer), rapid completion (speeding), primacy and recency effects (Krosnick et al. 1996; Baker et al. 2010). The direction of biases from these mode effects are more difficult to predict *a priori*.

1.2 Existing literature on mode effects in online panels

While there has been substantial research exploring mode effects of more traditional survey modes, there has been little research exploring those effects in online panels. The following studies give some insight into effects of mode in online panels, both probability-based ones (Knowledge Networks, Longitudinal Internet studies for the Social Sciences (LISS)) and a non-probability online panel (Harris Interactive online panel), while not all of them randomly assigned respondents to modes.

Dennis et al. (2005) conducted a study on mode effects in probability-based online panel surveys, controlling for sample origins. Regarding the differences between samples, they noted that the differences in answers might be attributed to data collection modes when they are, in fact, a result of differences in the representativeness of the samples. Sample composition differences, as well as panel conditioning and panel attrition in online panel research, might contribute to the differences in survey responses observed for the different modes of collection. After controlling for demographic characteristics and panellists' survey experience, the observed mode effects were significant for several survey items. The reason for that might have been a tendency to select positive responses (on a scale) in telephone interviews, as well as the visual-aural differences (e.g., a feeling thermometer displayed online).

Duffy et al. (2005) carried out a similar study, but they aimed to identify the relative impact of sample and mode effects in online panel (volunteer/opt-in) and face-to-face surveys. They concluded that there were two competing effects when comparing online and face-to-face data collection. Online panels attracted more knowledgeable and viewpoint-oriented respondents on the one hand, whereas face-to-face techniques produced greater social desirability effects. While sometimes those two effects appear to balance, sometimes they did not.

De Leeuw et al. (2019) investigated measurement error in probability-based online panels, i.e., the relationship between mode effects and question format effects. In contrast to the other two studies on mode effects in online panels, respondents in the LISS panel were randomly assigned to online and telephone modes. While there was little evidence of interaction effects between mode and question format, they found small but consistent question format effects and mode effects, namely reliability, acquiescence response bias and choosing extreme response categories. Telephone mode respondents provided less consistent responses, showed a greater tendency to acquiesce, and more often chose extreme response categories than online mode respondents.

1.3 Methodology for assessment of mode effects

Mode effects can be divided into three components: coverage mode effects, nonresponse mode effects (both selection effects), and measurement mode effects (Buelens et al. 2012; Schouten et al. 2013). The most prevalent approaches to studying mode effects have been testing for differences in data quality indicators, such as those for data completeness (e.g., item nonresponse rate), response accuracy (e.g., in comparison to benchmarks), and reliability (e.g., scaling properties), as well as testing for differences in response distributions of survey items (De Leeuw & Zouwen 1988). Jäckle et al. (2010) reported that, in practice, mode effects are commonly tested using a variety of statistical tests: t-tests or chi-square test with weighted data, binomial, ordinal and multinomial logistic regressions, partial proportional and proportional odds models, ordinary least squares models (OLS), or structural equation modelling (SEM).

On the other hand, the study of mode effects does not come without challenges: sample compositions might differ between modes due to differential nonresponse, differences in responses might impact only certain estimates, and identifying the net mode effects requires careful experimental designs (Jäckle et al. 2008). To successfully identify the net mode effects, various aspects would have to be controlled for, including the difference in coverage of the mode and the differences in mode preferences (Buelens et al. 2012; Schouten et al. 2013). In quasi-experimental survey designs, in which respondents are not randomly assigned to modes, there are two parallel and possibly competing sources of differences in responses in different modes: mode effects and mode self-selection effects (Suzer-Gurtekin et al. 2018).

Vannieuwenhuyze and Loosvelt (2012) discussed three methods to disentangle measurement mode effects from selection effects. Mixed-mode (MM) Calibration generally tries to render both mode groups comparable on a set of variables (by weighting) assuming that the remaining differences are caused by measurement effects, while Extended MM comparison is based on comparing mixed-mode data with comparable single-mode data. On the other hand, Extended MM Calibration predicts the respondent mode group in the comparable single-mode data. Each one of those methods have notable disadvantages – MM Calibration is based on an unrealistic assumption (i.e., high difficulty of finding a set of mode-insensitive variables properly explaining self-selection effect), Extended MM Comparison can only compare two modes, and both Extended MM Comparison and Calibration require an availability of comparable single-mode data (Vannieuwenhuyze & Loosvelt 2013, 99-101).

1.4 Data matching methods

In observational studies where random assignment is absent, individuals ending up in different groups (called treated and control) may differ in terms of observed, unobserved, and unobservable

characteristics. The two groups may not therefore have the same outcome in the absence of treatment, and causal effects cannot be estimated without careful statistical controls (Rosenbaum 2020). To deal with the absence of random assignment, one quasi-experimental technique is matching, where the control group is made to look more like the treatment group across observed characteristics. Different matching methods such as propensity score matching, propensity score weighting, Mahalanobis distance matching or coarsened exact matching have been suggested in the literature (King & Nielsen 2019 Rosenbaum 2020).

In observational studies which by definition lack random assignment, individuals ending up in different groups may differ in terms of covariates, are not directly comparable, and it is challenging to estimate causal effects without multivariate matching (Rosenbaum 2020). To deal with the absence of random assignment, different methods for casual inference such as propensity score matching, propensity score weighting, doubly robust estimation (combining outcome regression and propensity scores), Mahalanobis distance matching or coarsened exact matching have been suggested in the literature (King & Nielsen 2018; Rosenbaum 2020; Funk et al. 2011)). More recently, more classic data matching methods listed above have been expanded to so-called machine learning methods and techniques such as random forest, Dynamic Almost-Exact Matching with Replacement (D-AEMR), genetic matching, or a combination of different classic or supervised learning methods (Sizemore & Alkurdi 2019).

Out of more classic distance models, propensity score matching might be the most popular method used in quasi-experimental studies and has been available for almost 40 years (Rosenbaum & Rubin 1983), but not without any sceptics who are more in favour of alternative methods, such as Mahalanobis distance matching or coarsened exact matching (King & Nielsen 2018). Besides matching methods based on modelling, data can be matched with so-called stratification, namely with exact matching and coarsened exact matching (Sizemore & Alkurdi 2019). Exact matching (EM) is a statistical technique for matching on discrete metric with a meaningful set of predictors, and is rarely feasible in real data sets as it can result in an empty set in a multivariate setting with a number of continuous covariates (King & Nielsen 2018). Coarsened exact matching is a Monotonic Imbalance Bounding matching method which coarsens each variable, i.e., recoding each continuous variable by grouping substantively indistinguishable values into the same value, and later applies exact matching to the coarsened data (Lacus et al. 2012).

1.5 Aim of the study

With this background in mind, the general objective of this paper is to study the severity of mode effects in probability-based online panel research, while comparing methods for improving causal inference in the study of mode effects with quasi experimental design. There has been limited research on measurement mode effects in probability-based online panels, while they are different from many other mixed-mode surveys for the following reasons (but not limited to): (1) the ability to measure change in time over short time intervals, (2) a possibility of switching modes for reasons such as to minimise nonresponse or to accommodate respondents' preferences regarding privacy, (3) a possibility of using a uni-mode approach (e.g., for rapid data collection, self-administration only to collect data on very sensitive topics). These differences warrant investigating measurement mode effects in probability-based online panels in more detail.

Using data from two studies using the same questionnaire and collected by the same social research organisation using a mix of modes, we would particularly like to answer the following research questions:

1. *How significant are differences in distributions of response variables commonly explained by satisficing and associated with the mode of data collection?*

By answering this question, we would like to establish how severe measurement mode effects related to satisficing can be in online panel research combining the online mode with an offline mode. We will test for differences in data quality indicators and for differences in response distributions of survey items, as suggested by Jäckle et al. (2008). We will compare satisficing-driven mode effects between the online mode and two offline modes and discuss different mixed-mode designs from the measurement mode effects perspective.

2. *How significant are differences in distributions of response variables commonly explained by social desirability and associated with the mode of data collection?*

In addition to studying mode effects related to satisficing, we would like to determine the severity of measurement mode effects related to social desirability in probability-based online panel mixed-mode research design. Again, we will discuss mixing modes from the measurement mode effects perspective.

3. *To what extent can self-selection effect in a quasi-experimental design be controlled with different matching methods to help identify mode effects-related differences in distributions of response variables?*

We will also test a new combination of approaches, including data matching, to investigate different solutions in studying measurement mode effects in a quasi-experimental design. The aim of this study and the contribution of this paper is not only establishing the extent and severity of measurement mode effects in online panel research, but also identifying methods and techniques offering practical solutions to studying mode effects in mixed-mode approaches not allowing for random assignment of participants to survey modes. In particular, we would like to present evidence on controlling for mode self-selection effect with data matching, and the success in disentangling measurement mode effects from coverage and nonresponse mode effects. In that case, our findings on mode effects could be more in line with the literature on measurement survey errors in mixed-mode survey research.

2 Methods

2.1 Data

We will analyse the Life in Australia™ Wave 2 (2017) and Online Panels Benchmarking Study 2015 (OPBS) (Pennay et al. 2016b) unit record files. The data collection was conducted by the Social Research Centre (SRC) with the support of the ANU Centre for Social Research and Methods. The OPBS data were primarily collected for a benchmarking study by the SRC (probability-based sampling) and via five opt-in online panels (nonprobability-based sampling). We will only use the probability component of the study. The findings of OPBS also provided the grounds for the introduction of a national probability-based online panel in Australia (Life in Australia™; Kaczmirek et al. 2019). We will also analyse the Wave 2 survey data, which used the same questionnaire as for OPBS. Although studying mode effects in an online panel setting is the primary focus of this paper, the OPBS data is

included to: (1) increase subsample sizes, (2) investigate mode effects in paper self-administered mode (PAPI), (3) include mail mode as another (control) self-administered mode in relation to the telephone mode, (4) extend the findings from probability-based online panel research to the other types of mixed-mode research, including web-push surveys.

2.2 Samples, subsamples and data collection modes

The OPBS study comprised three probability-based samples of the Australian population aged 18 years and above. The Wave 2 study comprised a mixed mode probability-based sample. Data collection was carried out between October and December 2015 (OPBS) and in January 2017 (Wave 2). The OPBS surveys used the following designs (Pennay et al. 2016a):

1. Address-Based Sampling (ABS) using the Geocoded National Address File (G-NAF) sampling frame. The G-NAF is the authoritative list of Australian addresses, with more than 13 million physical address records including geocodes (Australian Government 2020) (online, telephone, mail modes).
2. Standalone dual-frame random digit dialling (DFRDD) CATI Survey.
3. Recruitment at the end of an established DFRDD survey/piggyback recruitment (online, telephone, mail modes).

The ABS survey and the survey using piggyback recruitment allowed mixed modes of completion. For the ABS survey, sample members were initially approached by mail, but some responded online (39%) or to outbound telephone reminders (24%). The piggyback-recruited respondents mostly responded online (52%) or via phone (41%) (Pennay et al. 2018).

Response rates (AAPOR RR3) were 12.4% for the DFRDD piggyback sample, 17.9% for the standalone DFRDD sample and 26.5% for the ABS sample. Cumulative response rate (CUMRR2, see Callegaro & DiSogra 2008) as a product of recruitment, profile, retention and completion rates was 12.0% for Wave 2. About 14% of Wave 2 panellists responded by telephone.

