# Approaches to dealing with survey errors in online panel research

Sebastian Kocar

August 2021

A thesis submitted for the degree of
Doctor of Philosophy of The Australian National University.

## Candidate Declaration

Except where otherwise indicated, this thesis is my own original work. This thesis has not been submitted by me in whole or part for a degree or diploma in any university or other tertiary education institution.

Chapters 3, 4, 5, 6, 7, 8, and 9 of this thesis are based on journal articles and methods papers that I co-authored with my panel members and research collaborators, as follows:

- Chapter 3 was co-authored with Dr Lars Kaczmirek, and I contributed roughly 80–90% of the material in that chapter.
- Chapter 4 was co-authored with Paul J. Lavrakas, PhD, and I contributed roughly 70–80% of the material in that chapter.
- Chapters 5, 6, and 8 were co-authored with Professor Nicholas Biddle, and I contributed roughly 80–90% of the material in those chapters.
- Chapter 7 was co-authored with Professor Nicholas Biddle and Dr Benjamin Phillips, and I contributed roughly 80% of the material in that chapter.
- Chapter 9 was co-authored with Dr Bernard Baffour, and I contributed roughly 70% of the material in that chapter.

Sebastian Kocar

05/08/2021

## Acknowledgments

I would like to thank my primary supervisor and chair of my supervisory panel, Professor Nicholas Biddle, for his supervision, support, co-authorship, and engaging with me in many interesting discussions throughout my candidature that were often not limited to survey methodology and social research.

I would also like to thank my supervisory panel and my research collaborators for providing support and advice, and reviewing my papers/chapters. My special acknowledgment goes to the Social Research Centre for providing access to their data and a number of their staff for providing valuable insight into commercial social research.

I have to acknowledge the Department of Education, Skills and Employment for funding my PhD program with the Australian Government Research Training Program Scholarship, and the ANU Centre for Social Research and Methods for their resourcing of my research program. I would also like to acknowledge Ms Hilary Bek, the Expert Editor team, and Biotext for providing copyediting and proofreading advice.

Last but not least, I have to thank my wife Gözde for believing in me and giving me both space and unlimited support to do something I love, and my parents for successfully convincing me over the years that education is, in fact, very important.

## Abstract

Survey research is a relatively young field, and online surveys including online panel surveys are now routinely used for collecting survey data. We distinguish between different types of online panels, and this thesis is focused on both probability-based and nonprobability-based general population panels. To increase the quality of online panels in the era of nonresponse, more methodological research is needed, and that is the focus of the research in this thesis.

To investigate approaches to dealing with survey errors, the Total Survey Error paradigm as a conceptual framework is applied, and both errors of representation and errors of measurement are the subject of this research. One of the contributions of this thesis is a review and discussion of a range of data sources and methodology which can be used in the study of survey errors. The other theoretical and practical contributions, presented within three groups, are related to the investigation of individual types of survey errors in online panel research.

First, worldwide probability-based online panels are identified, and their methodological approaches to recruitment and data collection reviewed and compared as part of a meta-analysis. The study shows high levels of heterogeneity in both recruitment rates and recruitment solutions, as well as explains variability of recruitment rates. The other studies on errors of representation present evidence on how online panel paradata can be effectively transformed and used to identify about three in four nonrespondents in a subsequent panel wave, and answer the question of why people participate in online panel surveys while presenting evidence on how social-psychological theories can explain survey participation in a longitudinal design.

Second, two studies focus on measurement error in probability-based online panel research due to mixing modes. The study on measurement mode effects shows how measurement error is present in the case of a lack of measurement equivalence between modes, and presents evidence on how applying matching methods (like coarsened exact matching) quite effectively controls for self-selection bias due to non-random assignment of online panellists to modes. The study on individual-level measurement mode effects presents a newly identified source of measurement error in online panel survey, that is, panel measurement mode effects. It also conceptualizes and showcases how panel conditioning can be a factor of two measurement aspects. These results are later related to a trade-off between representation (undercoverage) and measurement bias.

Third, the thesis studies two cost- and time-efficient approaches to online data collection – nonprobability online panels and a fairly new combination of random digit dialing, text message invitations, and web-push methodology. The study on nonprobability panels, which are generally considered as less accurate but cheaper than probability-based panels, investigates post-survey adjustment methodology to improve inference in nonprobability samples. It presents evidence on how

accuracy can be improved under different external data access scenarios. The study on a new approach to online survey data collection shows very low response rates, and outlines effective solutions to increase response (such as advance SMS and reminders). It also presents evidence on the fairly high accuracy of the proposed approach, which seems to be feasible for continuing recruitment to a probability-based online panel.

In the final section of the thesis, the cost dimension of online survey research is discussed, the requirement of collecting data from the offline population in probability-based online panel research from different perspectives is challenged, and the theoretical contributions of this research are explained in more detail.

## Contents

## Figures

## Tables

## Abbreviations

| | |
|---|---|
| AAE | average absolute error |
| AAPOR | American Association for Public Opinion Research |
| ABS | Australian Bureau of Statistics |
| A-BS | address-based sampling |
| AIC | Akaike Information Criteria |
| ANU | Australian National University |
| ARB | Average absolute relative bias |
| ATP | American Trends Panel |
| CAPI | Computer-assisted personal interviewing |
| CASI | Computer-assisted self-interviewing |
| CATI | Computer-assisted telephone interviewing |
| CAWI | Computer-assisted web interviewing |
| CDT | cognitive dissonance theory |
| CEM | coarsened exact matching |
| CH | compliance heuristics |
| CI | confidence interval |
| COMR | completion rate |
| CSRM | Centre for Social Research and Methods (ANU) |
| CUMR | cumulative response rates |
| D-AEMR | Dynamic Almost-Exact Matching with Replacement |
| DFRDD | dual-frame random digit dialing |
| DiSC | Dominance, Influence, Steadiness, Compliance |
| EET | economic exchange theory |
| ELIPSS | Enhancing Learning by Improving Process Skills in STEM |
| ELSST | European Language Social Science Thesaurus |
| EM | exact matching |
| ESB | English-speaking background |
| ESOMAR | European Society for Opinion and Marketing Research |
| EU | European Union |
| F2F | face-to-face |
| FDR | false discovery rates |
| FFRISP | Face-to-Face Recruited Internet Survey Platform |
| G-NAF | Geocoded National Address File |
| GIP | German Internet Panel |
| GREG | generalized regression estimation |
| HILDA | Household, Income and Labour Dynamics in Australia Survey |
| IDI | in-depth interviews |
| ISO | International Organization for Standardization |
| IVR | interactive voice response |
| LGBTQ | lesbian, gay, bisexual, transgender, and queer or questioning |

| | |
|---|---|
| LISS | Longitudinal Internet studies for the Social Sciences |
| LOTE | language other than English |
| LST | leverage-saliency theory |
| MDM | Mahalanobis distance matching |
| MM | mixed mode |
| MRP | multilevel regression and poststratification |
| NDSHS | National Drug Strategy Household Survey |
| NESB | non-English-speaking background |
| NHS | National Health Survey |
| O4TS | Open Four Temperament Scales |
| OCEAN | Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism |
| OLS | ordinary least squares |
| OPBS | Online Panels Benchmarking Study |
| ORR | overall recruitment rate |
| OS | operating system |
| PAPI | Paper-and-pencil personal interviewing |
| PCA | principal components analysis |
| PROR | profile rate |
| PSM | propensity score matching |
| PSW | propensity score weighting |
| RAA | reasoned action approach |
| RDD | random digit dialing |
| RECR | recruitment rate |
| RETR | retention rate |
| RMSE | root mean square error |
| RQ | research question |
| RR | response rates |
| SEIFA | Socio-Economic Indexes for Areas |
| SEM | structural equation modeling |
| SET | social exchange theory |
| SMS | short message service |
| SMD | standardised mean difference |
| SPT | self-perception theory |
| SRC | Social Research Centre |
| TSE | Total Survey Error |
| UK | United Kingdom |
| US/USA | United States of America |

# Chapter 1 Introduction: online panels and survey errors

Surveys are a key tool and one of the most commonly used methods to study society, to test theories of behavior, and to estimate different population parameters. They are a systematic method for gathering information with an aim to construct quantitative descriptors of population attributes, and can measure either everyone in the population or a sample of the population (Groves 2004). Compared to most other social science disciplines, survey research is a relatively young field with three distinctive stages of development: the first era between 1930 and 1960 (the era of innovation), the second era between 1960 and 1990 (the era of expansion) and the third era after 1990. In this last era, response rate continued to deteriorate, costs increased, face-to-face (F2F) interviews continued to reduce in volume and the traditional telephone surveys declined in coverage due to the introduction of mobile phones (Groves 2011). Surveys are at a turning point as a result of both societal and technological changes, and few technologies have had as much of an impact as the internet on survey data collection (Couper 2017). We can argue that in the age of smartphones, the 2020s should be focused on optimization of mixed-mode designs, with an aim to achieve better coverage and lower nonresponse (Dillman 2018). That should be especially true for self-administered web surveys[1].

Since the beginning of this century, more survey research has been moving online, in part because that mode of survey data collection can be considered as cheaper, less time consuming, easy to implement, is computerized with an ability to use multimedia, and flexible with regard to time and geography. Also, increasing percentages of people have access to the internet, as well as its use for exchanging opinions, resulting in the internet enabling access to unique populations (Callegaro et al. 2015; Wright 2005). As large-scale face-to-face or mixed-mode probability-based surveys as an accurate instrument for measuring population parameters are unlikely to retain a central role in empirical survey research, there are opportunities for less-expensive methods, including probability and nonprobability online panels (Couper 2017). These online methods are now routinely used for collecting survey data and have been increasing in numbers. They are a form of access panel, working as a sample database of respondents who can be selected for survey participation, and can include large numbers of panellists (Callegaro et al. 2014). They can be drawn up to meet specific needs of studies, the final sample can vary based on the studied population or survey topic (Baker et al. 2010), and they present opportunities for longitudinal analysis. While it is difficult to pinpoint the start of online panel research (some consider Dutch Telepanel as the pioneer (Callegaro & DiSogra 2008)), the period between the mid-1990s and until about 2005 is considered the era of large growth (Callegaro et al. 2014). Since then, various solutions have been developed in online panel research. There are different types of

[1] Web surveys, also known as internet or online surveys, collect data from respondents via the internet. Sometimes, web survey respondents are also sampled via the internet (Nathan 2008).

panels using distinctive sampling strategies while collecting data from different populations. Even if panels are of the same type and study the same population – for example, probability-based general population online panels – there are often substantial differences in their recruitment and data collection strategies, as well as panel management.

Regarding panel composition, we distinguish between general population panels, specialty panels, proprietary panels, and election panels (Baker et al. 2010). The most common type is general population panels, which consist of a diverse range of people from the general population, sometimes including smaller subpopulations that are normally hard to reach (Callegaro et al. 2014). Those types of online panels are the main subject of this thesis. Regarding selection to the sample, the most common types of online panels are nonprobability-based[2] ones, with probability-based panels remaining quite rare and smaller in size (Baker et al. 2010), especially in lower population countries like Australia (Pennay et al. 2016). The most important distinction between the two types is in their sampling methodologies: while nonprobability-based online panels accept practically any adult who sees an open invitation and is willing to join the panel, probability-based online panels invite only those selected with established probabilistic sampling methodologies[3], such as address-based sampling (A-BS)[4], area sampling[5], or random digit dialing (RDD)[6] (Callegaro et al. 2014). As a result of those notable methodological differences, probability-based panels are considered of higher quality and overall accuracy than volunteer panels (e.g., Baker et al. 2010; Yeager et al. 2011). Despite probability-based panels being perceived as possible alternatives to more expensive traditional probability-based survey methods (Bosnjak et al. 2016), and although there are several clear advantages of online panel surveys – such as fast data collection, lower cost, and sampling efficiency (Callegaro et al. 2014) – there are also notable limitations and shortcomings of this kind of research, many of which are shared with longitudinal[7] data more broadly. Lastly, online panels can be divided into online panels collecting data online only, some of which provide 'offline' respondents technology to respond online, and online panels using mixed-mode[8] strategy by interviewing panellists in an 'offline' mode (for example, Paper-and-pencil personal interviewing (PAPI)) (Blom et al. 2016). While opt-in panels tend to use one mode only, it is also plausible for an opt-in panel to use mixed modes of data collection, even if it is quite

[2] Nonprobability-based panels are also known as volunteer, opt-in or access panels (Baker et al. 2010). In this thesis, these terms are used interchangeably.

[3] In probability samples and in contrast to nonprobability samples, "each member of the population has a known non-zero probability of being chosen into the sample" (Hade & Lemeshow 2008).

[4] Address-based sampling "involves the selection of a random sample of addresses from a frame listing of residential addresses" (Link 2008a).

[5] Probability sample based on area sampling is a sample with geographic areas sampled with a known probability, and are usually a part of multi-stage or cluster designs (Hall 2008).

[6] RDD is an approach for drawing a sample from the frame or set of telephone numbers (Brick 2008).

[7] Longitudinal studies are those involving multiple measurements on a sample of individuals over a period of time (Kalaian & Kasim 2008).

[8] Mixed-mode or multimode surveys collect data for a single project using more than one survey mode by combining different ways of survey data collection (Link 2008b).

rare. These differences in panel coverage and their effects on data accuracy have many practical implications, thus they prompt further investigation.