To isolate the mode of data collection while controlling for sample origins and socio-demographic characteristics, variables for mode and origins were derived. The samples are then uniquely defined as presented in Table 1.

Table 1 Original subsamples

Data source	Sampling frame (mode)	Sample origin	Mode
Probability-based online panel Life in Australia™ (Wave 2)	DFRDD/panel (CAWI) (n=2,166)	1	1
	DFRDD/panel (CATI) (n=414)	1	2
Online Panels Benchmarking Study 2015 (OPBS)	ABS/standalone (CAWI) (n=208)	2	1
	ABS/standalone (CATI) (n=128)	2	2
	ABS/standalone (PAPI) (n=202)	2	3
	DFRDD/standalone (CATI) (n=601)	3	2
	DFRDD/piggybacked (CAWI) (n=292)	4	1
	DFRDD/piggybacked (CATI) (n=228)	4	2
	DFRDD/piggybacked (PAPI) (n=40)	4	3

CATI = Computer-assisted telephone interviewing; CAWI = Computer-assisted Web; DFRDD = dual-frame random digit dialling

While those samples differ based on the sampling approach applied (Sample origin in Table 1), we combined samples based on the mode used (Mode in Table 1), since all of the surveys were probability-based. For example, to identify mode effects, CATI mode respondents from ABS, standalone DFRDD CATI, DFRDD piggybacking, and DFRDD-recruited panel (Wave 2) samples will be compared to the PAPI respondents from ABS and DFRDD piggybacking. For more detail, see Table 2.

Table 2 Subsamples by mode combined for this data matching and mode effects analysis

Sample by mode	Source	Sample origin
CAWI (n=2,666)	Wave 2 (n=2,166)	1
	OPBS ABS (n=208)	2
	OPBS DFRDD/piggybacked (n=292)	4
CATI (n=1,371)	Wave 2 (n=414)	1
	OPBS ABS (n=128)	2
	OPBS DFRDD/standalone (n=601)	3
	OPBS DFRDD/piggybacked (n=228)	4
PAPI (n=242)	OPBS ABS (n=202)	2
	OPBS DFRDD/piggybacked (n=40)	4

ABS = Address-Based Sampling; CATI = Computer-assisted telephone interviewing; CAWI = Computer-assisted Web; DFRDD = dual-frame random digit dialling; OPBS = Online Panels Benchmarking Study 2015; PAPI = paper self-administered mode

By combining two studies and four samples into three targeted subsamples[2], we increased the statistical power as well as enable mode effects analysis for three distinctive modes.

[2] Potential temporal effects due to the time gap in data collection, as well as sample composition effect (in OPBS studies), were controlled by including *sample source* variable (see samples in Table 2) as a predictor/control in all regression models (for more information, see subsection 2.4 Data Analysis and Table 6 in the Appendix).

2.3 Data matching methods

We will also test five different methods to control for the absence of non-random assignment of respondents to modes, four of them being data matching methods. In practice, this is not uncommon as the literature recommends reporting results based on multiple matching methods since the conclusions might be very sensitive to matching algorithm choices (Leite 2016). The matching methods were chosen based on reviews of King and Nielsen (2016) and Sizemore and Alkurdi (2019). In this study, we used arguably the most traditional matching methods for casual inference due to a high availability of information in the literature on how to apply those methods in practice.

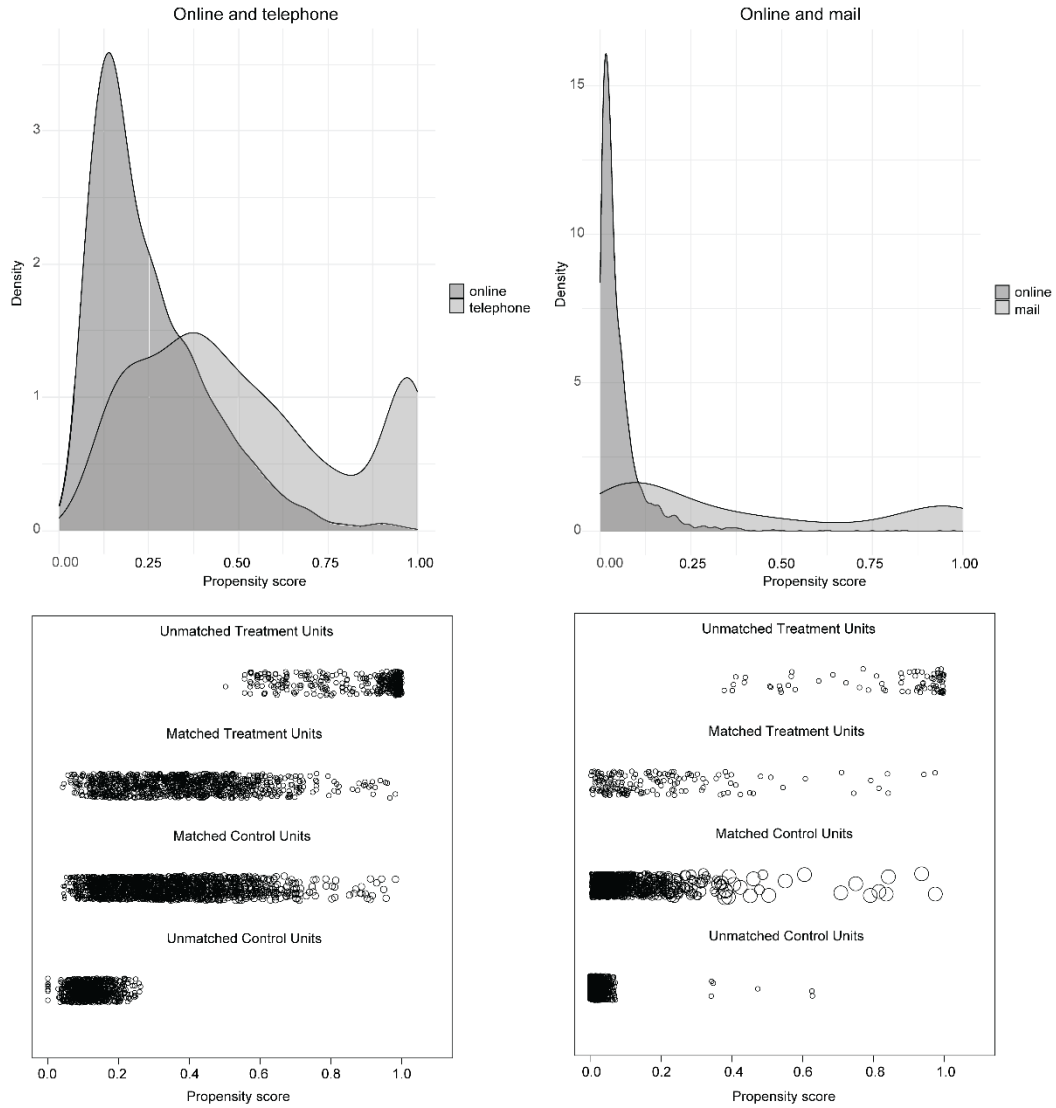
1. Socio-demographic controls in regression models without matching

This approach is similar to poststratification weighting and MM Calibration (see Vannieuwenhuyze & Loosvelt 2012), and it was conducted to identify differences in distributions of response variables as a result of both measurement mode effects and mode selection effects due to non-random assignment of respondents to modes (see Suzer-Gurtekin et al. 2018). With this approach, we also aimed to identify items which should and should not be used as covariates in data matching in the next steps, since the differences in distributions, consistent with the literature, might be pure measurement mode effects and not self-selection bias. Excluding so-called outcome variables is a standard approach in data matching (King & Nielsen 2018).

2. Propensity score matching (PSM)

We decided for a greedy approach based on propensity score, which were calculated manually using R. In addition to the variables selected for matching, there are two parameters that should be selected to control for the size of the final samples and their imbalance (MatchIt package): (1) the maximum allowed distance between matched units (caliper), and (2) the maximum number of units that could be matched to one unit from the other sample (ratio). We carefully investigated different combinations of parameters (caliper 0.05-0.25, ratio 1-5:1 [CAWI-CATI], 10-20:1 [CAWI-PAPI]) and reviewed: standardised mean difference (SMD) as a measure of imbalance after matching, the variability of weights (not to overly inflate variance of estimates), and the final sample sizes (to keep enough statistical power). The following parameters seemed to represent the most optimal solution for our cases: caliper 0.05, ratio 3:1 (CAWI-CATI), and caliper 0.05, ratio 15:1 (CAWI-PAPI). The selected ratio for PAPI mode was much greater since the initial sample of PAPI respondents was much smaller than both CAWI and CATI samples (see Table 2). The matching results in Figure 1 show that there was a notable overlap of propensity scores between the modes, but quite some imbalance in distributions. This was even more apparent for the CAWI-CATI samples than the CAWI-PAPI samples. In the end, the vast majority of cases, which could not be matched, were pruned to improve balance and not because they were off the region of common support (i.e., the area where the densities of the estimated propensity scores for treatment and control groups overlap).

Figure 1 Initial distribution of propensity scores (histogram), propensity scores before and after matching for two pairs of samples (visual presentation of PSM solutions)



3. Mahalanobis distance matching (MDM)

To perform Mahalanobis distance matching, we used the MatchingFrontier R package developed by King et al. (2016). The problem of most matching methods is that they are designed to maximize one metric, such as Mahalanobis distance, but are judged against a different metric, such as standardised mean difference. While using Mahalanobis distance as a distance metric and Average Mahalanobis Imbalance as the imbalance metric, the software calculated optimal matching solutions for each possible sample size, constituting a frontier (King et al. 2016). In the end, we selected the subsample by pruning the same numbers of units as with PSM for comparability purposes. Technically speaking, this approach can be considered a hybrid between a classic distance model and machine learning, as there is some degree of algorithmic optimisation of individual matches (Sizemore & Alkurdi 2019).

4. Exact matching (EM)

The most notable issue with exact matching is the algorithm returning an empty set in a multivariate setting with a number of continuous covariates. Therefore, we carefully reviewed the differences

between the samples and selected the best predictors of group membership. Out of all covariates selected for matching (coloured blue in Table 6), only c2, the number of household members, is continuous. To control for the size of the final samples and their imbalance, one can make a decision on the number of covariates and the number of their categories by collapsing their values. The more covariates or their categories, the smaller the matched samples and lower imbalance, but also a decreased statistical power. With the eight selected covariates, we pruned a fairly comparable number of cases to PSM and Mahalanobis distance matched samples.

5. Coarsened exact matching (CEM)

We performed automated CEM with the same eight selected covariates as for exact matching for comparability. In contrast to EM, CEM coarsens each continuous variable by recoding it into homogeneous groups with very similar values grouped together, which prevents too many units with no perfect match to be pruned, something that could happen with EM (Iacus et al. 2012). We assumed that in order to observe notable differences between the methods, a decent proportion of covariates would have to be continuous, which was not the case in our study.

Table 3 summarises the data matching approaches and results. We purposely tried to optimise the data matching solutions while keeping the matched samples of fairly similar sizes for comparative purposes. The propensity score matching was carried out first, and the sample size of the most optimal PSM matching solution (also based on SMD and the variability of weights) was the reference sample size (about 72% online-telephone and 56% online-mail) for MDM, EM and CEM methods. By introducing this case pruning consistency across different matching methods, the loss of statistical power did not affect our conclusions on the adequacy of different matching methods in measurement mode effect analysis.