Online panels can be considered a hybrid between longitudinal and online survey research. As such, they are known for advantages and disadvantages related to both longitudinal and online surveys, some of which represent notable sources of survey errors in online panels (e.g., panel conditioning[9], nonresponse, voluntary attrition[10] (Kocar 2020)). In practice, organizations managing online panels have to monitor panel participation behavior, optimize survey completion, and control for changes in representation and active panel size. There are different types of online panel members based on their panel participation behavior[11]. Due to the differences in their (response) profile, personalities, non-demographic and socio-demographic characteristics (known as differential panel nonresponse/attrition (Callegaro et al. 2014)), good panel management and/or panel recruitment/refreshment is important to keep panel representative of the studied population.

There are three things valued above all in survey research: affordability, speed, and quality (Keeter 2019). And while online panels are known at least in theory, for lower cost and fast data collection (Callegaro et al. 2014) survey data quality is dependent on how well different sources of survey errors are managed and errors mitigated. To evaluate different approaches to dealing with survey errors in online panel research, I will use the Total Survey Error (TSE) Framework firstly introduced by Groves (2004, also see Groves et al. 2009; Groves & Lyberg 2010). In the TSE framework, errors are firstly divided into errors of representation[12] and errors of measurement[13]. Errors of representation are further divided into coverage error, sampling error, nonresponse error and adjustment error (also see Chapter 11 for a review of TSE in web surveys), while errors of measurement are further split into specification error, measurement error, processing error and inferential error (Lavrakas 2013).

Application of TSE in online panel research in surely not a new approach to improving the quality of data gathered with either volunteer or probability-based panels. However, one could argue that it is even more challenging to apply the TSE framework and study concurrent survey errors due to the time dimension, which is also associated with sources of survey error specific to longitudinal survey data collection. A high frequency of surveys, often infrequent panel refreshment, and predominantly online data collection can create additional challenges for data collectors as well. Undercoverage of people

[9] Panel conditioning is related to repeated surveys and results in respondent's previous participation/contact influencing their survey response (Cantwell 2008).
[10] Voluntary attrition is "the proactive action of panel members to contact the company and ask to be removed from the panel" (Callegaro et al. 2014).
[11] The literature lists the following types: stayers, slow starters, fast attritors, gradual attritors, lurkers (Lugtig 2014), sleepers, dozers, comatose, 'gold star' respondents, and backouts (Lavrakas et al. 2018) (see Appendix 4 for combined classification).
[12] Errors of representation apply "to the representation of the target population by the weighted net sample" (Fuchs 2008).
[13] Errors of measurement add "to the total error by affecting the edited survey responses obtained from a respondent" (Fuchs 2008).

without the internet, panel conditioning in monthly (or more regular) surveys, allowing switching between survey modes, self-selection in volunteer panels, and differential nonresponse/attrition are just some of them. Therefore, instead of studying all errors from TSE in detail, this thesis primarily investigates the errors which contribute the most to the total survey error in web surveys and online panel surveys. Also, in each of the chapters I and my co-author(s)[14] focus on new or adjusted approaches to identification, comparison, and mitigation of errors, bias, and sources of errors in online panel research. As online panel research is fairly new compared to most other more traditional survey modes, research designs and data collection approaches, there is still a lot to be learned about specific errors in online panel research.

In survey methodology, there are a number of theories explaining phenomena related to survey errors, in addition to the TSE conceptual framework. The broader theoretical frameworks include theory of survey participation (e.g., Groves et al. 1992), theory of sampling (based on probability theory) (e.g., Deming 1950; Neyman 1938), and theory of the response process (e.g., Tourangeau et al. 2000), while some other theories explaining narrower phenomena include satisficing theory (Krosnick 1991), leverage-salience theory (Groves et al. 2000), and benefit-cost theory (Singer 2011). To explain survey participation behavior, social-psychological theories such as social exchange theory, self-perception theory, and compliance heuristics are also applied in practice (for a full review, see Chapter 4). As the thesis investigates various types of survey errors and is based on an interdisciplinary approach, it applies and contributes to all of the theoretical and conceptual frameworks listed above. There is slightly more focus on theory of survey participation: nonresponse error is studied by looking at societal-level factors in a meta-analysis (country effect), at attributes of the survey design (response maximization strategies), at characteristics of the sample person (socio-demographic predictors of nonresponse and attrition), and at respondent-interviewer interaction (recruitment outcomes) as the dimensions of the theory.

This thesis and its chapters aim to answer the following overarching research question: *What are the most prevailing survey errors in online panel research (with the largest impact on data quality/accuracy), and what are the most suitable approaches to dealing with them, including those for identification, reduction, correction, and balancing of survey errors?* In the following paragraphs, I will briefly present the current state of online panel survey research, identify relevant research gaps related to different survey errors, outline the objectives of each chapter, explain how those chapters are related, and discuss the relevance of each study in this thesis for online panel research practice.

[14] Plural forms 'we', 'us', and 'our(s)' are used throughout the thesis for papers/chapters written in co-authorship. Singular forms 'I' and 'my' are used for single-authorship papers/chapters and my own ideas/findings/conclusions.

## 1.1 Data sources and methods to study survey errors in online panels (Chapter 2)

While this work is primarily focused on development of methodology to dealing with survey errors in panel research, one of the main goals and potential contributions to the literature is also to explore applicability of different data sources and methods to study errors in online panel surveys. This research is presented in Chapter 2. To identify the most suitable research methods for studying survey errors, a decision for an interdisciplinary and multimethod[15], as well as multi-mode research was made. In the end, disciplines such as survey methodology, survey statistics, data science, and social-psychology are intertwined in this thesis.

First, survey methodology and web survey methodology as research disciplines are predominantly based on quantitative evidence about particular methods since they are fundamentally quantitative research approaches. In an attempt to find new evidence which could answer questions that cannot be answered by quantitative data alone, qualitative data were collected. This made one of the studies presented in Chapter 4 a mixed-method research study combining both qualitative and quantitative evidence in a triangulation design. The other chapters are based on quantitative data from various data sources with different types of data.

Second, online panel research normally produces a variety of data sources suitable for methodological research, although survey data might have been primarily collected to study topical and not methodological issues. In addition to cross-sectional type of data collected with panel surveys (so-called waves), data files could be linked over time and longitudinal data created for time series or panel data analyses. As one of the advantages of computerized data collection, panels can automatically or semi-automatically collect data about each panellist's panel behavior, including device, questionnaire-navigation, and online panel paradata (Callegaro 2013). Moreover, survey experiments can be used with online panels to test for different methodological solutions, both in recruitment and survey completion stages, and to study both errors of representation and measurement. Lastly, data from nationally representative high-quality sources can be used to assess the relative accuracy of online panels.

Third, as the evidence based on online panel data from a particular country often cannot easily be generalized across the borders, data could be retrieved from different data and metadata sources, synthesized and used with a meta-analytical approach. Survey methodology is a discipline in which meta-analyses are regularly conducted to study different survey errors (mostly nonresponse errors), but they are still less prevalent than in some related disciplines, such as psychology (Čehovin et al. 2018) or population health. In the case of probability-based panels, most organizations managing them

[15] Multimethod research is the application different research methods or data to the investigation of research questions; a term mixed methodology is frequently used instead (Lewis-Beck et al. 2004).

are more open to publicly sharing their data and/or methodological information (such as response rates) than commercially oriented volunteer panels. This offers opportunities for exploring online panel research phenomena in different countries and on several continents.

## 1.2 Probability-based online panels and 'nonrecruitment' (Chapter 3)

There has been limited comparative evidence on the effectiveness of different recruitment and panel management strategies in probability-based online panel research, and even the studies which compared methodological approaches in different panels (e.g., Blom et al. 2016), presented localized and often mixed evidence. On the other hand, some of those panels carried out survey experiments to determine the best solutions to maximization of recruitment rates (e.g., Cornesse et al. 2021; Rao et al. 2010), but their findings were often country, survey-participation culture, or time specific in a fast-changing survey methodology environment. Also, the preliminary analysis revealed notable differences in how online panels carry out recruitment, as well as how they approach wave-by-wave survey data collection, and how they build long-term relationships with their panellists.

In Chapter 3, we primarily deal with the issues of efficient recruitment to probability-based online panels – that is, how to increase recruitment rates with various response maximization approaches. One could argue that recruitment is the most important step from the errors of representation TSE perspective. It is the first stage of the panel lifecycle; thus, any potential representation bias is carried over to the other stages, which might not be easy to identify, control, or correct for. The experience of recruited panellists in the recruitments stage can also affect their future experience and panel survey completion behavior, as well as their decision to opt-out of the panel. This is later discussed in Chapter 4. Also, there are not only TSE-related implications of recruitment to a panel, but cost-related implications, as low recruitment and survey completion rates, and high voluntary attrition rates, can all result in higher costs and additional efforts to maintain the panel representative of the population.

Since we collated evidence from the majority of probability-based online panels from around the world, we are able to present methodological differences between online panels from different times and different countries. Some online panels even changed their methodological approaches over time, and that further shows how online and online survey panel research evolved over time. In Chapter 3, I and my co-author present an overview of panels, their recruitment and panel management strategies, as well as the effectiveness of their approaches to recruitment. Since the analysis revealed differences between the panels with substantial effects on different survey errors and costs, the results from that chapter act as the basis on which the rest of the thesis is built.

## 1.3 Probability-based online panels and 'nonparticipation' (Chapters 4 and 5)

While the empirical evidence from Chapter 3 can reveal some important practical implications for existing probability-based online panels, as well as for organizations planning to establish a new panel, there is still not a profound understanding of why people join an online panel, why they participate in online panel research, why they decide to stop participating, and why some of them opt-out from the panel. At the same time, existing survey participation theory (for review see Albaum & Smith 2012; Keusch 2015), which is based on social-psychological theories, has a better ability to explain survey response (and nonresponse) in cross-sectional surveys than longitudinal or panel surveys. Expanding the theory to online panel surveys and identifying both motivational factors and barriers, is the subject of Chapter 4. To extend the findings on recruitment strategies presented in Chapter 3, we are discussing relevant evidence and practical solutions for all stages of online panel lifecycle, including survey completion and attrition.

Moreover, besides understanding the 'why' aspect of panel participation behavior, we also study how previous panel participation both influences and predicts future panel participation behavior. This is the subject of Chapter 5. Since online panel research offers new opportunities due to the abundance and the format of collected data (see Chapter 2 for more information), we attempted to take advantage of this fact and tried to identify new statistical solutions to predict panel survey nonresponse and voluntary attrition before they actually happen. As this is more of a prevention than intervention approach, this evidence (combined with evidence from Chapter 4) could have important practical implications for panel management attempting to increase survey completion and extend the average panellist's panel lifespan. Chapter 5 addressed an important gap in the literature and survey practice.

Keeping panellist in the panel longer does not have only positive effect on panel survey sample sizes, delayed panel refreshment or associated cost; so-called differential attrition can have a negative impact on representativeness of the panel, which can result in the introduction of nonignorable bias, and lower data quality and accuracy (Callegaro et al. 2014). This is not limited to primary socio-demographics (which can be adjusted for with weighting), and secondary demographics (which might be associated with primary demographics as weighting variables), but also to non-demographics (attitudinal, behavioral, knowledge or other factual variables). Hence, Chapter 5 discusses both the effect of socio-demographics on nonparticipation and their predictive power in combination with previous panel participation trends.

## 1.4 Probability-based online panels and undercoverage (Chapter 6)

Web surveys are still known for lower response rates than surveys using other data collection modes (Daikeler et al. 2020), and while nonresponse might be a greater reason for concern, coverage bias as the difference between web users and non-users still exists (Couper et al. 2018). Due to the fact that not every household or individual has access to the internet, probability-based online panels have to address the issue of undercoverage of those who cannot (or are unwilling to) respond to surveys online (Callegaro & DiSogra 2008).

The evidence from Chapter 3 reveals a fair amount of inconsistency between the online panels in practice. Some panels completely exclude the offline population as they do not offer survey completion using an offline mode. Other panels offer internet technology to their offline population or collect data offline, but the proportion of offline panellists in the online panel is much smaller than the proportion of people with no access to the internet in that particular country. This raises a question of how important coverage of offline population is in practice, and how much coverage bias is introduced if data are, conveniently, collected online only. Unfortunately, the evidence in the existing literature does not offer sufficient insight into how important it is to invest additional funds into collecting data from so-called 'offliners' and the effect on data quality, if not following the general recommendations on minimizing the defined undercoverage bias.

Chapter 6 addresses the problem of data quality/accuracy in case of single-mode (online) data collection in probability-based online panel research. Building on the evidence provided by Eckman (2016) who carried out analysis with Dutch data (LISS panel), we provide additional evidence on the extent of hypothetical additional representation bias in probability-based online panel research (for Australia). However, the expanded study on the effects of exclusion of the offline population does not only investigate the extent of hypothetical undercoverage bias, that is, how would the final results differ if only the data for the online population was included. Instead, we identified the need to introduce more complexity into addressing this issue. Thus, undercoverage bias was studied conditional on the size of the offline population, on the magnitude of differences between the population and on the survey topic, as well as relative to nationally representative estimates from high-quality government surveys (benchmarks). As such, it provides more comprehensive evidence on this type of survey bias than the existing literature. By studying nonresponse and undercoverage in Chapters 4–6, the most prevalent sources of representation bias are covered in this thesis.

## 1.5 Probability-based online panels and measurement errors (Chapters 7 and 8)

Another type of error which has been previously studied in survey data and longitudinal survey data literature, but to a lesser extent in online panel research, is measurement error. There are different

types and sources of measurement error in web surveys and in mixed-mode online panels, including item nonresponse, inconsistent responding, straightlining, and fast completion (Tourangeau et al. 2009). Since studying every source of measurement error would be out of scope of this thesis, Chapters 7 and 8 focus on those aspects of online panel research which introduce new challenges for measuring topical issues (over time) accurately. These chapters are partly associated with coverage bias analyzed in Chapter 6 – if an online panel is mixed-mode, undercoverage bias can be mitigated – but it becomes more challenging to offer measurement equivalence between modes. Fundamentally, we trade measurement bias for representation bias.

In Chapter 7, we focus on measurement mode effect[16] as a result of mixing modes in probability-based online panel research. There is an ongoing debate regarding whether an offline mode should be offered to offline respondents or if offline respondents should be instead provided technology, that is, internet access or a device like a tablet with access to the internet, to participate in online surveys. For example, Enhancing Learning by Improving Process Skills in STEM (ELIPSS) panel from France provided a tablet to their panellists which should guarantee measurement equivalence (Blom et al. 2016) and Face-to-Face Recruited Internet Survey Platform (FFRISP) from the United States (US) offered a $500-worth laptop (Krosnick et al. 2009). On the other hand, these extra recruitment and data collection expenses could be spent on other panel management solutions. Therefore, we aim to establish the severity of the issue from a measurement error perspective, and assess it in combination with undercoverage bias (if excluding the offline population). To study measurement mode effect in mixed-mode research lacking random assignment of respondents to modes, which can be a problem in mixed-mode research including web-push strategy, methods and techniques for controlling self-selection to mode can be used. While matching methods[17] are commonly applied in fields such as epidemiology or behavioral economics, they are not as often used in survey methodology and statistics, and this chapter addressed an important methodological research gap.

In Chapter 8, we introduce and investigate two issues related to measurement error in (online) panel survey research, both of which are specific to the probability-based online panel studied in this thesis. If an organization managing the panel allows their panellists to switch modes over time (e.g., initially Computer-assisted telephone interviewing (CATI) and Computer-assisted self-interviewing (CASI)) or uses CATI to occasionally interview online panellists (as a reminder followed by an interview), the effects of this mode-switch could potentially be observed in their answers in a longitudinal design. We call these phenomena 'panel mode effects', and they are specific to mixed-mode panel survey

[16] Measurement mode effect is an influence of the survey mode of data collection on measurement. Measurement error is related to mode and can be linked to the questionnaire, interviewer or respondents (Jans 2008).

[17] Matching is a method for preprocessing data to reduce imbalance and to improve casual inferences. Examples of matching methods are propensity score matching, exact matching, and coarsened exact matching (King & Nielsen 2019).

research. The other issue normally observed in longitudinal studies is panel conditioning. In comparison to annual longitudinal/panel studies, this source of error might be even more severe if survey data are collected more frequently and identical questions asked with shorter time gaps. To the best of our knowledge, Chapter 8 presents the first study investing both of those sources of measurement error concurrently.

## 1.6 Nonprobability-based online panels and total survey error (Chapter 9)

While Chapters 3-8 investigate different phenomena in probability-based online panel research, Chapter 9 focused exclusively on nonprobability-based online panels. On one hand, some advantages of those panels are even clearer than of probability-based panels, such as convenience (self-selection process of recruitment) and low costs of data collection (Baker et al. 2010). On the other hand, there has been extensive research (scientific articles and conference presentations) on the lower accuracy of nonprobability-online panels in comparison to their probability counterparts (e.g., Chang & Krosnick 2009; MacInnis et al. 2018; Malhotra & Krosnick 2007; Yeager et al. 2011). As a response of that, research has been carried out to determine if and to what extent can the accuracy of nonprobability online surveys be increased (e.g., DiSogra et al. 2011; Dutwin & Buskirk 2017; Mercer et al. 2018), but often presenting mixed evidence and partial solutions to mitigating bias in nonprobability samples.

In addition to not identifying consistent and efficient ways of improving accuracy of nonprobability-based online panel surveys, the existing literature is lacking some key methodological conventions on how to approach the problem of dealing with representation bias in those sample surveys. Thus, in this chapter we attempt to add new post-survey adjustment methodologies to the range of methods previously used in this space, test new approaches of identifying the best covariates, and look at the problem from the external data availability perspective. Our investigation is purposely building on the existing research in this space, while adding complexity to the study of bias in nonprobability samples, and presenting a new country example which might, to some extent, differ from the findings from the US.

Chapter 9 is an important component of this thesis as it does not only address the problem of improving quality of non-demographic estimates from nonprobability-based online panel surveys, it also provides valuable evidence on the future role of nonprobability-based online panels in social, academic and government-funded research. Simply put, if methodologically sound and consistently efficient ways of improving the quality of samples not based on probabilistic principles (in the design, data collection, post-survey adjustment stages) can be identified, that type of panels can be used more frequently as a cheaper alternative to market research and polling spaces.

## 1.7 New alternative probability-based online data collection approaches and total survey error (Chapters 10 and 11)

Due to the lack of comprehensive internet sampling frame with every member having a non-zero chance of being selected (Kennedy et al. 2016), recruitment to probability web surveys should be carried out offline using one of more 'traditional' survey modes, that is, PAPI, CATI or Computer-assisted personal interviewing (CAPI) (Callegaro et al. 2014). For that reason, probability-based online panels consistently use offline methods to recruit their panellists, and repeated data collection from the same respondents justifies the effort and financial investment into offline recruitment to a panel. While similar approaches are used in push-to-web surveys, the same strategies are often not time- and cost-effective for cross-sectional probability web-only surveys. As a result, new probability sampling and data collection approaches have emerged, including Interactive Voice Response (IVR) surveys, text message surveys, or text-to-web surveys.

Also, as a result of a presence of various survey errors in probability-based online panels studied in Chapters 5, 7 and 8 – namely attrition, measurement mode effects and panel conditioning – I decided to investigate a cross-sectional single-mode alternative to repeated panel web data collection. While text-message surveys and text-to-web surveys have previously been used, the proposed approach of combining text-to-web with random digit dialing sampling of mobile numbers has only been outlined in Bucher & Sand (2021) to the best of my knowledge. As this is a fairly new combination of survey design solutions, very little evidence exists on the suitability and quality of this type of data collection. Besides the fact that prior consent to text survey invitations is required in a number of countries (Fordyce et al. 2020; Kongsgard et al. 2014), this approach might suffer from high nonresponse, undercoverage of people without access to the internet and a smartphone, and a generally low accuracy as a result of those limitations. Therefore, I experimentally collected the data to study total survey error in a cross-sectional smartphone survey. Chapter 10 focuses on factors affecting nonresponse in an RDD-sampling SMS-invitation web-push survey, including the effect of nonresponse and undercoverage on socio-demographic representation bias, and Chapter 11 investigates relative accuracy of various non-demographic estimates from the same survey (a benchmarking approach).

On the other hand, the studied combination of approaches to data collection might be useful for other purposes as well, as being a cost- and time-efficient mean of recruitment to a survey. For example, in Chapter 2 we identified different recruitment practices, including those using CAPI, PAPI, CATI, or IVR recruitment modes. However, Life in Australia™ might have been the first probability-based online panel to carry out SMS push-to-web recruitment at the end of 2020 (Phillips et al. 2021). Hence, the findings from my non-panel study have practical implications for probability-based online panel research as well.

## 1.8 References


Albaum, G., & Smith, S. M. (2012). Why people agree to participate in surveys. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 179-193). Springer.

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

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

Bosnjak, M., Das, M., & Lynn, P. (2016). Methods for probability-based online and mixed-mode panels: Selected recent trends and future perspectives. *Social Science Computer Review*, *34*(1), 3-7.

Brick, J. M. (2008). Random-digit dialing. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 675-678). Sage.

Bucher, H., & Sand, M. (2021). Exploring the Feasibility of Recruiting Respondents and Collecting Web Data Via Smartphone: A Case Study of Text-To-Web Recruitment for a General Population Survey in Germany. *Journal of Survey Statistics and Methodology (2021) 00, 1–12.*

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

Callegaro, M. (2013). Paradata in web surveys. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information Improving surveys with paradata: Analytic uses of process information* (pp. 261-279). Wiley.

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look at the future. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 1-22). John Wiley & Sons.

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

Cantwell, P. J. (2008). Panel conditioning. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 566-567). Sage.

Cornesse, C., Felderer, B., Fikel, M., Krieger, U., & Blom, A. G. (2021). Recruiting a Probability-Based Online Panel via Postal Mail: Experimental Evidence. *Social Science Computer Review*. https://doi.org/10.1177/08944393211006059

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology, 43*, 121-145.

Couper, M. P., Gremel, G., Axinn, W., Guyer, H., Wagner, J., & West, B. T. (2018). New options for national population surveys: The implications of internet and smartphone coverage. *Social Science Research*, *73*, 221-235.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641-678.

Čehovin, G., Bosnjak, M., & Lozar Manfreda, K. (2018). Meta-Analyses in Survey Methodology: A Systematic Review. *Public Opinion Quarterly, 82*(4), 641-660.

Daikeler, J., Bosnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513-539.

Deming, W. E. (1950). *Some theory of sampling*. Wiley.

Dillman, D. A. (2018, October 2). *How Web-Push Surveys are Changing Survey Methodology* [Seminar presentation]. NatCen/ESS/City University methods seminar, City University of London, London, United Kingdom.

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

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

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

Fordyce, E., Bilgen, I., & Stern, M. J. (2020, May 11-14). *The Use of Synchronous and Asynchronous Text Messaging During Survey Recruitment and Screening* [Conference presentation]. 75th annual conference of the American Association for Public Opinion Research, Virtual.

Fuchs, M. (2008). Total Survey Error (TSE). In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 896-902). Sage.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475-495.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *The Public Opinion Quarterly, 64*(3), 299-308.

Groves, R. M. (2004). *Survey errors and survey* costs (Vol. 536). John Wiley & Sons.

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

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, *74*(5), 849-879.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861-871.

Hade, E. M., & Lemeshow, S. (2008). Probability sample. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 621-623). Sage.

Hall, J. (2008). Area probability sample. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 33-36). Sage.

Jans, M. (2008). Mode effects. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 475-480). Sage.

Kalaian, S. A., & Kasim, R. M. (2008). Longitudinal studies. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 439-440). Sage.

Keeter, S. (2019). *Growing and Improving Pew Research Center's American Trends Panel.* Pew Research Center.

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

Keusch, F. (2015). Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, *65*(3), 183-216.

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

Kocar, S. (2020). Attrition. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), *SAGE Research Methods Foundations*. https://www.doi.org/10.4135/9781526421036926973

Kongsgard, H. W., Syversen, T., & Krokstad, S. (2014). SMS phone surveys and mass-messaging: promises and pitfalls. *Epidemiology: Open Access*, *4*(4). http://dx.doi.org/10.4172/2161-1165.1000177

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*(3), 213-236.

Krosnick, J. A., Ackermann, A., Malka, A., Yeager, D., Sakshaug, J., Tourangeau, R., DeBell, M., & Turakhia, C. (2009, August). *Creating the Face-to-Face Recruited Internet Survey Platform (FFRISP)* [Workshop presentation]. Third Annual Workshop on Measurement and Experimentation with Internet Panels, Santpoort, The Netherlands.

Lavrakas, P. J. (2013). Presidential Address: Applying a Total Error Perspective for Improving Research Quality in the Social, Behavioral, and Marketing Sciences, *Public Opinion Quarterly*, 77(3), 831–850. https://doi.org/10.1093/poq/nft033

Lavrakas, P. J., Kaczmirek, L., Myers, P., & Pennay, D. W. (2018, May 16-19). *An experiment to reduce noncompliance in an online probability-based panel: the challenges of dozer, sleeper, comatose, and backout panelists* [Conference presentation]. 73rd annual conference of the American Association for Public Opinion Research, Denver, United States of America.

Lewis-Beck, M., Bryman, A. E., & Liao, T. F. (2004). *The Sage encyclopedia of social science research methods*. Sage Publications.

Link, M. W. (2008a). Address-based sampling. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 7-8). Sage.

Link, M. W. (2008b). Mixed-mode. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 472-475). Sage.

Lugtig, P. (2014). Panel attrition: separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research, 43*(4), 699-723.

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

Malhotra, N., & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. *Political Analysis*, *15*(3), 286-323.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For weighting online opt-in samples, what matters most.* Pew Research Center.

Nathan, G. (2008). Internet Surveys. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 356-359). Sage.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*(201), 101-116.

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

Phillips, B., Dove, C., Myers, P., & Neiger, D. (2021, March 4-5). *Expansion of an Australian probability-based online panel using ABS, IVR and SMS push-to-web* [Conference presentation]. CIPHER 2021 Conference, Virtual.

Rao, K., Kaminska, O., & McCutcheon, A. L. (2010). Recruiting probability samples for a multi-mode research panel with internet and mail components. *Public Opinion Quarterly*, *74*(1), 68-84.

Singer, E. (2011). Toward a benefit-cost theory of survey participation: Evidence, further tests, and implications. *Journal of Official Statistics*, *27*(2), 379–392.