Table 3 Data matching parameters and sample sizes

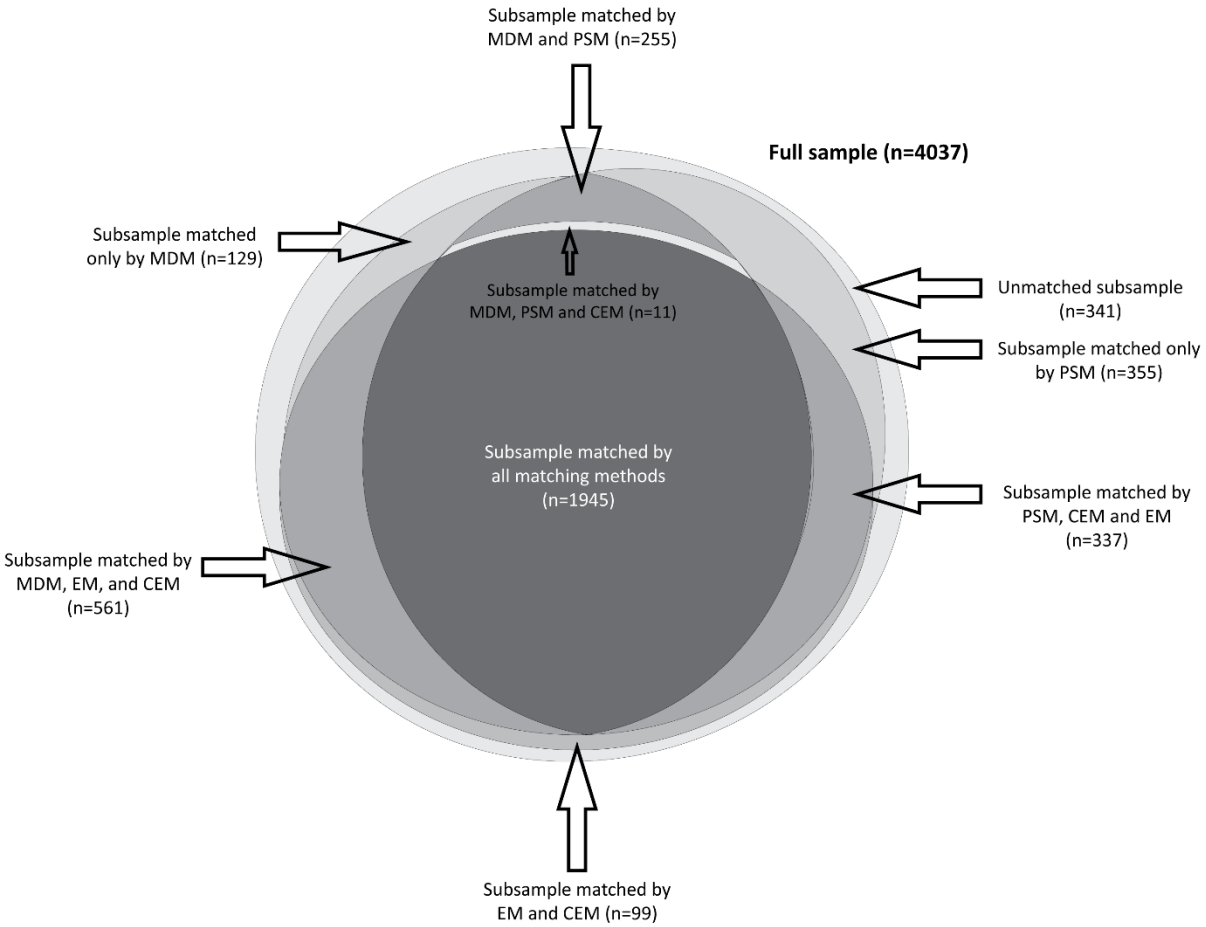
Matching method *(R package used)*	**Online-telephone samples matching**				**Online-mail samples matching**			
	Original sample size	**Matching parameters**	**Matched sample size**	**Average SMD[3]**	**Original sample size**	**Matching parameters**	**Matched sample size**	**Average SMD**
Propensity score matching *(MatchIt)*	4,037 (2,666 CAWI + 1,371 CATI) *Average SMD = 0.223*	ratio=3 caliper=0.05	2,904 (1,913+991)	0.055	2,908 (2,666 CAWI +242 CATI) *Average SMD = 0.355*	ratio=15:1 caliper=0.05	1,617 (1,453+164)	0.141
Mahalanobis distance matching *(MatchingFrontier)*		minimizing Average Mahalanobis Imbalance	2,904 (2,074+831)	0.135		minimizing Average Mahalanobis Imbalance	1,617 (1,499+118)	0.227
Exact matching *(MatchIt)*		8 matching variables	2,942 (2,045+897)	0.125		8 matching variables	1,666 (1,522+144)	0.226
Coarsened exact matching *(CEM)*		8 matching variables, 1 of them coarsened	2,957 (2,056+901)	0.125		8 matching variables, 1 of them coarsened	1,680 (1,531+149)	0.228

CATI = Computer-assisted telephone interviewing; CAWI = Computer-assisted Web; SMD = standardised mean difference

[3] All covariates not affected by measurement mode effects described in the literature on measurement error, were included in the calculation of average SMD in Stata 15. The selection criteria for these covariates were the same as for choosing covariates for PSM logit model (see subsection 2.6).

There are three findings worth mentioning. First, MDM method with a calculation of frontiers and minimisation of Average Mahalanobis Imbalance for a sample of particular size ended up including a higher proportion of online respondents for both mixed-mode approaches in comparison to the other three methods. Second, while the sample balance estimated with average standardised mean difference (SMD) was better for CAWI-CATI than CAWI-PAPI both before and after matching, PSM stood out as the method improving balance substantially better than MDM, EM and CEM. This could be explained by the fact that we were able to include all covariates (n=30) in PSM logit models, but not with the other three methods. For that reason, SMD might be a biased indicator of the quality of data matching when comparing these methods, and the results on measurement mode effects after matching should be the most suitable quality evaluation approach in our particular case. Third, all methods using the same parameters and/or ranges of matching variables pruned significantly more cases when matching CAWI-PAPI samples (more than 4 out of 10 cases) than when matching CAWI-CATI samples (less than 3 out of 10 cases). This indicates that the respondents who preferred to respond via PAPI were more different to online respondents than those in favour of CATI, which is apparent from Figure 1 as well.

Figure 2 Matching results for the online-telephone sample



In Figure 2 we present the results of data matching for online-telephone samples, i.e., the number of cases that were included in matched samples by various methods. There were 1,945 cases matched by all four different methods, which is about 48% of the whole sample of online and telephone respondents. Moreover, there is more than 99% overlap between the samples matched by EM and CEM. Out of all four methods, PSM stands out as the method with the highest number of cases not matched by any other method (n=355) while 561 cases were matched by the other three methods excluding PSM. With only 341 cases out of 4,037 not being matched by any of the methods, and a significant proportion of the sample being matched by one or two out of four methods, we could expect quite different results when studying mode effects in online panels.

2.4 Data analysis

Since the calculation of propensity scores requires complete data and we do not want to exclude cases with a small number of missing values, we will use random forest imputations, suitable for categorical data and provided by R package *missforest*. Data with imputed values will be used for matching purposes only, but not when testing for differences in distributions to identify item-level mode effects.

While the matching part of the analysis was done in R using the packages listed in Table 3, Stata 15 statistical software was used to carry out the statistical analysis to investigate mode effects. After matching samples, regression analysis was conducted separately for CAWI-CATI and CAWI-PAPI matched samples.

To identify any distributional differences between data collection modes, we carried out multivariate analysis, i.e., binomial logistic regression for binary response variables, ordinal logistic regression for mostly scalar variables, multinomial logistic regression for nominal response variables with more than two levels, and OLS for continuous variables, as suggested by Jäckle et al. (2010). To study different types of mode effects, in a limited number of cases we carried different regression modelling for the same variables, e.g., multinomial regression (primacy, recency) and ordinal regression (social desirability) for frequency and typical amount of alcohol consumption. To study item nonresponse and non-differentiation, we used full (non-matched) samples, while conducting bivariate tests – chi-square test and t-test for independent samples.

As the differences in distributions will be tested for a number of items, that increases the probability of an observed difference between groups being attributed by chance, which is indicated by a p-value measure (see Johnson (2013) for p-value selection review). To avoid reporting false positives and rejecting true null hypotheses, we examined the distribution of p-values, so called p-curves. With the assumption that every p-value is equally likely to be observed if the studied phenomenon is zero, we concluded that about one-third of all p values between 0.01 and 0.05 and about 7% for p <= 0.01 would be attributed by chance and not to real self-selection or measurement mode effects (see Head et al. 2015 for more information). To reduce false discovery rate and avoid Type I errors, we decided to run the Benjamini-Hochberg procedure (Benjamini & Hochberg 1995) instead of reporting statistically significant results at commonly used p=0.01 and p=0.05 levels. We reported statistically significant regression coefficient at false discovery rates (FDR) of 0.05, 0.1, and 0.2. For example, FDR value of 0.1 means that 1 out of 10 discoveries would be false and, in our particular case with about 150 regression coefficients compared, FDR=0.1 transformed into p values between 0.005 (online-mail, CEM) and 0.047 (online-telephone, socio-demographic controls). These differences indicate that the conventional p level selection could result in biased hypothesis testing. We will be careful in interpreting results significant at FDR=0.2 level, marked with a dagger. We will review both changes

in statistical significance as well as coefficients, since data matching and thus sample size reduction can lead to a loss of statistical power.

2.5 Selection of items as outcome variables and controls in regression models

As Jäckle et al. (2010) stated, identifying mode effects in practice often requires all or most items in the survey being tested for differences. Therefore, we will test the majority of items in the merged OPBS/Life in Australia™ Wave 2 data file. The Health, Wellbeing and Technology Survey questionnaire used in both surveys for comparability purposes consisted of 36 questions. The questionnaire included 12 topical questions (substantive measures) and a number of demographic questions with six primary and 13 secondary demographics measures (Pennay et al. 2018, 6-7). Out of all available items in the unit record files, we purposely selected all topical items and eight of the most relevant and/or possibly sensitive demographic items (a total of 35 items).

There are a number of different dimensions of measurement mode effects in survey research. In this paper, we will focus on measurement dimensions related to satisficing and social desirability, such as primacy and recency (e.g., for life satisfaction measured with a 10-point scale), response non-differentiation (also called partial straightlining) and actual straightlining, as well as item nonresponse (for all items with a particular focus on potentially sensitive items, such as income). Although some authors such as Dillman et al. (2014, 404–415) suggest using the same question format and wording in all modes, this is often not possible with non-substantive answer options such as 'don't know' and 'cannot say', or when answers should be displayed in visual modes and not read in aural modes. Therefore, we also selected all items with any kind of format differences to study question format effect, although those differences might not be significant enough to identify any relevant effects due to the carefully prepared questionnaire design taking multi-mode data collection into account.

Moreover, the selection of omitted reference mode and the selection of base outcome dependent variable value should be explained. In regression analyses, either binary logistic, multinomial logistic and ordinal logistic analysis, the online mode was primarily chosen as the reference group for mode effects comparisons since the difference between online and offline modes is key in probability-based online panel research (see Table 6 in Appendix). In terms of the base outcome selection, while the model would report the same differences between groups no matter the dependent variable value selected, the interpretation of coefficients differs based on the selection. In our case, when there was no reason to believe that a specific category would be selected with a higher probability due to a measurement mode effect (e.g., the first listed category due to primacy effect or the last listed category due to recency effect), the category with the highest frequency/modal category was primarily selected as the base outcome. Otherwise, the most 'neutral', middle category, or the second most common answer was chosen as the reference group. Modal categories are often selected in practice since confidence intervals decrease with larger subsamples.

Lastly, in binomial, ordinal, multinomial logistic and OLS regression analyses, the same socio-demographic controls will be used as for weighting in the OPBS and Life in Australia™ Wave 2 studies, namely: age, gender, education, state, country of birth (Australia, English speaking background, non-English speaking background), and telephone status (mobile, landline, dual user) (Pennay et al. 2018).

2.6 Selection of items as controls in data matching

To identify the net effect of mode, careful experimental designs controlling for other characteristics of the samples are required (Jäckle et al. 2010). The ABS and DFRDD piggybacked surveys in OPBS applied a survey design in which the sampled respondents were not assigned to modes randomly but selected the prefered way of participating themselves. That closely resembled a real-world offline recruitment to a probability-based online panel with an alternate non-CAWI mode in which respondents choose whether to be surveyed in CAWI or a non-CAWI mode, as was the case in Life in Australia™. To eliminate the mode self-selection effect, we had to distinguish between variables with a higher propensity to be affected by measurement mode effects consistent with the literature (e.g., grid/matrix questions, sensitive questions, questions with longer lists of answers), and those affected by mode-self-selection effects (e.g., 'webographic' variables[4]), which only seem to be associated with the effects associated with presentation of questions or the type of survey administration. This is consistent with recommendations of King and Nielsen (2018).

To match the samples, a number of controls had to be selected in attempts to correct for any mode self-selection bias. The literature has provided some evidence on how online and offline respondents may differ in various behavioural, factual and attitudinal dimensions, such as financial and health-related indicators (Couper et al. 2007) or the use of technology (Duffy et al. 2005). A number of items related to those topics were included in the Health, Wellbeing and Technology Survey questionnaire.

For PSM, the literature generally suggests to select the majority of available items to match the respondents participating in different modes, apart from items that are clearly affected by measurement mode effects described in the literature (Dutwin & Buskirk 2017). We identified those items with the first approach to control for self-selection bias, i.e., using socio-demographic controls in regression models. To decrease the bias in comparing samples, we did not include variables subject to satisficing, social desirability, and other measurement bias. Matching on items affected by mode effects could decrease or even eliminate the potential to identify real mode-effects after controlling for self-selection effect. Instead, as Dutwin and Buskirk (2017) proposed, we also matched samples on so-called 'webographic' variables, besides socio-demographics used in the non-matching poststratification weighting approach. If the variables were both subject to mode effects, or if key variables distinguished the samples by mode as previously reported in the literature (e.g., early adopter items), we derived, where possible, total scores which should be less sensitive to mode effects. We selected the same range of variables for MDM and recoded categorical variables into sets of dummy variables as MDM is not suitable for categorical data matching. See Table 6 for more information – PSM and MDM variables are coloured green.

Instead, EM and CEM find matches based on covariates. If the numbers of selected covariates are high, this results in few successful matches, especially for exact matching method. Therefore, we preliminarily modelled differences between online and offline respondents (binary logistic regression), identified the regressors which distinguished the groups best, and used them in matching (see Table 6 in the Appendix for the final list). Running the same model with the selected regressors only, we noticed a very little decrease in pseudo-R^2 values compared to the full model.

[4] Webographic variables are items measuring behaviors and attitudes towards new products, new brands, deals and discounts (Dutwin and Buskirk 2017). DiSogra et al. (2011, 4505) call them 'early adopter' items, and early adopters are defined as "consumers who embrace new technology and products sooner than most others".

In this paper, we purposely tried to maximise the potential of each matching method to reduce self-selection bias, which is why we decided not to use the same range of covariates for matching methods based on fundamentally different principles.

3 Results

In this section, we will present the result of all analyses, separated into subsections by the mode effects study approaches: (1) using socio-demographic controls only (covariate approach), (2) PSM, (3) MDM, (4) EM, and (5) CEM.

3.1 Unit nonresponse, non-substantive answering, and non-differentiation

We studied item nonresponse, non-substantive answering, non-differentiation and straightlining without using any of the five proposed approaches dealing with the quasi-experimental design. Most importantly, the analysis was carried out before imputing missing values for matching purposes.

We observed some notable differences between modes, starting with non-differentiation (see Table 5 in the Appendix). PAPI mode has the highest percentage of respondents who selected the same category for all ordinal items of questions. About half of mail mode respondents varied their answers for both early adopter and Kessler 6 Psychological Distress Scale items, and 11.2% should be, based on our criteria, classified as straightliners. That is more than twice as much as for CATI and almost four times as much as for CAWI. The differences between telephone and online mode respondents originated in a higher propensity to non-differentiate to K6 questions (more questions, longer scale compared to early adopter) in the telephone mode.

The PAPI mode has the highest average across-all-items skipped questions (2.85%) and analytically missing values (2.89%, those also include non-substantive answers in our analysis). While the PAPI respondents have a higher propensity not to respond to a question, CATI and CAWI respondents have a tendency to not provide a substantive answer instead (e.g., responding with ‘don’t know’). That could be explained with question format effect, i.e., not offering non-substantive answers in particular modes. Furthermore, there are few differences between CAWI and CATI modes in the total propensity for missingness (see Table 4).

The differences between modes are even more significant for sensitive items. PAPI respondents have a consistently higher proportion of analytically missing values, and there are no statistically significant differences between CATI and CAWI respondents (see Table 4). About 4% of the PAPI mode respondents did not provide information about the frequency of drinking alcohol or the amount consumed, and about 7% of them did not provide an answer to at least one K6 item. The income variable stands out as the variable with the highest missingness rate with 13.3% overall. Interestingly, PAPI respondents had a lower propensity not to provide information to the income question. The reason for that might be that online panel participants, knowing that they will be studied over a period of time, had higher privacy concerns in the panel profiling stage.

3.2 Mode effects observed with the non-matching approach

Regarding primacy and recency, as well as probabilities of selecting a specific answer on a scale, we noticed a number of statistically significant differences in distributions of response variables between CAWI, CATI and PAPI modes (see Table 6). However, in most cases those differences cannot be explained with measurement mode effect phenomena described in the literature and some of them

should be attributed to unknown mechanisms, whether self-selection or other measurement effects. There were a total of 27 items (out of 35) with distributional differences identified by regression modelling which we attributed to those unobserved mechanisms. We would expect that CATI respondents would typically have a greater tendency to select the last offered category (i.e., response recency) and both the first and the last categories (extreme category responding), but this was mostly not the case in this study (see Table 6). As a good example of that inconsistency, the CATI respondents had a higher or lower probability of selecting various response options for early adopter items: the first category, strongly agree (a1c), the last category, strongly disagree (a1c), the third category (disagree, a1b, a1d) compared to the second category, agree, and the online mode as reference groups. On the other hand, the propensity was higher for CATI respondents to select the last category *10–completely satisfied* (life satisfaction), and the extreme categories *excellent* and *poor* (general health), compared to both online and mail respondents. This indicates some response recency in telephone surveys. The same can be concluded for a number of Kessler 6 items, while the results indicate some extreme category responding could alternatively be self-selection effects. Generally speaking, it seems that CATI interviewers encouraged more dispersed distributions than we observed in self-administered modes. Moreover, there seems to be some evidence for primacy in PAPI surveys (smoking, household structure). At the same time, CATI respondents had a similar propensity for choosing the first offered answer, which indicates a self-selection effect (e.g., daily smokers are generally more inclined to respond offline). Because there is much more measurement equivalence between CAWI and PAPI modes, not many differences in distributions can be attributed to measurement mode effects.

Besides analytically missing values, differences in responding in different modes due to the question sensitivity were studied with the same regression modelling, controlling for socio-demographics (see Table 6). The results show that the offline PAPI and CATI respondents (some of which were not offered to respond online) were both more likely to report higher frequency and quantity of tobacco and alcohol use. These results imply some fundamental differences between online and offline respondents which could be a result of mode self-selection effect. Assuming that offline respondents are somewhat similar no matter the offline mode (PAPI or CATI), we found some interesting evidence. We noticed that PAPI respondents tend to report higher levels of those harmful behaviours than telephone respondents. Responding to an interviewer might be related to underreporting of particular harmful behaviours compared to responding in the self-administered mode. Further, CATI respondents had a higher propensity to say that they no longer drink than the respondents responding in the other two modes. These differences are small, but they indicate the presence of measurement errors, associated with question sensitivity to socially desirable responding or privacy. We worked with questions with supposedly low sensitivity, and the differences in distributions driven by social desirability (or satisficing) might be much greater if survey questions were more sensitive. Other than that, we did not observe many interpretable effects of modes on measurement. Overreporting satisfaction in CATI mode can be a result of social desirability (see Kocar & Biddle 2020, 2), but there were no statistically significant differences in the averages for life satisfaction and some other variables potentially sensitive to social desirability (e.g., the combined Kessler 6 psychological distress measure).

The results show some additional differences between the modes in income (lower for CATI respondents), Indigenous status (higher for the CATI respondents), and private health insurance (lower for offline respondents), for which there are no theoretical measurement mode effect

foundations. Further, we can observe significant mode self-selection effect for CATI and PAPI modes compared to the online mode – there are more respondents with no internet connection, those who access the internet less frequently or who do not use it for particular purposes.

Generally speaking, it seems that online respondents are significantly different to offline respondents, and offline respondents seem to be much more homogenous, no matter the mode of survey administration (CATI, PAPI). Those differences could lead to incorrect identification of measurement errors, or lack thereof. The items listed above, for which there could be no distribution differences explained by the differences in survey administration modes, will clearly have to be included as controls in matching. Ideally, matching methods would remove self-selection bias, keep the measurement differences between modes consistent with the literature on measurement mode effects, and possibly reveal additional measurement errors due to differences in survey administration. The next four subsections are focused on those changes as a result of data matching.

3.3 Mode effects observed after propensity score matching

The results show that PSM helped reduce the self-selection effect/bias to some extent. While this is not an optimal measure, we can report that, out of 27 variables with distributional differences attributed to unobserved mechanisms, the effects were still present for 19 items after matching (online telephone). For matching covariates, the self-selection effect reduction was, as expected, better than for non-matching covariates, albeit not perfect. It seems that PSM reduced imbalance better for CAWI-PAPI samples, but since many coefficients did not change much, this can as well be attributed to the reduction of sample size. With almost 50% less cases in the matched CAWI-PAPI sample we lost quite some statistical power to observe differences with small effect sizes.