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

Tourangeau, R., Groves, R. M., Kennedy, C., & Yan, T. (2009). The presentation of a web survey, nonresponse and measurement error among members of web panel. *Journal of Official Statistics*, *25*(3), 299-321.

Wright, K. B. (2005). Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, *10*(3), JCMC1034, https://doi.org/10.1111/j.1083-6101.2005.tb00259.x

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

# Chapter 2 Data and methods

As previously explained, this thesis is based on interdisciplinary, multimethod and multi-mode methodological research, and a variety of data sources and analytical methods are applied. In the end, they are also assessed to determine their suitability for the study of survey methods. To analyze different phenomena in online panel research, data sources such as quantitative online panel survey data, data from survey experiments, paradata, in-depth interview data, and synthesized evidence can be used with analytical methods such as multivariate analysis, panel data analysis, qualitative data analysis, benchmarking analysis, or meta-analysis.

The research included in this thesis is data-driven and evidence-based, and was dependent on accessibility of secondary data and a budget available to collect primary data to address additional gaps in knowledge. In the chapters in which quantitative data was analyzed, the hypothetico-deductive model as a theory-testing method was used, since the hypothesis and research questions were based on theory in (online panel) survey methodology (see Chapter 1). The study for which qualitative data was predominantly collected and analyzed (presented in Chapter 4) was based on grounded theory and as such applied inductive reasoning to extend social-psychological theory to an online panel setting. While I identified a number of other research questions, answering some of them would require access to more secondary online panel data and additional primary data collection (for example, with survey experiments).

The data, methodological and analytical approaches were carefully selected to target phenomena specific to online panel research. Cross-sectional data were used to study survey errors consistently present in panel surveys, longitudinal data and panel data analysis were applied to study phenomena specific to the longitudinal component of surveys, and survey experiments were carried out to target and compare particular methodological solutions. On the other hand, we collected and analyzed qualitative data for theory construction/extension purposes, and we used a meta-analytical approach to both synthesize evidence and to generalize findings worldwide. In the following paragraphs, I describe all data sources and analytical methods used in this thesis in more detail.

## 2.1 Quantitative data

There are various types of data and different classifications can be used: verbal, numeric and graphic, textual, categorical and ranked, quantitative and qualitative, and so on. The distinction between qualitative and quantitative data is not absolutely determinative, since both quantitative and qualitative analytical methods could be used with many kinds of data (Vogt et al. 2014). In this thesis, the majority of the chapters are based on quantitative evidence. Thus, quantitative analytical methods

used to address the most important research gaps in online (panel) survey research are applied to the following quantitative data sources: cross-sectional online panel survey data, longitudinal online panel survey data, data from survey experiments, paradata, personality assessment data, and data with nationally representative benchmarks. These data sources are presented below.

### 2.1.1 Cross-sectional online panel survey data

Cross-sectional data are collected from survey respondents at one point in time, and those types of studies normally collect self-reported data (attitudes, beliefs, opinions, values), while time is not considered one of the study variables but is rather assumed to have random effect (Liu 2008). Cross-sectional online panel survey data are topical survey data collected from online panellists in a particular survey wave, and are as such different to longitudinal data and paradata.

In addition to topical research, they can be analyzed for different purposes in methodological research, including (but not limited to): (1) in the study on measurement mode effect (measurement error), (2) to study the necessity for mixed-mode data collection in online panel research (coverage error), (3) to study representation bias due to attrition over time in a time-series fashion (nonresponse error), and (4) to assess accuracy of different samples (total survey error). In this thesis, cross-sectional online panel survey data are analyzed to study measurement, coverage, and total survey error. All cross-sectional survey data sources from this thesis, including non-panel surveys, are presented in Table 2.1.

Table 2.1: Cross-sectional online panel survey data sources used in this thesis

| Study | Surveys | Survey mode(s) | Analytical methods | Survey errors (chapter) |
|---|---|---|---|---|
| Life in Australia™ surveys | Online Panels Benchmarking Study 2017 Replication (W2)[18], Waves 1, 3, 10, 19, 21, 22[19] | Online, telephone | Univariate, bivariate, multivariate analysis; benchmarking analysis, post-survey adjustment methods (weighting, matching methods), data simulation | Coverage error (Chapter 6) Measurement error (Chapter 7) Representation error (Chapter 9) |
| Online Panels Benchmarking Study 2015[20] | Dual Frame RDD survey | Telephone | Univariate, bivariate, multivariate analysis; benchmarking analysis, post-survey adjustment methods (weighting, matching methods) | Measurement error (Chapter 7) Representation error (Chapter 9) |
| | Address-based sampling survey | Telephone, postal, online | | |
| | RDD 'piggybacking' | Telephone, postal, online | | |
| | 5 nonprobability online panel samples | Online | | |

With cross-sectional online panel survey data, a variety of statistical methods can be used to study approached to dealing with survey errors. For example, in Chapter 6, we used univariate, bivariate and

[18] DOI: 10.26193/YF8AF1
[19] DOIs: 10.26193/JFWRPI, 10.26193/EL5WHN, 10.26193/7OP0TI, 10.26193/LEWZYX, 10.26193/XHORAI, 10.26193/IRSDS8
[20] DOI: 10.4225/87/FSOYQI

multivariate analysis, as well as data simulation in R and benchmarking analysis, to determine the extent of undercoverage bias if offline population is excluded from probability-based online panel research. On the other hand, we also used a variety of weighting and matching methods to control for self-selection effect in the study on mode effect (Chapter 7), and to mitigate representation bias in nonprobability samples (Chapter 9).

### 2.1.2 Longitudinal online panel survey data

In contrast to cross-sectional online panel survey data, longitudinal data from online panels provide a better basis for measuring change over time since we can control the impact of omitted variables and generate more accurate predictions (Hsiao 2007). As such, they can also be used in methodological research to study survey error-related change over time. Online panels provide opportunities to collect data on the same concepts and topics from the same respondents in different points in time, but that could also be a source of survey error specific to longitudinal research. One of those sources is panel conditioning, which can increase measurement error. If switching modes over time is allowed, potential measurement 'panel' mode effect should be taken into consideration. In this thesis, we used the same survey items from different online panel survey waves, listed in Table 2.2.

Table 2.2: Longitudinal online panel survey data sources used in this thesis

| **Study** | **Surveys** | **Repeated survey items** | **Analytical methods** | **Survey errors (chapter)** |
|---|---|---|---|---|
| Life in Australia™ surveys | Waves 1, 3, 7, 10, 14[21] | Satisfaction with the way country is heading, the most and the second most important problem facing the country, party preference | Univariate, bivariate, multivariate analysis (pooled); panel data analysis (fixed- and random-effect) | Measurement error (Chapter 8) |

Besides 'static' multivariate analysis (e.g., pooled logit regression analysis), panel data regression modeling can be used with panel/longitudinal data. To control for the impact of unobserved heterogeneity and obtain valid inference on structural parameters, different models can be used, namely random-effect, fixed-effect and mixed-effect models (Hsiao 2007). What model is the most suitable in a given situation is sometimes challenging to determine, but the Hausman test for endogeneity (Hausman 1978) can be applied in choosing the right model to control for unobserved heterogeneity.

[21] DOI: 10.26193/ZXF0SQ

### 2.1.3 Data from survey experiments

In addition to the analysis of topical cross-sectional panel survey data, survey methodological solutions can be tested with so-called survey experiments. Also known as 'population-based survey experiments', they are administered to a representative population sample, and they use survey sampling methods to collect data from experimental subjects which are randomly assigned to conditions of the experiment (Mutz 2011, p. 2).

Survey experiments are a method regularly used in disciplines such as political science, sociology, psychology, economics and so on (Mutz 2011, p. 5), but they can also be used for other purposes in (online panel) survey methodology. An example of that is the study of mode effect in mixed-mode online panels (measurement error, see Dennis et al. 2005). In this thesis, I used quantitative survey data from a survey experiment briefly described in Table 2.3 to study response maximization approaches of a fairly new approach to probability-based online data collection. Since I was interested in methodological solutions, I did not use topical survey data (see Chapter 11, total survey error), but rather created a dataset with survey response and response maximization approaches as attributes (for more detail, see Chapter 10, Subsection 'Survey experiment').

Table 2.3: Survey experiment data used in this thesis

| **Study** | **Surveys** | **Mode** | **Experimental groups** | **Analytical methods** | **Survey errors (chapter)** |
|---|---|---|---|---|---|
| RDD SMS Web-push Survey project | Survey on Wellbeing, Health and Life in general 2020[22] | online | Day and time of invitation, type of SMS invitation, SMS content, incentives, reminders | Bivariate, multivariate analysis; benchmarking analysis | Nonresponse error (Chapter 10) |

Similar to the statistical methods and techniques applied to other types of cross-sectional survey data, a number of different analytical methods can be used with survey experiment data. In Chapter 10, bivariate, multivariate and benchmarking analysis was carried out to answer the proposed research questions.

### 2.1.4 Paradata

Paradata are defined as data containing information about data collection process and are also known as process data. They can be used to either control the data collection process, or to evaluate the quality of the collected survey data (Heerwegh 2008) – for example, to study speeding as a source of measurement error. There are two broad types of paradata in web surveys: questionnaire navigation paradata, and device-type paradata. In online panel surveys, there is a separate class of paradata,

[22] Access to be provided by the Australian Data Archive in the future.

which includes information such as survey topics, number of panel surveys completed, last survey completed, or number of invitations to panel surveys (Callegaro 2013).

In this thesis, we focus on paradata specific for online panels, so-called online panel paradata, which measure panellist panel participation-related behavior throughout the panel lifecycle. We will try to examine their practical use in panel management. The paradata used in Chapter 5 and presented in Table 2.4 are from the only Australian probability-based online panel.

Table 2.4: Online panel paradata sources used in this thesis

| Study | Time series | Online panel paradata items | Analytical methods | Survey errors (chapter) |
|---|---|---|---|---|
| Life in Australia™ surveys | Waves 1–30 (December 2016 – August 2019) | Survey invitation, survey outcome (interview, refusal, non-contact, non-refusal, other), charity donations | Univariate, bivariate, multivariate analysis (pooled); panel data analysis (fixed- and random-effect) | Nonresponse error (Chapter 5) |

Just as with longitudinal online panel survey data, panel data regression modeling can be used with online panel paradata, as the data are easily transformed into a panel form. Also, we can derive panel variables, such as completion rate by certain wave or consecutive surveys with nonresponse, and use them in panel data models as outcome variables and regressors.

### 2.1.5 Personality assessment data

Personality assessment is a set of procedures for studying and comparing personal characteristics and capacities of people. The objective of personality assessment is to integrate personality-based information into conclusions and recommendations. Data on personality characteristics are collected in disciplines such as medicine and health care, education, forensic science, and in organizational settings. In each of those disciplines, they are collected for different purposes; for example, in educational settings to identify the need for counselling and special services, and in organizational settings to evaluate candidates for promotion or employment (Weiner & Greene 2017).

There are a variety of personality assessment tests, and some of the most commonly used in organizational and similar settings are Myers-Briggs Type Indicator, Big 5 Personality Traits (also known as OCEAN or NEO PI), Open Four Temperament Scales (O4TS), and DiSC test. In survey methodology, personality assessment testing has previously been carried out to study the relationship between respondents' personalities and nonresponse, and the most popular test seemed to be the Big 5 Personality Traits (Goldberg 1992). To assess another personality test, which is predominantly used in industry to determine future behavior as a result of different personalities (Jones & Hartley 2013), we instead experimentally administered DiSC test on a smaller sample of panellists who also participated

in qualitative interviews (see Chapter 4). The details of this personality assessment are available in Table 2.5.

Table 2.5: Personality assessment data

| Personality assessment | Dimensions | Mode | Sample size | Analytical methods | Survey errors (chapter) |
|---|---|---|---|---|---|
| DiSC | Dominance, Influence, Steadiness, Compliance | Online | n=14 | Univariate, bivariate analysis | Nonresponse error (Chapter 4) |

### 2.1.6 Data sources with nationally representative benchmarks

Estimating total survey error is often challenging when not knowing the 'truth', which is in turn possible by having access to the most accurate estimates from representative high-quality survey data sources (including censuses). For example, studying measurement mode effect, we can observe statistically significant differences between the compared modes. However, we cannot know for sure what estimates are more accurate without comparing them to so-called benchmarks, that is, estimates representative for the studied population.

In this thesis, we use a number of data sources with nationally representative benchmarks, including Australian Census 2016, National Drug Strategy Household Survey 2019, National Health Survey 2017–18, and so on (for a complete list, please see Appendix 11). In three different chapters, we used those benchmarks to study the necessity for mixed-mode data collection in online panel research (coverage error, Chapter 6), comparing post-survey adjustment methods to mitigate bias in nonprobability samples (total survey error, Chapter 9), and investigating the relative accuracy of the RDD SMS web-push survey (total survey error, Chapter 11).

## 2.2 Qualitative data

Since survey methodology and survey statistics as research disciplines are predominantly based on quantitative data collection and quantitative evidence, more quantitative than qualitative data are analyzed in this thesis to study some key issues in online panel research. However, we purposely used qualitative methods to address certain research gaps which could not be investigated using quantitative data alone. While various types of qualitative methods exist, such as in-depth interviews, focus groups, ethnography, analysis of documents, discourse analysis and visual data analysis (Silverman 2015), our targeted data collection was limited to two qualitative approaches: analysis of answers to an open-ended question (so-called verbatims) at the end of a probability-based online panel survey, and in-depth interviews with panellists from the same online panel (see Chapter 4). The

methodology of that data collection and qualitative data analysis is briefly discussed in the following paragraphs.