Some of the remaining statistical differences between modes were self-selection effects for CATI mode: frequency of accessing the internet, also for particular purposes (less use), incidence of smoking and amount of alcohol consumption (higher) and income (still lower for CATI). While there were about the same proportions of daily drinkers in online and telephone samples after matching, the propensity to drink daily increased significantly in the PAPI sample relative to the CAWI sample after matching. The same conclusion can be made for those respondents who reported 'fair health' in comparison to 'good health' as a reference category. We observed extreme category responding for two out of five K6 items in telephone surveys even after matching, which could be interpreted as measurement mode effects. Moreover, after pruning we could not confirm most of previously identified satisficing or social desirability. We only noticed that the propensity to report 'no longer drink' (variable b6) was still higher for the CATI sample, which could be both an indicator of social desirability or recency.

However, while the methods reduced imbalance between the samples for the matching variables, it seemed to introduce randomness for some of the variables we purposely excluded from the matching model, and left the other coefficients relatively unchanged. We also noticed that, by pruning about 28% (CAWI-CATI) and about 45% (CAWI-PAPI) of units from the original samples, the evidence indicates that a large portion of retirees were removed from the offline samples, since they were less frequent internet users. Consequently, the propensity to have income in the $300–$399 a week range decreased significantly after matching in both offline mode groups. This indicates that PSM, as a form of indirect matching, can remove particular hidden subgroups from the final sample used for analysis and bias the socio-demographic or socio-economic representativeness of the analytical sample.

3.4 Mode effects observed after Mahalanobis distance matching

As MDM is indirect matching like PSM but with a different distance measure and since we used the same matching covariates, we also observed quite similar mode and self-selection effects as for PSM. Out of 27 variables with putative self-selection effects, some effects were still present for 20 items after matching (CAWI-CATI). There was a significant overlap between the remaining self-selection effects after both MDM and PSM, although we showed that a fairly large proportion of all cases were not matched by both of the distance models (see Figure 2).

The results also present evidence that MDM kept the distributional differences, which could be attributed to measurement mode effects, better than PSM. For example, after MDM we can still identify recency MDM (life satisfaction scale) and extreme category responding (general health scale) in the telephone mode, but not social desirability – no statistical significance for 'had alcoholic drink' variable and 'no longer drink' answer after matching.

3.5 Mode effects observed after exact matching

After EM the results show that, in addition to providing an almost perfect balance on matching covariates, matching EM online-telephone samples helped reduce the self-selection effect better than PSM or MDM – for 14 out of 27 variables with distributional differences attributed to unobserved mechanisms, there were no statistically significant differences anymore. It was also more in line with our expectations and more consistent with the literature on measurement errors due to mixing modes. For example, the differences which could be attributed to measurement mode effects were kept in the CAWI-CATI sample after EM, but eliminated with PSM: self-reported health (extreme category responding), life satisfaction scale (recency), and had alcoholic drink (possible social desirability).

There is also some evidence on mode effects being observed after matching that could not be previously identified. After EM, the results show that the mail respondents answered to the income question with the first category with a much higher propensity than for the second, third, fourth and some other categories. This indicates primacy, even compared to the other self-administered mode, and is in line with item-nonresponse or non-differentiation in PAPI mode (see Tables 4 and 5, related types of satisficing in self-administered surveys).

Last but not least, in contrast to PSM and similarly to MDM, EM did not seem to exclude as many people receiving pension, therefore the propensity for participants receiving $300–$399 a week did not increase significantly, *ceteris paribus*.

3.6 Mode effects observed after coarsened exact matching

In our data, a very small proportion of variables were continuous, and with our logit regression models, all but one variable out of the eight we selected for matching was categorical (coloured green in Table 6). In our methods assessment case, as previously explained, it means that there is almost no difference between the units matched and weighted by EM and CEM. Since only one out of the carefully selected covariates was continuous, the matching results were very similar to exact matching, with only 14 additional matches due to coarsening. The similarity can also be seen by the correlation coefficient for EM and CEM weights, which equalled 0.987 for the CAWI-PAPI matched sample and 0.995 for the CAWI-CATI matched sample. Consequently, the findings related to

controlling for mode self-selection with CEM to study mode effects, are the same as for EM (see Table 6 for all coefficients).

4 Discussion and recommendations

Mixed-mode surveys are increasingly common and seem to be the standard in probability-based online panel research. Panel organisations providing measurement equivalence like ELIPSS panel (e.g., tablets for all panellists) are more of an exception than not (see Kaczmirek et al. 2019, 4–5). The evidence from this research, as well as from the study on longitudinal panel mode effects (Kocar & Biddle 2020), suggests that while effects of modes on measurement can definitely be observed in probability-based mixed-mode research, the impact on the results is mostly relatively minor. However, that surely does not mean that methods for identification and adjustment should not be investigated and developed further as measurement mode effects can be very item specific.

In this study, we tried to identify mode effects using five distinctive approaches dealing with non-random assignment of respondents to modes. After carrying out the first, non-matching method, the evidence suggested that mode self-selection appeared to be the main reason for the differences in response variable distributions between the modes, which was previously reported by Dennis et al. (2005). This could lead to incorrect assumptions on measurement mode effects which could actually be self-selection bias, or vice versa. We conducted the rest of the study having in mind at least three possible and overlapping applications of data matching in mixed mode online panel research to deal with measurement errors: (1) in the questionnaire development stage to achieve measurement equivalence in both data collection modes (e.g., pilot testing on a smaller sample of onliners and offliners), (2) in longitudinal studies using a mixed mode online panel allowing for mode switching (see Kocar & Biddle 2020), and (3) in mode effect testing with an aim to adjust for mode effects (see Kennedy et al. 2012; Kolenikov & Kennedy 2014). However, the evidence from this study can be used for similar applications in longitudinal or mixed-mode cross-sectional research, particularly with web-push approaches.

We studied mode effect with a number of different approaches and methods. Firstly, we used non-matched data to investigate satisficing related sources of measurement error – non-differentiation/straightlining, item nonresponse and providing non-substantive answers. Mail mode was the mode with much higher propensity for those types of satisficing, especially for more sensitive items. There were few differences between CAWI and CATI modes, but we noticed that fewer online mode respondents non-differentiate. This is consistent with the findings of Dennis et al. (2005), but not in line with the findings of De Leeuw et al. (2019), who found evidence that telephone respondents provided less consistent responses. We also found that mail respondents (paper administration) have a higher propensity to skip a question while telephone and online respondents (computer-assisted) have a tendency to not provide a substantive answer instead, which is a form of question format effect. All in all, it seems to be much easier to find evidence on 'technical' effects of mode (e.g., missingness) than 'distributional' effects of modes (e.g., primacy, social desirability).

However, we could also find some evidence of distributional types of measurement mode effects, and they varied by different matching approaches. With the first approach, we identified some recency and extreme category responding in telephone surveys, consistent with the findings by De Leeuw et al. (2019). We observed potential primacy in PAPI surveys (again, consistent with the findings by Dennis et al. 2005), as well as potential social desirability. Most of the distributional differences were

attributed to mode self-selection effect and we later decided to control it by data matching, which includes pruning units and weighting with particular matching methods.

PSM successfully removed a portion of self-selection bias, but also affected the ability to identify mode effects. The reason for that might be that the method does not match directly on target variables. Since the matches are made based on propensity scores, they perform much better on covariates with greater contribution to the propensity score, but could bias the other variables which happen to be somewhat associated with the probability of a unit being pruned. This cannot be fully controlled with PSM, especially in the context of studying mode effects and excluding variables sensitive to the effects from the matching model. For those and other reasons, some literature suggests not using PSM (e.g., King & Nielsen 2019), and we have to agree with their recommendations to some extent. On the other hand, the same authors (King & Nielsen 2019) advised using MDM as an alternative classic distance model, but we found fewer advantages to that matching method than their study might have suggested. While more of measurement mode effects consistent with the literature were kept in the matched data than with PSM, MDM failed to remove as much self-selection bias than the other distance-based method. EM and CEM performed equally well, since only one of eight matching variables was continuous and needed to be slightly coarsened. Both stratification methods helped reduce the self-selection effect better than the distance-based methods and the findings on measurement mode effects were more in line with our expectations based on the literature review, especially compared to PSM. After EM and CEM matching, we could still report some extreme category responding and recency in the telephone mode and, additionally, primacy in the mail mode. Lastly, we have to be aware that with these approaches, we try to find the 'truth' on the presence of mode effects, which still cannot be fully confirmed until we conduct a sophisticated fully randomised mixed-mode survey experiment, or get access to a very similar single-mode dataset with estimates representative for the studied population. Data matching seems to help investigate measurement mode effects, but our evidence suggests that it is far from being perfect. The most convincing evidence of that imperfection is the removal of particular hidden subgroups by PSM (i.e., retirees in a particular income group). On the other hand, we have to note that working with small treatment group subsamples makes studying mode effects even more challenging, although this should not be limited to our data matching exercise. The PAPI subsample often did not offer enough statistical power for mode effects estimation, especially after pruning almost 50% of all units. Combined with more measurement equivalence between online and mail modes, it was difficult to disentangle a lack of statistical power from measurement mode effects and mode self-selection effects. This is a relevant limitation for probability-based online panel research as offline samples are often small compared to online samples in countries with high internet penetration rates.

If we wanted to mix modes to cover the offline population, and we had to choose from the measurement mode effect perspective (that is, leaving aside budget and other concerns), then the telephone mode should probably have a slight advantage over the mail one. In this study, we found evidence of less non-differentiation and item nonresponse in the telephone data collection. The exception to this general recommendation would be surveys with socially sensitive questions, although we found little evidence on social desirability. Based on the theory on measurement mode effects, as well as the evidence from this study, the mail mode as a self-administered mode seems to offer more measurement equivalence to the online mode, but at the expense of different forms of satisficing compared to the telephone mode. And, as De Leeuw (2005) explained, mixing modes can compensate for weaknesses of each single modal method. Taking into account the geographical size

of the country and other disadvantages of mailing survey questionnaires, the telephone mode should remain the preferred mode to collect data from the offline population in probability-based online panel research in countries with large land mass like Australia or the United States.

Although measurement mode effects, as defined and described in the literature, can be observed in this study using different approaches, they are not as apparent or prevalent as other authors in this space suggested. Again, there are no benchmarks for the 'truth' available. One of the reasons for the lack of identified mode effects might be that the questionnaire used to collect the data used in this study was carefully designed. We could argue that questionnaire design followed general suggestions to minimise measurement differences across all survey modes: very similar question and visual format, wording and conversational clues, the questionnaire was purposely designed with mixing in mind, etc. (for details see Dillman et al. 2014, 404–415). The other reason could be that mode effects are question-specific and the likelihood of a mode effect depends on the nature of the question (Kennedy et al. 2012). It is possible that the items available for this study were not very sucseptible to measurement mode effects; in the questionnaire, there were no extremely sensitive questions, items with long ranges of nominal or ordinal responses (except for the life satisfaction item), or questions to be answered in a very socially desirable fashion. In addition, the length of the survey should not have encouraged the same extent of satisficing as longer surveys. We suggest researchers look for opportunities to repeat this analysis on questions more prone to mode effects, on long survey questionnaires, and with more continuous variables (if good predictors) to fully utilise the potential of CEM.