### 2.2.1 Qualitative in-depth interviews

Qualitative in-depth interviews are one of the most popular research methods. There are three types of interviews: structured, semi-structured, and open-ended. While structured interviews are quantitative in nature, semi-structure and open-ended interviews conducted on smaller samples are more common in qualitative research. Their advantages are to collect 'richer' data on certain phenomena, due to an ability to have direct access to what people do in real life, as well as economic data collection in terms of time and resources (Silverman 2015).

In Chapter 4, we analyzed qualitative data collected with semi-structured in-depth interviews. They were conducted on a small sample (n=15) of panellists of the only probability-based online panel from Australia. The results of this analysis were later combined with results from the analysis of open-ended survey questions, as well as those on personality assessment (see Subsection 2.1.5). Generally speaking, the study presented in Chapter 4 was a mixed-method study. Some key characteristics of the qualitative in-depth interview data and the analytical approach are presented in Table 2.6.

Table 2.6: Open-ended survey data and qualitative data analysis

| **Study** | **Qualitative method** | **Sample** | **Analytical approach** | **Survey errors (chapter)** |
|---|---|---|---|---|
| Social-psychological aspects of online panel participation | Semi-structured in-depth interviews | Life in Australia™ panellists: frequent-respondents (n=6), stopped-responding panellists (n=5), nonrespondents (n=14) | Interview transcription, coding of transcripts, synthesis of codes, interpretation and generalisation | Nonresponse error (Chapter 4) |

### 2.2.2 Open-ended survey questions

Open-ended questions are often included in surveys, normally at the end of the questionnaires, for respondents to make additional comments or to respond to questions for which there are not pre-categorized responses. After coding is done in a similar way to coding of in-depth interviews, coded responses or scores can be used as quantitative survey data (demographics or substantive items) (Bazeley 2013), and analyzed with quantitative analysis methods and techniques.

In Chapter 4, we analyzed open-ended survey data to present evidence on motivational factors for online panel members to join the panel and participate in panel surveys. The main advantages of this qualitative approach were the ability to collect in-depth data without having a list of pre-categorized responses (in a closed-ended multi-answer question), and to collect data from a large sample

(n=1500+) of panellists, something that cannot be done in practice with in-depth interviews. The characteristics of these open-ended survey data are presented in Table 2.7.

Table 2.7: Open-ended survey data and qualitative data analysis

| Study | Wording of open-ended question | Sample size | Analytical approach | Survey errors (chapter) |
|---|---|---|---|---|
| Life in Australia™ Wave 28[23] | We would like to understand why you chose to be part of Life in Australia™ and what, if anything, you value about being part of it.<br>What does being part of Life in Australia™ mean to you or what motivates you to participate in the surveys? | n=1,557 | Coding of verbatims, coding consistency analysis, factor analysis of codes (dimension reduction), univariate analysis | Nonresponse error (Chapter 4) |

## 2.3 Data from systematic reviews or research syntheses

Systematic review or research synthesis are types of literature reviews. In practice, there is a very little difference between systematic reviews and research syntheses – systematic reviews are carried out for evidence-based practical applications, while research syntheses are closer to basic research and not necessarily tied to practical applications. Meta-analyses as a subtype are quantitative evidence-based reviews or syntheses (Vogt et al. 2014). Although it is quantitative in nature, meta-analysis is methodologically quite different to other quantitative analytical approaches and as such deserves to be presented in more detail separately.

Meta-analysis is the statistical synthesis of the data from primary studies and is used to address similar questions to those from primary studies. In a meta-analysis, each of those studies receives a weight, and this weighted statistical analysis should provide an objective and replicable framework. Meta-analyses are used in many fields of research and for variety of reasons, including to synthesize evidence and to explain phenomena with covariates in a meta-regression as a type of a meta-analysis (Borenstein et al. 2011).

In this thesis, meta-analysis including meta-regression is used to study recruitment rates in probability-based online panels (nonresponse error). The main advantage of this approach is the ability to study online panel recruitment outcomes worldwide and is, unlike the rest of the data analyzed in this thesis, not limited to Australia. However, synthesizing evidence from different studies, or better to say different online panels, does not come without challenges. This is described in detail in Chapter 3, while Table 2.8 presents key characteristics of the synthesized evidence in a form of a quantitative data file.

[23] DOI: 10.26193/SYYFCS (survey metadata only).

Table 2.8: Data for meta-analysis

| Data file | Studies | Effect size | Moderators | Analytical methods | Survey errors (chapter) |
|---|---|---|---|---|---|
| Data on recruitment rates in probability-based online panel research | 23 online panels, 95 recruitment rates | Single group summaries (recruitment rate) | Year of recruitment, type of recruitment incentives, total guaranteed incentives amount, recruitment mode, multiple-mode recruitment, end-of-survey recruitment, North American panel indicator | Average single group summaries, time-series analysis of single group summaries, meta-regression | Nonresponse error (Chapter 3) |

## 2.4 References


Bazeley, P. (2013). *Qualitative Data Analysis: Practical Strategies* (1st ed.). SAGE Publications Ltd.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Callegaro, M. (2013). Paradata in web surveys. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information Improving surveys with paradata: Analytic uses of process information* (pp. 261-279). Wiley.

Dennis, J. M., 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. *60th Annual Conference of the American Association for Public Opinion Research*, 1-26.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological assessment*, *4*(1), 26-42.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, *46*(6), 1251–1271.

Heerwegh, D. (2008). Paradata. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 574-575). Sage.

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *Test, 16*(1), 1-22.

Jones, C. S., & Hartley, N. T. (2013). Comparing correlations between four-quadrant and five-factor personality assessments. *American Journal of Business Education*, *6*(4), 459-470.

Liu, C. (2008). Cross-sectional Data. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 170-171). Sage.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton University Press.

Silverman, D. (2015). *Interpreting qualitative data.* Sage.

Vogt, W. P., Gardner, D. C., Haeffele, L. M., & Vogt, E. R. (2014). *Selecting the right analyses for your data: Quantitative, qualitative, and mixed methods*. Guilford Publications.

Weiner, I. B., & Greene, R. L. (2017). *Handbook of personality assessment*. John Wiley & Sons.

# Chapter 3 A meta-analysis on worldwide recruitment rates in 23 probability-based online panels, between 2007–2019

## 3.1 Introduction

Online panels are growing in numbers and represent an alternative to more traditional survey modes and cross-sectional survey research in general. While most online panels are nonprobability-based panels, so-called volunteer, access or opt-in panels, there are a number of probability-based online panels, which tend to have fewer members (Baker et al. 2010). Those panels recruit their members using "offline" methodologies such as random digit dialing or address-based sampling (Callegaro et al. 2014, p. 7), hoping to overcome the issue of low external validity that nonprobability panels face (Lugtig et al. 2014), but there are various other differences between probability online panels in how their members are recruited (Blom et al. 2016). The decisions on the recruitment approach have to balance data quality with costs and time-efficiency. The rise of online research has been partly due to panels being time and cost effective, and due to increasing nonresponse in other modes (Baker et al. 2010). However, a number of studies have shown that probability-based panels also face low response in the recruitment phase (e.g., Blom et al. 2016; Lugtig et al. 2014; Rao et al. 2010), which is consistent with an issue of lower response rates in web surveys compared to other survey modes (most recently Daikeler et al. 2020). While it has been reported that the association between unit nonresponse and representation bias is weak at best (Groves & Peytcheva 2008) and response rates as an indicator of panel quality are only the tip of the iceberg when it comes to representation (Callegaro et al. 2014), other studies with online panels have shown that nonresponse bias exists as early as at the recruitment stage and cannot be entirely eliminated even if non-internet households are provided with the required technology (Lugtig et al. 2014).

Hence, understanding what affects recruitment rates to optimize recruitment would have many positive implications, from lowering costs to mitigating overall representation bias. This meta-analysis focuses on unit (non)response, as one of sources of survey error (see Total Survey Error framework, Groves et al. 2009), and determinants of that type of nonresponse in recruitment, as the first phase in the online panel lifecyle.

## 3.2 Background

As of 2020, there has not been a single meta-analysis on survey errors in probability-based online panel research, which seems to be a general issue in survey methodology; Čehovin et al. (2018) concluded that in survey methodology there are fewer meta-analysis than in related disciplines like psychology. Cross-country comparative research on recruitment in probability-based research has been limited –

to the best of our knowledge, only Blom et al. (2016) and Greaves (2017)[24] presented detailed comparative studies. In contrast, there are a number of meta-analyses on response rates in online surveys (Cook et al. 2000; Daikeler et al. 2020; Manfreda et al. 2008; Mavletova & Couper 2015; Shih & Fan 2008) and systematic analyses of factors affecting response rates in web surveys (e.g., Fan & Yan 2010). This study will be built on the methodology of those studies, and will exploit the advantages of a meta-analytical approach, which is to synthesize and to generalize.

Response rates in probability-based online panel research can be dissected into recruitment rates (RECR), profile rates (PROR), completion rates (COMR), and retention rates (RETR), while the product of these four are defined as cumulative response rates (CUMR) (Callegaro & DiSogra 2008). Baker et al. (2010) reported how response rates in probability-based online panel research seem to be much higher than in nonprobability-based opt-in panels[25]. Yet, they are far from being predictable and there are little generalizable findings on the best predictors of response rates. In the most comprehensive comparative study so far, Blom et al. (2016) described and compared four probability-based panels, including their offline recruitment procedures, inclusion of the offline population, and recruitment response rates. They reported recruitment rates as a product of recruitment and registration rates. This product of rates is also known as overall recruitment rate or ORR (USC Dornsife Center for Economic and Social Research, n.d.). The overall recruitment rates were between 18.1% (AAPOR RR4, German Internet Panel) and 48.3% (AAPOR RR3, LISS)[26], while the analyzed panels substantially differed in mode of offline recruitment, types of incentives, and invitations (Blom et al. 2016). Kaczmirek et al. (2019) collected and described overall recruitment rates for 12 probability-based online panels and reported recruitment rates (RETR x PROR) between 6.5% (American Trends Panel) and 54.3% (Social Science Research Institute panel).

Moreover, factors affecting recruitment rates have been studied with survey/recruitment experiments to determine the most optimal approach to recruitment, but with mixed evidence and without a cross-country comparative component. Rao et al. (2010, Gallup Panel) concluded that mail recruitment nets a higher panel response than telephone recruitment and that advance letter, incentive, and telephone follow-up conditions positively influence recruitment rates. Scherpenzeel and Toepoel (2012, LISS) reported no differences in recruitment rates between telephone and face-to-face recruitment, and between different content of the advance letter. On the other hand, they also reported effectiveness of unconditional/prepaid incentives, which was consistent with findings from Blom et al. (2015,

[24] Blom et al. (2016) carried out a comprehensive but European-centred cross-national comparative methodological study of four European probability-based panels, and Greaves (2017) prepared a systematic global review of probability-based online panels as a feasibility study for setting up a national online panel.
[25] Sample yields are reported instead in opt-in panels.
[26] In our meta-analysis, these kinds of differences in reporting of recruitment rates (AAPOR RR1, RR2, RR3 or RR4 (The American Association for Public Opinion Research 2016)) are being carefully addressed in the analysis section.

German Internet Panel) and DiSogra et al. (2009, KnowledgePanel), who also reported higher response rates if increasing the incentives amount ($1 to $5), while they did not find any effects of sending an advance postcard on recruitment rates.

With this background in mind, we would like to answer the following research questions (RQ):

*RQ1: What overall recruitment rates can be expected in probability-based online panel research?*

*RQ2: How do recruitment mode, type of incentives, incentives amount, and other recruitment strategies impact the variation in the overall recruitment rate in probability-based online panel research?*

First, we would like to calculate both the average overall recruitment rates, and report minimum and maximum recruitment rates. We would also like to investigate changes in average overall recruitment rates over time, as the literature suggests that response rates have declined steadily over years, year of data collection is a strong predictor of response, and the phenomena are not specific to a particular survey mode (Couper 2017; Stedman et al. 2019). Knowing what overall recruitment rates to expect has important cost, time, and sample size implications for providers of newly established online panels.

Second, we would like to present evidence on the most optimal methdological solutions for recruitment to probability-based online panels between 2007 and 2019. So far, recruitment rates have been investigated with experimental studies within single online panels which tested different recruitment strategies but offered lesser generalizability (e.g., DiSogra et al. 2009; Rao et al. 2010). Our study will be the first one presenting evidence not limited to a specific country context. These findings could be used by any provider of probability-based online panel (active, new) interested in optimizing recruitment to their panel from both survey error and cost perspectives.

## 3.3 Method

In this study, we will synthesize evidence from various sources, including scientific articles, methodological reports and panel websites, which reported or studied recruitment rates in probability-based online panel surveys. Based on the inclusion criteria (see below) we defined what types of panels are eligible for our study and conducted an online search for probability-based online panels (globally), for both active and no longer active panels. In the retrieval stage, we tried to locate recruitment rates for as many recruitment waves of as many online panels as possible. We later selected moderators for meta-regressions and coded the data. In the last stage, we analyzed the data and interpreted the results. These steps are explained in more details in the next paragraphs.

### 3.3.1 Meta-analytic sample, inclusion criteria and retrieval

Our initial meta-analytic sample of probability-based online panels was made out of all panels for which we found at least some basic information online, either from their documentation, websites, or

a mention in any scientific publication/documentation including conference or workshop presentations. Nonprobability-based panels, also known as volunteer, opt-in or access panels, were excluded from the retrieval, coding or analysis due to their predominantly commercial nature, non-probabilistic sampling, and recruitment approaches and strategies very different to those in probability-based research. Out of general population panels, specialty panels, proprietary panels, and online community panels (Callegaro et al. 2015, p. 206), we included only general population panels in this study. For example, we did not include specialty panels focusing on a particular population subgroup, such as those with the population defined as 50 years of age and older; thus, Singapore Life Panel (Singapore Management University, n.d.) and a specialty panel managed by AARP (n.d.) were excluded from the target sample.

Our focus was on studying overall recruitment rates (ORR), which measure the final recruitment outcome, i.e., registration of recruited respondents as panel members. Probability-based online panels differ substantially in recruitment steps, phases and strategies, and the effectiveness of recruitment is captured best in ORR. This will be explained in more detail in subsection 'Computation of effect sizes'. In the coding and analysis stages, we included only those probability-based online panels (see Table 3.3 in the Appendix 3), and their recruitment waves/years as units, for which we could locate or obtain overall recruitment rates, and for which there was sufficient methodological explanation of recruitment strategies for moderator selection and coding.

Due to the specifics of our meta-analitical sample, the retrieval phase differed from those from more traditional meta-analyses – we firstly had to identify online panels to later find information on recruitment events as 'studies' and to extract recruitment rates as effect sizes. To identify worldwide online panels, we carried out an online search for probability-based panels in Google and Google Scholar search engines. We used the following keywords: "probability-based online panel", "online panel", "probability-based web panel", "web panel", to identify as many as possible, later excluding all panels which did not apply a probabilistic approach to sampling. We also reviewed relevant articles, conference presentations, feasibility studies and publicly available lists of probability-based online panels, such as those from Blom et al. (2016), Greaves (2017), Kaczmirek et al. (2019), and WebSM (n.d.).

After the identification of all probability-based panels that matched the definition, i.e., general population panels either active or no longer active, we located any publicly available recruitment (RECR) and profile (PROR) rates by reviewing the following sources of information: online panel websites, scientific articles, methodological reports, project reports, panel promotional material, and presentations at events such as scientific conferences or workshops. In case no rates had been made publicly available, or if they had been calculated using an alternative methodology (e.g., household

instead of individual person recruitment rates), we contacted the panels' management and asked them to assist us by sharing their overall recruitment rates and/or to provide more information and clarifications. Only recruitment rates calculated consistently with our criteria (see the 'Computation of effect sizes' subsection for more information) and for which the selected moderators could be coded (see the 'Moderators, predictors and coding' subsection for more information) were included in the end. The information was received/retrieved between July 2019 and December 2020, and recruitment rates for recruitment waves finished by the end of 2019 are included in this study. The review of all identified probability-based panels (n=28) and the availability of their recruitment rates for this meta-analysis is presented in Table 3.3 in the Appendix 3.

### 3.3.2 Computation of effect sizes

Single group summaries are meta-analytical effect sizes[27] in this study, and we synthesized and analyzed ORR as single groups summaries. We used ORR as a product of RECR and PROR (see Equation 3.1 below), consistently with how Understanding America Study (USC Dornsife Center for Economic and Social Research, n.d.) use it in their reporting.

$$ORR = RETR * PROR \qquad (3.1)$$

In our meta-analysis models, we purposely did not analyze recruitment rates (RETR) and profile rates (PROR) separately, as there were notable differences in how organizations managing online panels approached recruitment. There are three predominant strategies used in practice: (1) one stage recruitment with recruitment and profiling carried out in one interview, (2) one stage recruitment without subsequent profiling, and (3) two stage recruitment with separate recruitment and profile interviews. This becomes even more complex in the case of end-of-survey recruitment, so-called piggybacking recruitment. Consequently, for some panels we could obtain the final panel registration rates only, as RETR=ORR, and not separate RETR and PROR.

Secondly, to guarantee comparability of ORRs, we had to make sure that all of them were calculated using the same formula. For example, Blom et al. (2016) presented recruitment rates from four probability-based panels using three different response rate formulas: AAPOR RR3 (LISS, ELIPSS), AAPOR RR4 (GIP), and AAPOR RR5 (GESIS Panel). The difference between the different calculations is in the treatment of cases of unknown eligibility, exhibited in the *e* value, and in the treatment of partial interviews. To solve this problem, all ORR from our study were calculated using AAPOR RR1 equation with all cases with unknown eligibility treated as eligible (see Equation 3.2 below). This adjustment

[27] Borenstein et al. (2011) list single group summaries as a type of effect sizes, but they also explain that they are, strictly speaking, not effect sizes as effects imply relationships. In this article, we use the 'effect sizes' term as it is more common for meta-analyses.

substantially reduces the effect of the quality of sampling frames in terms of coverage, which can be very country specific. The adjustment increases the robustness of cross-country comparative results.

$$ORR = \frac{registered\ panellists}{(all\ sampled - ineligible)} \qquad (3.2)$$

Lastly, we decided to include all individual waves of recruitment from all eligible panels to use all available information. If there was an experimental study embedded within a particular recruitment wave, we included individual ORRs for all relevant experimental groups separately (e.g., ORR for $0 incentives group, ORR for $1 incentives group, and ORR for $5 incentives group in Gallup 2007 recruitment round (Rao et al. 2010)). However, since certain panels recruit new members continuously throughout the year with no methodological changes over time, the total annual ORR were calculated for those panels to enable comparison with panels carrying out a shorter time-period ad-hoc recruitment, refreshment, or replenishment.

### 3.3.3 Moderators, predictors and coding

As we will carry out meta-regression analysis to answer the research question on the impact of recruitment strategies on overall recruitment rates, we selected a number of moderators as predictors of effects sizes as outcomes in our meta-regression models. The selection of moderators was based both on the theoretical review and on the availability of information, either published online or provided by the organizations managing probability-based online panels. The methodological detail shared by the panel organizations in their materials and publications differed substantially between the organizations, thus we had to carefully review the available information and find comparable moderators. We were also careful not to select too many moderators as predictors of ORR, as that could lead to overfitting of the models. The literature suggests it is preferable to have an appropriately large ratio of studies to covariates, such as 10 or more subjects per covariate, although there are no hard rules (Borenstein et al. 2011). In the end, the following moderators were selected:

- *Year of recruitment:* while some online panels shared annual ORRs, we found information on ad-hoc recruitment waves for other panels and if their recruitment extended into the first few months of the following year, we recorded the recruitment *start year*
- *Type of recruitment incentives – No incentives, Conditional monetary, Unconditional monetary, Prize draw/lottery*: this multiple answer variable was recoded into a set of dummies, as some panels used a mixture of conditional and unconditional monetary incentives

- *Total guaranteed recruitment incentives amount*: this was the sum that an invited respondent would have received for registering as a panel member, and any amount in other currency was recoded to US dollars[28]; for prize draw/lottery (tickets), the coded amount was $0
- *Recruitment mode – Interactive voice response (IVR), telephone, face-to-face (F2F), mail/postal*: this multiple answer variable was later recoded into a set of dummies, as some panels used a mixture of modes to recruit panellists
- *Multiple-mode contact/recruitment*: this binary variable (0=no, 1=yes) was created to identify either: (1) the same panellists being approached using different recruitment modes (recruitment via mail followed by F2F recruitment in case of non-contact), or (2) the same panellists being contacted/reminded using different channels (e.g., advance letter-mail combined with F2F recruitment mode or mail recruitment combined with follow-up telephone calls)
- *End of survey recruitment*: also known as piggybacking recruitment, this binary variable (0=no, 1=yes) was created to distinguish between 'standard' recruitment, i.e., respondents were primarily contacted to be recruited to the panel, and recruitment at the end of a different cross-sectional survey, i.e., respondents were primarily contacted to participate in a different survey and had to respond first before any attempts to recruit them to a panel were made
- *North American panel:* the preliminary analysis has shown a substantial difference in mean ORRs between online panels from North America and panels from the rest of the world; by including this moderator, we will attempt to control for some continental specifics.

### 3.3.4 Data analysis and weighting

To analyze the data, we used the statistical environment R and specifically the R package for meta-analyses called *metafor* (Viechtbauer 2010). We used the package to calculate average effect size, cumulative effect sizes, prepare visualisations of results (forest plots), and conduct meta-regressions. We carried out outlier detection analysis (influence diagnostics) such as standardized residuals, DFFITS and Cook's distance, to identify outliers to be excluded from meta-regression models. Outliers in our meta-regressions can be present due to various reasons: incorrectly calculated rates, overreporting of ORR, and there might be 'outlier' countries in terms of their privacy and confidentiality, information-sharing and survey-participation culture. All those potential sources of discrepancies make carrying out outlier detection even more important in efforts to mitigate bias.

[28] We extracted and calculated recruitment incentives amount from the available materials to the best of our ability; we used the exchange rates from 1 December 2020 and rounded amounts to USD values with no decimals (no inflation adjustment). We acknowledge the fact that exchange rates fluctuated over time, and that the same dollar amount does not represent equal value to respondents from different analyzed countries.

We assume that the true effect size, i.e., ORR, varies from one panel to the next, and from one country to the next due to differences in the cultures discussed above. Thus, we will carry out random-effects analysis. Under the random-effect model, the weight assigned to each study is calculated as presented in Equation 3.3 (Borenstein et al. 2011):

$$W_i^* = \frac{1}{V_{Y_i}^* + T^2} \qquad (3.3)$$

where $V_{Y_i}^*$ is the within-study variance for study $i$ and $T^2$ is the between-studies variance. However, due to fairly large samples of potential survey respondents (up to more than 1 million sampled numbers in Probit panel recruitment) resulting in very small within-study variances ($V_{Y_i}^*$), as well very large between-study variance ($T^2$, see the 'Overall recruitment rates' subsection in Results), all ORR would end up having almost the same weight and the final results would match those from unweighted samples. Moreover, there were notable differences in the number of recruitment events (waves/rounds for ad-hoc or years for continuous recruitment) between panels, also due to failed attempts to obtain longer ORR time-series from some organizations managing online panels. Thus, we will adjust weighting to ensure that each of the 23 panels will have an equal overall contribution in our models. Assuming that the true effect size for all recruitment events within the same panel is the same, we will calculate within-panel weights as the inverse of their variance, similar to the calculation made under the fixed-effects model presented in Equation 3.4 (Borenstein et al. 2011):

$$W_i^* = \frac{1}{V_{Y_i}^*} \qquad (3.4)$$

The total sum of all within-panel weights will be the same for all 23 panels with available ORR, no matter the length of their ORR time-series or the total sample sizes. We will calculate the weight for a recruitment wave/year $i$ of panel $j$ with $k$ recruitment events as presented in Equation 3.5:

$$W_{ij}^* = \frac{1}{V_{Y_{ij}}^* \sum_{i=1}^{k} \frac{1}{V_{Y_{ij}}^*}} \qquad (3.5)$$

We consider this an adequate adjustment based on the characteristics of our meta-analytic sample.[29]

[29] We will also present unweighted results in the Appendix 3. Comparison of unweighted and weighted results will showcase any effect of weighting on our findings.

### 3.3.5 Publication bias and sensitivity analysis

The publication bias which is estimated and reported in most of more traditional meta-analysis studies, cannot be observed in this type of research and cannot be studied with traditional publication bias approaches, such as funnel plot analysis. Instead, the bias might have originated in a lack of transparency or willingness to report recruitment rates (AAPOR RR1) on the websites of organizations managing probability-based online panels, in their reports, conference presentations or scientific papers. Further, panel organization not being willing to share information with the authors of this study, even after several contacts via various channels, might be a source of bias.

Out of 28 probability-based online panels, we could not locate or were not provided any ORR for five online panels (18%). Out of the remaining 23 online panels, we could find (or were provided) very limited numbers of ORR samples for five additional panels (18%). The average ORR for panels (publicly) sharing very limited information was about 3% points lower than the average ORR for panels (publicly) sharing more or complete information. On the other hand, AmeriSpeak for which we could not obtain AAPOR RR1 ORR, reported 36.9% AAPOR RR3 ORR for 2014-2015 (Montgomery et al. 2016) and 33.7% AAPOR RR3 ORR for 2014-2017 (weighted for selection probabilities) (Bilgen et al. 2018). This seems to compare favorably to the average RR1 ORR for all panels in our study (also after adjusting for unknown eligibility).

Even without in-depth analysis, we can report that online panels from the US and Sweden, the countries with a combined 50% of all probability online panels in the world, were less likely to release or share their recruitment rates. In more competitive markets with a higher portion of commercially oriented and less academically oriented panels, organizations seem to be more hesitant to publicly release methodological information indicating data quality not to affect their competitive advantages.

## 3.4 Results

This section presents expected recruitment rates in probability-based online panels, factors affecting them, and outline the most effective recruitment strategies. We will present changes of effect sizes over time, descriptive statistics for panel characteristics and moderators, and meta-regression analysis with ORR as outcome variables and moderators as predictors. We will start off by presenting basic characteristics of panels, relevant to the further analysis of ORR.