Identifying mode effects in multi-mode surveys is a difficult problem analytically, which is why there are still no straightforward, guaranteed-to-work solutions and any existing approach involves some degree of compromise to internal or external validity. However, one of the important finding of this paper is that it is even more challenging to investigate mode effects in probability-based online panel studies without carrying out data matching. In an optimal randomised design, all online and offlline respondents would have to have an equal probability of being assigned to either the online or the offline mode. However, this kind of randomisation is almost impossible, since most offline respondents cannot or refuse to respond online. There may also be quite large nonresponse for those who would normally respond online and are approached to complete offline. Even if it was possible, such survey designs have been rare in panel studies due to high costs and extensive effort to implement (Cernat et al. 2016). In theory, onliners and a little portion of offline respondents could be randomized to modes, but the results on mode effects could not be generalisable due to the non-randomized offline subsample. We would sacrifice external validity for internal validity. A possible solution to that is using matching methods, with exact matching and especially coarsened exact matching being better solutions than distance model matching. In that case, it seems to be possible to partially disentangle mode effects from subsample composition effects, i.e., the unobserved mechanisms for selection of the mode after controlling for demographics. If we manage to do that, then the adjustment for mode effects, suggested in some literature (e.g., Kennedy et al. 2012; Kolenikov & Kennedy 2014), could represent added value in online panel research as the accuracy of the estimates could be improved. Further, other matching methods, such as machine learning matching methods, and propensity scores weighting could be evaluated for that purpose. To further improve matching quality and balance, combining the results of different matching methods, similarly to using an ensemble of methods in machine learning to improve the accuracy of estimation and prediction, should be considered. It also has to be determined what matching parameters work best,

and how much pruning is needed to achieve enough sample balance to reduce more self-selection bias, while not affecting the potential to identify measurement mode effects. That would be a nice data simulation exercise. All in all, the study of measurement mode effect seems to be an interesting space for further development in mixed-mode and online panel research from the methodological perspective.

5 References

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

Australian Bureau of Statistics (2018b, October 2). Internet Activity, Australia. Retrieved from https://www.abs.gov.au/statistics/industry/technology-and-innovation/internet-activity-australia/latest-release.

Australian Communications and Media Authority. (2020). Mobile-only Australia: living without a fixed line at home. Retrieved from https://www.acma.gov.au/publications/2020-12/report/mobile-only-australia-living-without-fixed-line-home.

Australian Government (2020). PSMA Geocoded National Address File (G-NAF). Retrieved from https://data.gov.au/data/dataset/19432f89-dc3a-4ef3-b943-5326ef1dbecc.

Baker R, Blumberg SJ, Brick JM, Couper MP, Courtright M, Dennis JM, ... & Kennedy C (2010). AAPOR report on online panels. *The Public Opinion Quarterly* 74(4):711–781.

Benjamini Y & Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* B 57: 289-300.

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

Beulens B, van der Laan J, Schouten B, van der Brakel J, Burger J & Klausch TG (2012). *Disentangling mode-specific selection and measurement bias in social surveys*. Discussion paper, Statistics Netherlands, The Hague.

Callegaro M & DiSogra C (2008). Computing response metrics for online panels. *Public Opinion Quarterly* *72*(5):1008–1032.

Cernat A, Couper MP & Ofstedal MB (2016). Estimation of mode effects in the health and retirement study using measurement models. *Journal of Survey Statistics and Methodology* 4(4):501–524.

Couper MP, Kapteyn A, Schonlau M & Winter J (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research* 36(1):131–148.

De Leeuw D (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics* 21(2), 233–255.

De Leeuw E, Hox J & Scherpenzeel A (2011). Mode effect or question wording? Measurement error in mixed mode surveys. In *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 5959–5967. Online verfügbar unter https://www. amstat. org/sections/srms/Proceedings/y2010/Files/400117.

De Leeuw ED, Hox J & Scherpenzeel A (2019). Mode effects versus question format effects: an experimental investigation of measurement error implemented in a probability-based online panel.

In: Lavrakas P, Traugott M, Kennedy C, Holbrook A, de Leeuw E & West E (eds), *Experimental methods in survey research: techniques that combine random sampling with random assignment*, John Wiley & Sons, New York, 151–165.

De Leeuw E & van der Zouwen J (1988). Data quality in telephone and face-to-face surveys: A comparative analysis. In Groves, RM, Biemer, PP, Lyberg, LE, Massey, JT, Nicholls II, WL, & Waksberg, J (eds) *Telephone survey methodology*, John Wiley & Sons, New York, 283–299.

Dennis JM, Chatt C, Li R, Motta-Stanko A & Pulliam P (2005). Data collection mode effects controlling for sample origins in a panel survey: telephone versus internet, In: *60th Annual Conference of the American Association for Public Opinion Research*, Miami Beach, FL.

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

Dillman DA, Smyth JD & Christian LM (2014*). Internet, phone, mail, and mixed-mode surveys: the tailored design method*, John Wiley & Sons, New York.

DiSogra C, Cobb C, Chan E & Dennis JM (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In *Joint Statistical Meetings (JSM), Survey Research Methods*, 4501–4515.

Duffy B, Smith K, Terhanian G & Bremer J (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research* 47(6):615.

Dutwin D & Buskirk TD (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability Internet samples to low response rate probability samples. *Public Opinion Quarterly* 81(S1):213–239.

Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA & Davidian M (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology* 173(7):761–767.

Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E & Tourangeau R (2009). *Survey methodology*. Hoboken.

Head ML, Holman L, Lanfear R, Kahn AT & Jennions MD (2015). The extent and consequences of p-hacking in science. *PLoS Biology* 13(3), e1002106

Iacus SM, King G & Porro G (2012). Causal inference without balance checking: coarsened exact matching. *Political Analysis* 20(1):1–24.

Jäckle A, Roberts C & Lynn P (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review* 78(1):3–20.

Johnson VE (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* 110(48):19313–19317.

Kaczmirek, L, Phillips B, Pennay DW, Lavrakas PJ & Neiger D (2019). Building a probability-based online panel: life in Australia. *CSRM and SRC Methods Paper No.2/2019,* Australian National University, Canberra.

Kennedy C, Ackermann A, Turakhia C, Emerson M & James A (2012). Mode effects measurement and correction: a case study. Presented at the *2012 Annual Meeting of the American Association for Public Opinion Research*, Orlando, FL.

King G & Nielsen R (2016). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):1–20.

King G, Lucas C & Nielsen R (2016). MatchingFrontier: Automated matching for causal inference. *R package version*, *2*(0).

King G & Nielsen R (2019). Why propensity scores should not be used for matching. *Political Analysis* 27(4):435–454.

Kocar S & Biddle N (2020). Panel mixed-mode effects: does switching modes in probability-based online panels influence measurement error? *CSRM and SRC Methods Paper No.1/2020.* Australian National University, Canberra.

Kolenikov S & Kennedy C (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology* 2(2):126-158.

Krosnick, J A, Narayan, S, & Smith, WR (1996). Satisficing in surveys: initial evidence, *New directions for evaluation, Special Issue: Advances in Survey Research,* 70:29-44.

Lavrakas PJ (2008). *Encyclopedia of survey research methods*, Sage Publications.

Leite W (2016). *Practical propensity score methods using R*, Sage Publications.

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

Pennay D, Borg K, Neiger D, Misson S, Honey N & Lavrakas P (2016b). *Online Panels Benchmarking Study, 2015.* doi:10.4225/87/FSOYQI, ADA Dataverse, V1.

Pennay DW, Neiger D, Lavrakas 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 No.2/2018.* Australian National University, Canberra.

Perrin N & Bertoni N (2017). *Converting mail mode panelists to web and measuring their early internet experiences*, Pew Research Center, Washington, DC.

Rosenbaum PR & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41-55.

Rosenbaum PR (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application* 7:143–176.

Schouten B, van den Brakel J, Buelens B, van der Laan J & Klausch T (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research* 42(6):1555–1570.

Sizemore S & Alkurdi R (2019). Matching Methods for Causal Inference: A Machine Learning Update. Available at https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/.

Suzer-Gurtekin ZT, Valliant R, Heeringa SG & de Leeuw ED (2018). Mixed-mode surveys: design, estimation, and adjustment methods. In: Johnson, TP, Pennell, B-E, Stoop, IAL & Dorer, B (eds) *Advances in comparative survey methods: multinational, multiregional, and multicultural contexts (3MC)*, John Wiley & Sons, New York, 409–430.

Tourangeau R, Rips LJ & Rasinski K (2000). *The psychology of survey response*, Cambridge University Press.

Vannieuwenhuyze, JT & 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.

Appendix

Table 4 Analytically missing values (all [54] and for selected sensitive items) (%)

Mode	Skipped	Non-substantive answers	Total analytically missing	Sensitive items (average % of missing values [skipped or non-substantive])				
	mean % per unit	mean % per unit	mean % per unit	K6score	b4 - smoking	b6 - frequency drinking alcohol	b7 - alcohol consumption	d16 - income
Mail[A]	2.85%[BC]	0.04%[BC]	2.89%[BC]	7.02%[BC]	2.06%[BC]	4.41%[BC]	3.86%[BC]	7.43%[BC]
Telephone[B]	0.01%[AC]	0.91%[A]	0.92%[A]	2.55%[A]	0.29%[A]	0.27%[A]	1.08%[A]	13.78%[A]
Online[C]	0.12%[AB]	0.84%[A]	0.96%[A]	3.26%[A]	0.30%[A]	0.04%[A]	0.60%[A]	13.5%[A]
Total	0.20%	0.84%	1.04%	3.25%	0.39%	0.33%	0.91%	13.25%

[A B C] = indicate statistically significant differences between the groups at p=0.01 level, pairwise t-testing ([A]=mail, [B]=telephone, [C]=online)

Table 5 Non-differentiation statistics (%)

Mode	Non-differentiation status			
	No	Potential straightliner* (early adopter items only)	Potential straightliner* (K6 items only)	Straightliner** (both early adopter and K6 items)
Mail[A]	49.57%[BC]	19.83%	19.4%[BC]	11.21%[BC]
Telephone[B]	66.18%[A]	17.79%	11.2%[AC]	4.83%[AC]
Online[C]	69.46%[A]	20.15%	7.43%[AB]	2.96%[AB]
Total	67.32%	19.38%	9.29%	4.01%

*respondent selected the same value/answer for all early adopter items or all K6 items

**respondent selected the same value/answer for all early adopter items and then the same value/answers for all K6 items

[A B C] = indicate statistically significant differences between the groups at p=0.01 level, pairwise Chi-Square testing ([A]=mail, [B]=telephone, [C]=online)

Table 6 Differences in distributions, four approaches and methods, mode and mode self-selection effects