### 3.4.1 Methodological characteristics of panels

The results from Table 3.1 reveal many methodological differences between probability-based panels. For some panels, i.e., those which publicly share little methodological information and did not respond to our emails, we do not have complete information on their characteristics. Out of all 28 online panels, 23 or 82% were still active, and the vast majority of them (92%) conducted probability-based recruitment only. In terms of their recruitment mode, and based on the information we collected, many panels have used different recruitment modes, either in the same recruitment round, or switching between modes over time. The most commonly used mode was telephone (58%), followed by mail (50%) and F2F (42%). IVR still remains less popular for recruitment to a panel.

Table 3.1: Basic methodological characteristics of online panels related to recruitment

| | | n | % |
|---|---|---|---|
| Panel activity status | Active | 23 | 82% |
| | No longer active | 5 | 18% |
| Sampling approach(es) to recruitment | Only probability-based | 24 | 92% |
| | Mixed-sampling* | 2 | 8% |
| Ever used these recruitment modes | IVR | 1 | 4% |
| | Telephone | 15 | 58% |
| | Face-to-face | 11 | 42% |
| | Mail/postal | 13 | 50% |
| Ever used recruitment incentives | Yes | 12 | 52% |
| | No | 11 | 48% |
| Ever practiced piggybacking recruitment | Yes | 6 | 24% |
| | No | 19 | 76% |
| Normally practicing continuous recruitment | Yes | 6 | 24% |
| | No | 19 | 76% |
| Survey mode to collect data from the offline population | No offline mode | 5 | 19% |
| | Telephone | 7 | 27% |
| | Face-to-face | 1 | 4% |
| | Mail/postal | 2 | 8% |
| | CASI – Tablet/computer** | 11 | 42% |

*we only analyzed their probability-based recruitment, **provided by the panel organization at recruitment

Furthermore, about one-half of all probability-based online panels offered incentives in one or more recruitment events, about one-quarter of them carried out piggybacking recruitment and about one-quarter conducted continuous recruitment at any point in time instead of more common recruitment in annual or less frequent recruitment rounds. While about 1 in 5 panels collected data online only and about 2 in 5 online panels provided technology to their panellists to respond online, about 27% of all panels used telephone to collect data from so-called 'offliners'. Face-to-face and mail/postal as the offline modes were quite uncommon.

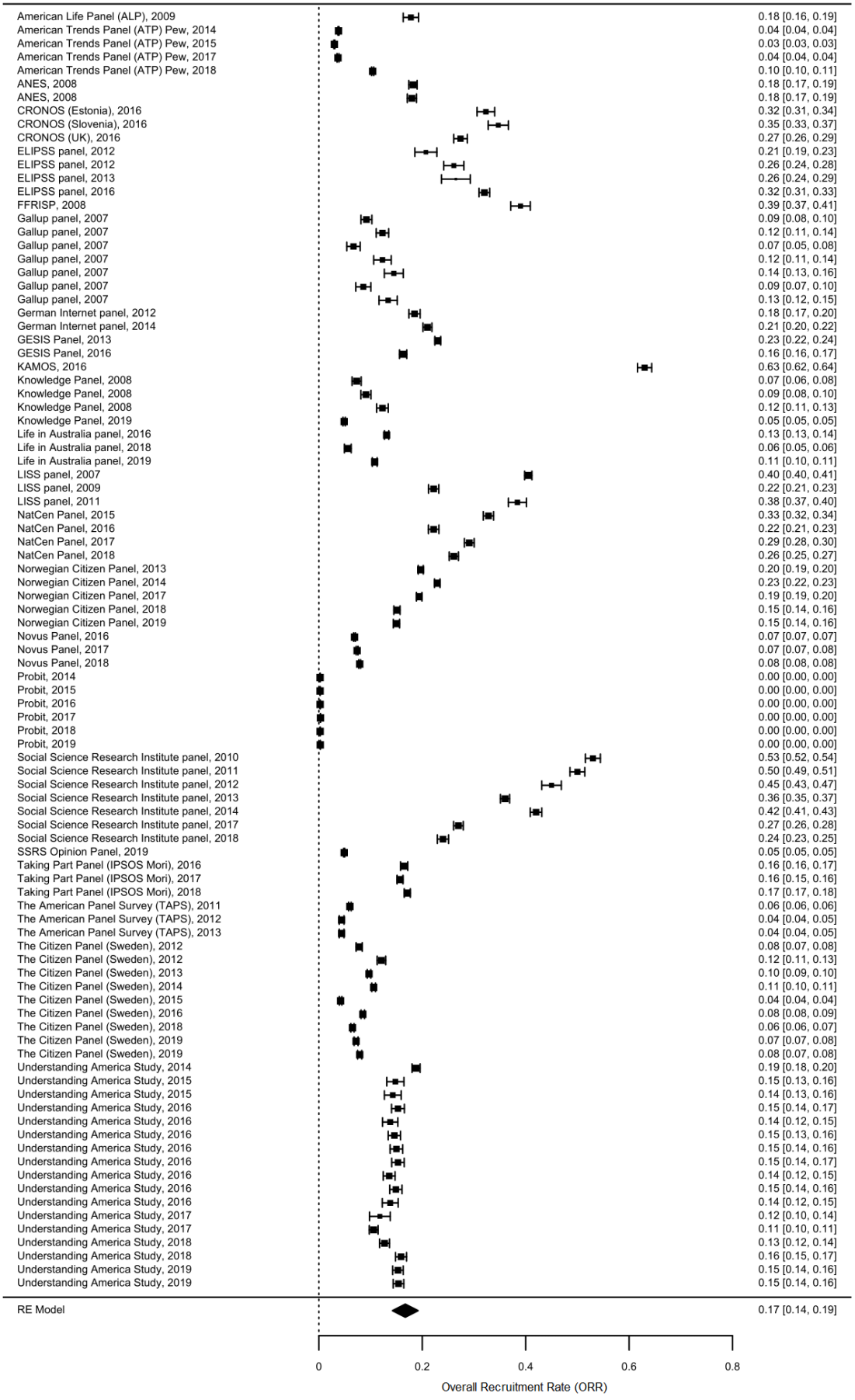
### 3.4.2 Overall recruitment rates

Overall recruitment rates for all recruitment events are presented in Figure 3.1; a forest plot shows the expected recruitment rates with unweighted pooled effect size. We can observe some implications for outlier detection. First, the results show a very high between-study heterogeneity with some very low (<0.5%, Probit) and some very high (>60%, KAMOS) ORR, which is consistent with the ratio of true heterogeneity to total observed variation ($I^2$=100%). Further, $T^2$ as the between-study variance of the model without moderators equalled to 0.0153, which is considerably more than the variance from any of the studies. In practice, this would make weighting under a random-effect model very inefficient and the selection of a fixed-effect weighting model seemed to be the correct one.

The very high between-study heterogeneity in practice means that the average unweighted pooled effect size estimate (16.7% [14.2%, 19.2%]) is a less important statistic in comparison to variability measures. On the other hand, we can observe much less within-panel heterogeneity over time (as explained in the 'Data analysis and weighting' subsection), with Social Science Research Institute Panel (Iceland) being the most notable exception – from 53% ORR in 2010 to 24% ORR in 2018 with no major methodological differences in recruitment strategy. For some other online panels, such as American Trends Panel (ATP), the differences in ORRs could clearly be attributed to notable changes in recruitment – between 2014 and 2017, ATP were recruiting via telephone with ORR between 3 and 4%, and in 2018, they switched to mail/postal recruitment with ORR increasing to more than 10%. For some panels like Probit (Canada), Novus Panel (Sweden) or Taking Part (UK), we can report very stable ORR over time. For some other panels, like Social Science Research Institute Panel and Norwegian Citizen Panel, we can identify a fairly consistent decrease of ORR over time, while ELIPSS (France) seems to be the only probability-based online panel in the world with a notable increase of ORR in more recent recruitment waves.

The results show very high levels of between-study heterogeneity, and with observed changes of both recruitment strategies and ORR over time (overall and individual-panel-level), further investigation should be done using a meta-regression. Different recruitment strategies could potentially explain a large portion of variability of the outcome variable ORR. However, due to some extremely low and extremely high ORR as shown in Figure 3.1, it is crucial to carry out outlier detection and influence diagnostics before constructing any meta-regression models.

Figure 3.1: Forest plot for overall recruitment rates (ORR), unweighted, ordered alphabetically



### 3.4.3 Cumulative overall recruitment rate

To extend the analysis of average ORR and its variability, we will present a cumulative forest plot capturing recruitment rate changes over time from all recruitment events combined. We are particularly interested in an overall trend; i.e., if any consistency with declining survey response rates in surveys in general can be observed.

Figure 3.2: Cumulative forest plot for ORR, weighted, ordered chronologically

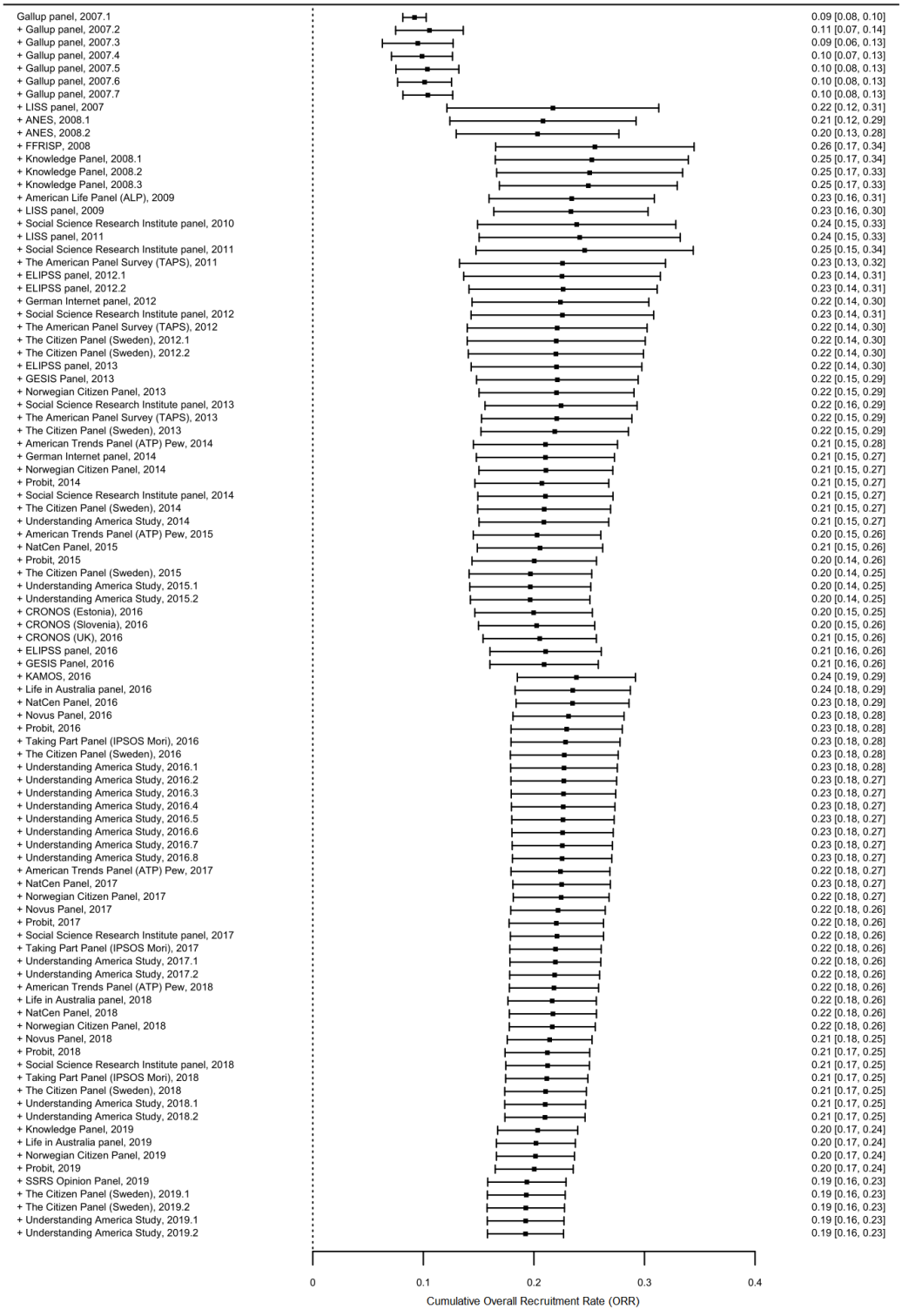


Figure 3.2 is also showing how changes in average ORRs would affect the results on pooled effect sizes over time, e.g., if this meta-analysis had been carried out a few years back. The cumulative forest plot presents weighted results (see Subsection 3.3.4 for more information). The average ORR in our study after weighting (see the forest plot in Figure 3.2) did not differ much from the unweighted results in the forest plot in Figure 3.1 (+2.5%), and most of that increase could be attributed to a decrease of contribution of Probit (5 recruitment years) and an increased contribution of KAMOS (single recruitment) to the pooled effect size.