Variable	Category	RM/ RC	Socio-demographic controls (no matching)		Propensity score matching		Mahalanobis distance matching		Exact matching		Coarsened exact matching	
		Online	Telephone	Mail	Telephone	Mail	Telephone	Mail	Telephone	Mail	Telephone	Mail
			Coef	Coef	Coef	Coef			Coef	Coef	Coef	Coef
a1a - early adopter - try new products early (multinomial regression)	Strongly agree		-0.16	-0.65	-0.35	0.00	-0.35	0.17	-0.29	-0.86	-0.29	-0.49
	Agree	RC										
	Disagree		-0.03	0.21	0.00	0.82†	-0.04	0.47	-0.01	0.42	-0.01	0.34
	Strongly disagree		0.11	0.49	0.04	0.59	0.03	0.52	0.07	0.62	0.07	0.55
a1b - early adopter - try new brands early (multinomial regression)	Strongly agree		0.22	0.25	0.43	0.70	0.53	0.92	0.50	0.48	0.50	0.93
	Agree	RC										
	Disagree		-0.23*	0.13	-0.15	0.83*	-0.17	0.38	-0.15	0.85*	-0.15	0.84†
	Strongly disagree		-0.02	0.38	-0.09	0.34	-0.04	0.35	-0.09	0.67	-0.10	0.67
a1c - early adopter - shopping for new things (multinomial regression)	Strongly agree		0.59**	-0.56	0.78**	-0.74	0.78**	-0.10	1.02**	-0.78	1.03**	-0.56
	Agree	RC										
	Disagree		0.28**	0.24	0.33*	0.26	0.38**	0.10	0.5**	0.16	0.5**	0.11
	Strongly disagree		0.73**	0.95**	0.53*	0.10	0.66**	0.60	0.94**	1.08†	0.92**	1.05†
a1d - early adopter - like to be first (multinomial regression)	Strongly agree		-0.08	0.76	-0.13	0.34	-0.11	-0.06	0.10	0.09	0.11	-0.05
	Agree	RC										
	Disagree		-0.36**	0.37	-0.16	0.48	-0.37*	0.45	-0.20	0.59	-0.2	0.43
	Strongly disagree		-0.21	0.64†	-0.11	0.28	-0.19	0.54	-0.04	0.80	-0.05	0.7
a1e - early adopter - like to talk about new things (multinomial regression)	Strongly agree		0.06	0.03	0.05	0.2	0.11	-0.40	0.06	0.11	0.06	0.35
	Agree	RC										
	Disagree		0.02	0.11	0.05	0.13	-0.04	0.32	0.03	0.91*	0.03	0.87*
	Strongly disagree		-0.04	0.29	-0.3	-0.62	-0.15	0.22	0.08	0.50	0.07	0.52
a1 - early adopter score (OLS regression)			-0.01	0.67**	-0.09	0.27	-0.08	0.55	0.03	1.03†	0.02	0.91
a2_1 - internet connection (broadband)	No	RC										
	Yes		-1.41**	-1.99**	-0.26	0.08	-0.97**	-1.08	-0.26	-0.79	-0.26	-0.83
a2_2 - internet connection (dial-up, ISDN)	No	RC										
	Yes		0.04	0.44	-0.12	-0.74	-0.05	0.95	0.47	1.67	0.48	1.64
	No	RC										

a2_3 - internet connection (mobile device)	Yes		0.61**	-0.19	-0.33**	-0.43	0.76**	-0.56	-0.19	-0.62	-0.19	-0.59
a2_4 - internet connection (no internet)	No	RC										
	Yes		3.00**	3.69**	0.46	-0.5	3.25**	outcome does not vary	1.51	outcome does not vary	1.52	outcome does not vary
a3 - using internet (multinomial regression)	Several times a day	RC										
	About once a day		0.66**	1.06**	0.36*	0.05	0.64**	1.16†	0.35†	0.64	0.35†	0.6
	Three to five days a week		1.16**	1.54**	0.33	0.37	0.97**	2.12	0.37	1.62†	0.38	1.63
	One to two days a week		1.88**	2.05**	0.53†	-0.56	1.43**	2.66	0.43	2.95†	0.43	2.74†
	Every few weeks		1.64**	2.76**	0.84†	0.31	2.37*	few units only	0.49	very few units	0.49	few units only
	Once a month		3.18**	2.96*	0.54	3.25	3.10†	no units	2.27	no units	2.27	no units
	Less often		2.88**	3.78**	1.3*	few units only	no units	no units	no units	no units	0.99	no units
	Never		4.95**	5.93**	0.8	few units only	4.34**	no units	0.99	very few units	0.01	few units only
a3 - using internet (ordinal regression)			1.8**	2.31**	0.43**	0.27	0.99**	1.64**	0.4*	1.10†	0.4*	1.07†
a4a - internet use (searching information)	Several times a day	RC										
	About once a day		-0.02	-0.08	-0.04	-0.24	-0.05	0.06	0.01	0.39	0.01	0.31
	Three to five days a week		0.55**	0.6	0.39†	0.85	0.32	0.71	0.37†	0.39	0.37†	0.32
	One to two days a week		0.93**	1.51**	0.52**	2.16**	0.55*	1.55†	0.54*	1.73*	0.52*	1.66*
	Every few weeks		0.5*	1.52**	-0.04	1.77†	0.23	1.66	-0.03	1.98	-0.04	1.98
	Once a month		1.43**	1.85**	0.67†	-0.55	1.23**	3.06	0.62	3.40	0.62	3.41
	Less often		2.43**	3.56**	1.37**	2.43†	2.18**	1.65	2.15**	2.81†	2.15**	2.56†
	Never		2.67**	1.21	4.01**	few units only	only few units	no units	2.88†	no units	2.87	no units
a4b - internet use (social media)	Several times a day		-0.83**	1.15	-0.88**	1.69	-1.11**	0.98	-0.99**	0.45	-0.97**	0.43
	About once a day		-0.6*	0.89	-0.81**	1.5	-0.9**	0.05	-0.75*	-0.43	-0.73*	-0.39

	Three to five days a week		-0.26	0.91	-0.35	0.88	-0.47	0.20	-0.66†	0.07	-0.65†	0.09
	One to two days a week		-0.27	0.81	-0.43	0.7	-0.36	0.38	-0.56†	0.14	-0.55	0.06
	Every few weeks		-0.6*	1.55†	-0.69†	1.7	-0.74*	0.83	-0.84*	-0.12	-0.83*	-0.06
	Once a month	RC										
	Less often		0.28	1.41	-0.02	1.32	-0.13	0.76	-0.11	1.00	-0.09	0.94
	Never		0.14	1.76*	-0.04	1.74	-0.3	0.82	-0.25	0.23	-0.24	0.2
a4c - internet use (financial transactions)	Several times a day		-1.06**	-1.78*	-0.74†	-1.16	-0.77†	-1.50	-0.9*	-2.69†	-0.91*	-2.61†
	About once a day		-0.98**	-1.64**	-0.83**	-0.86	-0.91**	-2.05	-1.13**	-3.1*	-1.14**	-2.83*
	Three to five days a week		-0.79**	-0.45	-0.59†	-0.32	-0.83**	-0.88	-0.73*	-1.58	-0.74*	-1.42
	One to two days a week		-0.67**	-0.78	-0.54†	0.32	-0.59†	-1.06	-0.66*	-1.49	-0.67*	-1.36
	Every few weeks		-1.1**	-0.6	-1.08**	0.05	-1.11**	-0.85	-1.3**	-1.84	-1.3**	-1.67
	Once a month	RC										
	Less often		0.78**	0.37	0.51†	-0.57	0.53	-0.93	0.2	-1.25	0.2	-1.19
	Never		0.26	0.67	0.12	0.59	0.34	0.34	0.11	0.02	0.12	0.08
a4d - internet use (blog/forum)	Several times a day		-0.94†	-0.21	-1.1†	0.31	-1.02	1.12	-0.84	-0.8	-0.79	-0.8
	About once a day		-0.71†	0.13	-0.61	1.79	-0.68	1.43	-0.83	0.2	-0.8	0.38
	Three to five days a week		-0.23	1.79**	-0.48	1.91†	-0.5	1.94†	-0.95†	1.4	-0.94†	1.47
	One to two days a week		0.25	1.34**	0.15	2.53**	0.15	2.44**	0.04	2.01*	0.08	1.99*
	Every few weeks		-0.66†	0.35	-0.87*	0.86	-0.8†	1.06	-1.13**	-0.25	-1.11**	-0.18
	Once a month	RC										
	Less often		0.19	-1.47**	0.04	-0.86	0.03	-1.63	-0.22	-1.34	-0.22	-1.4
	Never		0.85**	-0.63	0.73*	-0.46	0.7*	-0.96	0.4	-0.98	0.38	-1.02
a5 - no. of online surveys (OLS regression)			-0.17*	-0.15	-0.4*	-1.27	-0.06	-0.07	-0.49	0.04	-0.49	0.02
b1 - life satisfaction (multinomial regression)	Not at all satisfied 0		0.34	-1.12	0.54	few units only	few units only	few units only	-1.15	very few units	-1.16	few units only
	1		-0.05	0.04	-0.64	few units only	few units only	few units only	-2.44	0.62	-2.42	0.43
	2		1.11**	1.4	0.39	1.86	-0.14	few units only	1.82*	very few units	1.82*	few units only

	3		-0.12	0.27	-0.27	0.41	0.52	0.72	0	-0.33	0	0.07
	4		-0.22	-0.06	0.47	-0.69	-0.33	0.64	-0.18	-0.65	-0.18	-0.64
	5		0.3†	0.18	0.28	0.24	0.11	0.30	0.33	-0.23	0.33	-0.37
	6		-0.25	-0.25	0.28	-0.81	-0.24	-0.13	0.03	0.15	0.03	0.14
	7		-0.05	-0.12	0.11	-0.22	-0.19	0.07	-0.07	-0.3	-0.06	-0.32
	8	RC										
	9		-0.06	-0.04	-0.19	-0.05	-0.25	-0.18	-0.28	-0.4	-0.29	-0.39
	Completely satisfied 10		0.47**	0.02	0.03	-0.52	0.54**	-0.02	0.5*	-0.71	0.49*	-0.66
b1 - life satisfaction (ordinal regression)			0.07	0.01	-0.22†	0.14	0.16	-0.14	-0.07	0.01	-0.07	0.04
b2 - general health (multinomial regression)	Excellent		0.66**	-0.1	-0.12	-0.35	0.54**	-0.24	0.66**	0.17	0.65**	0.14
	Very good		-0.14	-0.69**	-0.03	0.01	0.02	-0.85†	0.02	-0.69†	0.01	-0.68†
	Good	RC										
	Fair		0.38**	0.38	0.23	1.51**	0.32	0.95	0.41*	0.74	0.4*	0.68
	Poor		0.88**	-0.53	0.22	0.96	0.96**	few units only	1.04**	0.2	1.04**	0.07
b2 - general health (ordinal regression)			0.19*	0.43**	0.2†	0.59†	0.01	0.82*	0.1	0.39	0.1	0.37
b3a - Kessler 6 nervous (multinomial regression)	All of the time		1.15**	few units only	1.43*	few units only	0.48	2.53	0.61	very few units	0.6	few units only
	Most of the time		0.12	-0.53	0.46	1.43	-0.06	-0.72	0.52	-0.06	0.52	-0.05
	Some of the time		-0.03	-0.06	0.11	1.15*	-0.15	0.21	-0.01	0.7	-0.01	0.69
	A little of the time		-0.48**	-0.62**	-0.34**	-0.2	-0.36**	-0.56	-0.4**	-0.32	-0.39**	-0.3
	None of the time	RC										
b3b - Kessler 6 hopeless (multinomial regression)	All of the time		1.6**	0.93	0.72	1.06	2.58†	few units only	0.69	0.52	0.7	0.46
	Most of the time		-0.06	-0.13	0.03	0.5	0.05	1.74	0.21	0.1	0.22	0.11
	Some of the time		0.29*	0.17	0.55**	-0.28	0.05	0.2	0.3	-0.12	0.3	-0.16
	A little of the time		0.05	-0.16	0.08	0.22	-0.11	-0.22	-0.03	-0.01	-0.03	-0.07
	None of the time	RC										
b3c - Kessler 6 restless or fidgety (multinomial regression)	All of the time		1.42**	0.22	1.67**	0.48	1.52*	few units only	2.36**	-1.38	2.32**	-1.33
	Most of the time		0	-0.03	0.12	1.04	-0.08	0.63	-0.01	-0.12	0	-0.06
	Some of the time		0.03	-0.56†	0.26†	-0.4	-0.16	-0.68	0.13	-0.33	0.13	-0.32

	A little of the time		-0.57**	-0.29	-0.35*	-0.28	-0.63**	-0.35	-0.48**	-0.29	-0.48**	-0.2
	None of the time	RC										
b3d - Kessler 6 depressed (multinomial regression)	All of the time		0.61	few units only	-0.09	few units only	few units only	few units only	-0.18	very few units	-0.18	few units only
	Most of the time		0.14	0.9	0.3	0.45	-0.82	few units only	-0.05	0.35	-0.06	0.33
	Some of the time		0.47**	0.03	0.52**	-0.05	0.19	-0.06	0.37	0.19	0.37	0.2
	A little of the time		0.1	0.03	0.33†	0.13	0.07	0.2	-0.03	0.57	-0.03	0.63
	None of the time	RC										
b3e - Kessler 6 everything effort (multinomial regression)	All of the time		0.89**	-0.25	0.59	-0.29	1.23**	0.96	0.88*	-0.75	0.89*	-0.78
	Most of the time		0.2	-0.03	0.18	0.47	0.09	0.37	0.31	0.12	0.31	0.03
	Some of the time		0.04	-0.31	0.21	-0.03	0.01	-0.36	0.1	0.19	0.1	0.11
	A little of the time		-0.34**	-0.15	-0.19	-0.52	-0.38**	-0.2	-0.33*	0.03	-0.33*	0.07
	None of the time	RC										
b3f - Kessler 6 worthless (multinomial regression)	All of the time		0.8†	0.46	0.57	-0.41	-0.03	few units only	0.51	very few units	0.51	
	Most of the time		-0.06	-0.05	-0.67	-0.76	0.01	few units only	0.01	0.61	0.02	0.63
	Some of the time		0.34*	0.1	0.41†	1.68†	0.08	0.29	0.24	0.21	0.24	0.38
	A little of the time		-0.1	0.22	0.14	-0.56	-0.05	0.28	0.02	-0.12	0.02	-0.12
	None of the time	RC										
K6 score (OLS regression)			-0.4*	0.2	-0.54*	-0.27	0.03	-0.31	-0.32	0.29	-0.32	0.29
b4 - smoking (multinomial regression)	Daily		0.79**	1.18**	0.63**	0.77	0.63**	1.21	0.54*	0.5	0.54**	0.57
	At least weekly (but not daily)		0.65†	1.42†	1*	2.61	0.94†	3.22	1.06*	2.01	1.06*	1.97
	Less often than weekly		0.39	-0.55	0.56	-0.58	0.4	0.64	0.42	0.04	0.43	0.02
	Not at all (but in last 12 months)		-0.14	0.38	-0.24	0.22	0.07	0.76	0.06	1.02	0.06	0.99
	Not at all (not in last 12 months)	RC										
b4 - smoking (ordinal regression)			-0.62**	-0.96	-0.51**	-0.71	-0.54**	-1.27†	-0.47*	-0.70	-0.48*	-0.74
b5 - had alcoholic drink	No	RC										

	Yes		-0.43**	-0.63**	-0.07	-0.06	-0.27	-0.63	-0.46*	-0.57	-0.47*	-0.52
b6 - frequency drinking alcohol (multinomial regression)	Every day		0.45**	0.43	0.16	1.29†	0.22	1.46	0.08	2.02*	0.11	1.97*
	5 to 6 days a week		-0.35†	-0.26	-0.45†	0.57	-0.43†	-0.41	-0.61*	-0.38	-0.57*	-0.38
	3 to 4 days a week		-0.08	0.26	-0.23	0.74	-0.24	0.05	-0.41†	0.22	-0.39†	0.16
	1 to 2 days a week	RC										
	2 to 3 days a month		-0.36*	-0.66	-0.52*	0.29	-0.44†	-0.28	-0.52*	-0.12	-0.5*	-0.26
	About 1 day a month		-0.16	-0.92†	-0.41	-0.69	-0.31	-0.41	-0.33	-0.98	-0.32	-1.03
	Less often		0.01	-0.51	-0.39†	0.26	-0.4	-0.4	-0.48†	-0.44	-0.48†	-0.47
	No longer drink		1.42**	-0.68	1.37*	1.07	1.07	0.54	1.49*	very few units	1.57*	-1.21
b6 - frequency drinking alcohol (ordinal regression)			-0.1	-0.57**	-0.14	-0.54	-0.16	-0.56	-0.1	-0.94*	-0.1	-0.89*
b7 – alcohol consumptions when drinking (multinomial regression)	9 or more drinks		1.55**	2.23**	1.35**	3.65†	1.1**	0.9	1.36**	1.34	1.36**	1.36
	7-8 drinks		0.3	2.05**	0.39	2.86†	-0.04	2.71	0.56	3.04†	0.53	3.08†
	5-6 drinks		0.91**	1.67**	0.72**	0.08	0.93**	1.22	0.77**	1.46	0.77**	1.4
	3-4 drinks		0.3*	0.42	0.18	0.09	0.21	0.5	-0.02	0.4	-0.01	0.34
	2 drinks	RC										
	1 drink		0.26†	0.06	0.16	-0.31	0.06	0	-0.03	0.34	-0.04	0.33
	Half a drink		-0.2	-0.45	-0.24	0.01	-0.7	-1.36	-0.31	0.12	-0.31	0.03
b7 – alcohol consumptions when drinking (ordinal regression)			-0.32**	-0.74**	-0.26*	-0.49	-0.32**	-0.69†	-0.28†	-0.46	-0.28†	-0.44
c1 - household structure	Person living alone		0.51**	0.66**	-0.03	-0.2	0.3†	0.27	0.06	0.44	0.06	0.49
	Couple living alone	RC										
	Couple with non-dependent child(ren)		-0.02	-0.54	-0.05	-0.77	0.03	0.89	-0.16	0.37	-0.17	0.41
	Couple with dependent child(ren)		-0.03	-0.27	-0.15	0.31	-0.27	-0.58	-0.22	0.1	-0.21	0.02
	Couple with both (dep, non-dep)		-0.28	-0.99	-0.09	-0.24	-0.57	-1.88	-0.26	-0.29	-0.26	-0.28
	Single parent with only non-		0.45†	0.11	0.17	0.57	0.90*	-0.44	0.32	0.33	0.32	0.36

	dependent child(ren)											
	Single parent with dependent child(ren) or both		-0.10	-1.02	-0.31	-1	-0.37	0.02	-0.22	0.7	-0.23	0.65
	Non-related adults sharing		-0.07	0	-0.18	0.17	0.07	1.69	-0.51	-1.25	-0.52	-1.2
	Other household type		0.31	-0.43	-0.02	0.1	-0.09	few units only	-0.2	-1.39	-0.2	-1.35
c2 - number of household members (OLS regression)			-0.04	-0.19**	0.06	0.02	0.02	-0.01	0	-0.07	0	-0.08
c3 - living at current address 5 years ago	No	RC										
	Yes		0.3**	0.6**	0.22	0.36	0.51**	0.46	0.07	0.41	0.07	0.39
d3 – highest level of schooling (multinomial regression)	Year 12 or equivalent	RC										
	Year 11 or equivalent		0.35*	0.76*	0.11	0.88	0.45†	1.29	0.15	0.93	0.15	0.94
	Year 10 or equivalent		0.45**	0.45†	-0.06	-0.23	0.37*	0.34	-0.06	-0.26	-0.06	-0.26
	Year 9 or equivalent		0.81**	0.68	-0.09	0.57	-0.11	1.20	-0.44	-0.74	-0.47	-0.66
	Year 8 or below		1.34**	0.68	0.21	-0.53	1.31*	2.47	-0.31	-1.58	-0.32	-1.51
	Did not go to school		2.42*	few units only	few units only	no units	no units	no units	no units	no units	no units	no units
d3 – highest level of schooling (OLS regression)			0.64**	0.46*	0.02	-0.27	0.35**	0.6	-0.27	0.27	-0.1	-0.25
d10 - Australian Citizen	No	RC										
	Yes		0	-1.09**	0.31	0.6	0.22	-0.43	0.56	-1.73†	0.56	-1.73†
d12 - LOTE	No	RC										
	Yes		0.13	-0.52	0.03	-0.59	0.07	-1.25	-0.02	-0.05	-0.02	-0.02
d13 - Indigenous status	No	RC										
	Yes		0.98**	-0.05	0.56	1.32	1.61**	no units	0.9†	-1.25	0.91†	-1.25
d15 – private health insurance	No	RC										
	Yes		-0.63**	-0.62**	-0.27†	-0.17	-0.3*	-0.72	-0.33†	-0.29	-0.32†	-0.26
d16 - income (multinomial regression)	$2,000+ per week	RC										
	$1,500 - $1,999 per week		-0.59*	-0.51	-0.68*	-1.2	-0.91**	-0.42	-0.43	-1.68*	-0.42	-1.47†

	$1,250 - $1,499 per week		-0.03	-0.4	-0.04	-1.57	-0.01	-0.31	0.33	-2.71*	0.33	-2.13†
	$1,000 - $1,249 per week		-0.16	-0.41	-0.21	-0.63	-0.25	-0.29	-0.02	-1.59†	-0.03	-1.47†
	$800 - $999 per week		-0.43	0.24	-0.54	-0.24	-0.43	0.42	-0.38	-0.12	-0.38	0.06
	$600 - $799 per week		0.03	-0.11	-0.21	-0.2	-0.16	0.31	0.08	-0.33	0.07	-0.22
	$400 - $599 per week		-0.03	-0.34	-0.23	-0.89	-0.48	0.04	-0.22	-1.67†	-0.23	-1.52†
	$300 - $399 per week		-0.72*	-0.38	-1.24**	-1.61†	-0.47	-0.04	-0.82†	-1.37	-0.81†	-1.19
	$200 - $299 per week		-0.15	-0.55	-0.32	-0.84	-0.51	-0.56	-0.64	-2.1†	-0.63	-1.89
	$1 - $199 per week		0.16	0.1	-0.34	0.39	-0.24	-0.82	0.08	-0.73	0.07	-0.57
	Nil income or negative income		0.82**	0.27	0.41	-0.1	0.28	-0.25	0.53†	-1.2	0.52†	-1.01
d16 - income (ordinal regression)			0.57**	0.1	0.31**	0.17	0.22†	-0.02	0.29*	-0.05	0.29*	-0.06

RM/RC=Reference mode/reference category, Coef=logit/multinomial/ordinal/multiple linear regression beta coefficients, **significant at false discovery rate (FDR)=0.05, *significant at false discovery rate (FDR)=0.1, †significant at false discovery rate (FDR)=0.2