The cumulative forest plot in Figure 3.2 chronologically presents accumulation of evidence on recruitment rates in probability-based online panel research. It primarily shows one notable trend: the pooled effect size becomes more precise over time with decreasing confidence intervals as a result of increasing sample size of ORR. We can also observe a fairly slow trend of decreasing pooled effect size between 2008 and 2015 (after FFRISP), and after 2016 (KAMOS). This trend would be more consistent if we removed both FFRISP and KAMOS data. FFRISP recruitment rate (39%) tends to be fairly high for North America, which can be explained with F2F recruitment and extremely high conditional incentives – $500 internet access equipment or $200 monetary incentives at recruitment. KAMOS recruited panellists in South Korea, which is the only (non-Middle-East) Asian country with a probability-based online panel, and the cultural differences might have been a source of such a high ORR (63%). However, completion rate (COMR) in the first post-recruitment KAMOS panel survey was quite low (50%) in comparison to most other panels (see Blom et al. 2016; Kaczmirek et al. 2019), which indicates that there might have been notable differences in how 'panel registration' was defined. We also observed that without the panel by far the lowest ORR, i.e., Probit with IVR recruitment mode, cumulative effect size would decrease in a more consistent fashion. These three online panels seem to be the best candidates to be identified as outliers with influence diagnostics.

### 3.4.4 Meta-regression analysis with moderators

Due to a very high between-study heterogeneity, as well as substantial differences in methodological approaches to recruitment to a probability-based online panel, meta-regression analysis is required to explain the variability of the outcome variable overall recruitment rate. We started off by carrying out outlier detection and influence diagnostics, which identified KAMOS and FFRISP recruitment as outliers. We removed them from the final data for meta-regression.

In Table 3.2, two meta-regression models are presented[30]:

[30] To check the robustness of the models we also ran models Model 3 with unweighted results and no interactions (every recruitment event with the same contribution) and Model 4 with unweighted results and interactions. These are presented in Table 3.4 in the Appendix 3. The results and conclusions do not substantially change in case of creating and using weights as explained in the 'Data analysis and weighting' subsection.

- Model 1 with weighted results and no interactions (every panel with the same contribution, recruitment event with within-panel contribution proportional to the inverse of their variance)
- Model 2 with weighted results and interactions.

Reviewing the moderators, we noticed that a combination of conditional and unconditional monetary incentives was used in 63% of all recruitment waves in which money was offered in exchange for joining the panel. Moreover, we observed that a telephone and mail mixed-mode recruitment was used in almost one-third of recruitment waves/years. Having evidence on statistically significant differences in the effect of incentives and modes on overall recruitment rates (see Models 1), we included conditional*unconditional incentives and telephone*mail mode interactions to Model 2 to explore the impacts of incentives and recruitment modes on ORR in more detail.

The results in Table 3.2 show that there are little differences between the models in portion of explained variability of the response variable ($R^2$, between 61.2% and 61.4%), and no differences in between-study variance of the model with moderators ($T^2$=0.005), and in the ratio of true heterogeneity to total observed variation ($I^2$=100%). Also, both tests for residual heterogeneity and for moderators are statistically significant for all models (including Models 3 and 4 in the Appendix 3), which shows that the effect sizes are heterogeneous.

We observe statistically significant effects of a number of moderators on ORR, and there are minor differences between weighted and unweighted models. Firstly, while the cumulative forest plot in Figure 3.2 indicated a trend of decreasing ORR over time, meta-regression modeling without interactions (Model 1) did not support our assumption that time is a statistically significant predictor of ORR (coefficients=-0.003, p=0.370 (Model 1) and p=0.482 (Model 2)). The reasons for decreasing recruitment rates might be in found in changes of recruitment strategies over time.

Table 3.2: Meta-regression models, weighted – random effect Models 1 and 2 with ORR as the outcome variable and moderators as predictors (n=93 recruitment waves/years)

| Moderators | Model 1 (Weighted, NO interactions) | | | Model 2 (Weighted, WITH interactions) | | |
|---|---|---|---|---|---|---|
| | Coef. | Standard error | p value | Coef. | Standard error | p value |
| Year of recruitment | -0.003 | 0.004 | 0.370 | -0.003 | 0.004 | 0.482 |
| Type of incentives (multiple answer variable) | | | | | | |
| No incentives | 0.170 | 0.056 | 0.003** | 0 | | |
| Lottery/prize draw | 0.219 | 0.069 | 0.001** | 0.050 | 0.040 | 0.207 |
| Conditional monetary incentives | 0.000 | 0.081 | 0.996 | | | |
| Unconditional monetary incentives | 0.100 | 0.062 | 0.105 | | | |
| *Both unconditional and conditional incentives* | | | | *0.170* | *0.056* | *0.003*** |
| *Conditional monetary incentives only* | | | | *-0.168* | *0.062* | *0.007*** |
| *Unconditional monetary incentives only* | | | | *-0.067* | *0.052* | *0.195* |
| Guaranteed recruitment incentives in USD | 0.006 | 0.005 | 0.240 | 0.006 | 0.005 | 0.294 |
| Recruitment mode (multiple answer variable) | | | | | | |
| Interactive voice response (IVR) | -0.134 | 0.054 | 0.014* | -0.128 | 0.068 | 0.059† |
| Face-to-face | 0.043 | 0.048 | 0.374 | 0.049 | 0.056 | 0.379 |
| Telephone | 0.032 | 0.035 | 0.359 | | | |
| Mail/postal | -0.091 | 0.033 | 0.006** | | | |
| *Both telephone and mail/postal* | | | | *-0.012* | *0.069* | *0.866* |
| *Telephone only* | | | | *0.040* | *0.059* | *0.502* |
| *Mail/postal only* | | | | *-0.083* | *0.062* | *0.179* |
| Multiple-mode contact/recruitment | -0.026 | 0.040 | 0.514 | -0.025 | 0.041 | 0.547 |
| Piggybacking recruitment | 0.001 | 0.044 | 0.977 | 0.002 | 0.045 | 0.961 |
| North American panel | -0.087 | 0.033 | 0.009** | -0.085 | 0.033 | 0.010* |
| Intercept | 0.085 | 0.076 | 0.260 | 0.244 | 0.080 | 0.002** |
| $T^2$ | 0.005 | | | 0.005 | | |
| $I^2$ | 1.000 | | | 1.000 | | |
| $R^2$ | 0.614 | | | 0.612 | | |
| Test for Residual Heterogeneity | QE(df = 79) = 19607.9, p < .0001 | | | QE(df = 78) = 19590.3, p < .0001 | | |
| Test of Moderators | QM(df = 13) = 114.2, p< .0001 | | | QM(df = 14) = 116.4, p< .0001 | | |

***p<0.001, **p<0.01, *p<0.05, †p<0.10; Coef. – regression coefficient

Furthermore, we observed some interesting evidence related to incentives use for recruitment purposes. The first results showed that not offering incentives did not have a negative impact on ORR in comparison to offering respondents to enter a prize draw or unconditional incentives. Not offering incentives (coefficient=0.170, p=0.003**) had a more positive effect on ORR (i.e., +17.0% in ORR) than offering conditional monetary incentives (coefficient=0.000, p=0.996). However, we have to take into account the effect of the incentives amount, knowing that *No incentives* or *Lottery/prize draw* always have a value of zero USD for *Guaranteed recruitment incentives in USD*, unlike the two types of monetary incentives. Moreover, adding interactions (see Model 2) explained the effective use of

incentives in more detail: unconditional incentives offered the best ORR maximization in combination with conditional monetary incentives (coefficient=0.170, p=0.003**), while offering only unconditional incentives (coefficient=-0.067, p=0.195) performed about as well as prize draws (coefficient=0.050, p=0.207) or even no incentives (the reference group, coefficient=0). Offering only conditional incentives proved to be the worst of all incentives use approaches (coefficient=-0.168, p=0.007**). The incentives amount, in our study conceptualized as the maximum guaranteed amount in USD (sum of conditional and unconditional incentives), did not seem to have a statistically significant effect on ORR (coefficients=0.006, p=0.240 (Model 1) and p=0.294 (Model 2)). Splitting the amount into conditional monetary incentives amount and prepaid monetary incentives amount, and possibly adjusting the amounts to the average wages in participating countries, could explain more variability. However, we would have to collect a larger sample of ORR to include new moderators.

In terms of the effect of recruitment modes on ORR, F2F (coefficient=0.043, p=0.374) and telephone (coefficient=0.032, p=0.359) recruitment modes performed similarly well, while mail mode (coefficient=-0.091, p=0.006**) and especially IVR (coefficient=-0.134, p=0.014*) recruitment mode performed significantly worse than the interviewer-administered recruitment modes (Model 1). In practice this means that, in comparison to F2F and telephone modes, between about 13% (mail) and about 17% (IVR) lower ORR could be expected if all other recruitment methodologies are being equal. Interestingly, multiple-mode contact/recruitment or piggybacking recruitment did not have a statistically significant effect on ORR. Hence, our results do not advise against using potentially more time- and cost-efficient piggybacking approaches to recruitment, from a response perspective.

On the other hand, including *North American panel* as a binary predictor of recruitment rates helped explain more variability. Initially we believed that lower ORR in North American panel recruitment could be attributed to unpopularity/impracticability of F2F recruitment mode, especially in comparison to most European Union (EU) based panels. However, the results showed that North American probability-based online panels, with panels from the US representing the vast majority of the group, generally have lower ORR across all modes and recruitment strategies compared to similar panels from other continents. The gap is about 9% on average, if all other methodologies are being equal (Model 1: coefficient -0.087, p=0.009**, Model 2: coefficient -0.085, p=0.010*).

Lastly, we would like to compare the performance of different meta-regression models to increase robustness of our findings and to think about how future meta-analyses having access to longer time-series could improve estimation. Initially, we considered using three-level meta-analytic models due to the panel nature of data – 23 panels with 95 recruitment events. However, we observed minor methodological changes in how most panels with 2+ recruitment waves/years in our study carried out recruitment, which leads to the problem of very low variability of moderators and an inability to fully

exploit the advantages of multi-level meta-regression. Moreover, while weighting had a small and often negligible effect on estimates (Models 1 and 2 compared to Models 3 and 4 in the Appendix 3), increasing standard errors contributed to the loss of statistical power and significance for a limited number of covariates, e.g., for unconditional monetary incentives in Model 1. This indicates that in meta-regressions with relatively small samples of studies but a number of relevant moderators (e.g., studies/moderator ratio being just below 10), constructing, presenting and comparing results of both weighted and unweighted models should be considered. On the other hand, including interactions to meta-regression models helped explain the effects of combining or not combining modes and incentives approaches on ORR. Model 4 with interactions also indicated a potentially significant effect of year of recruitment (negative effect) and monetary incentives amount (positive effect) on ORR, but at $p < 0.1^{+}$ level (see Table 3.4 in the Appendix 3). To minimize model selection effect on final findings, we would suggest researchers presenting different models and discussing any differences (as we did).

## 3.5 Discussion

This meta-analysis on recruitment rates in probability-based online panel research closes an important research gap and presents a number of practical solutions for online panel research practice. To the best of our knowledge it is the first meta-analysis on nonresponse error in online panels, as survey methodology is a discipline with fewer meta-analyses – in accordance with the findings of Čehovin et al. (2018). Besides offering methodological solutions to existing panel providers, findings of this study could inform a new online panel in the pre-recruitment phase regarding different recruitment strategies to be tested in their experimental recruitment study, similar to the pilot outlined by Struminskaya et al. (2014). Although the analysis is limited to recruitment events in online panel research, the findings on factors affecting ORRs can be extended to other data collection approaches, including those of cross-sectional nature and excluding the online mode.

In this meta-analysis, we firstly observed reasonably low average panel registration rates with a very high heterogeneity of effect sizes, even higher than heterogeneity reported in similar meta-analyses studying the difference in nonresponse between the online and other survey modes (most recently Daikeler et al. 2020). After removing two outliers, the average effect size was even lower, and by multiplying ORR with COMR and RETR, the average cumulative rate would reach single digit numbers for a notable portion of all online panels. That is consistent with findings on lower response rates in non-panel online surveys (most recently Daikeler et al. 2020). The major differences between ORR in our meta-analysis was the reason for changing the primary focus from estimating the true effect size to explaining the variability of ORR using meta-regression modeling. Assuming that the true effect size varied between probability-based online panels and countries, we were able to explain more than 60% of variability of ORR with the selected moderators. Respecting cultural differences between countries

with probability-based online panels, which was an unobserved heterogeneity we could not fully control for, we consider this a very good result.

Further, we identified a number of statistically significant predictors of ORR as part of recruitment strategies of organizations establishing or replenishing online panels. Some of the findings were in line with findings reported by panels conducting survey recruitment experiments. Our meta-analysis supported the findings of Scherpenzeel and Toepoel (2012) who reported no differences in recruitment rates between telephone and F2F modes. Our findings on the effectiveness of unconditional incentives were consistent with findings by Blom et al. (2015) and DiSogra et al. (2009), whose evidence was synthesized in our study as well, while our results showcased the combination of prepaid and conditional incentives as the best recruitment maximization technique. On the other hand, we did not observe the same positive effect of mail mode on recruitment rates in comparison to the telephone mode or the effect of follow-ups using a different mode on ORR, reported by Rao et al. (2010). As such, our meta-analysis contributes to mixed evidence in survey methodology on the use of incentives and different survey recruitment modes. One of the reasons for mixed evidence in studies/panels included in this meta-analysis could also be bias known as simultaneity bias; simultaneity is where the predictor variables are jointly determined by the outcome variable, and is one cause of endogeneity. In this particular study, while incentives could in theory positively affect ORR consistent with most literature on survey response (see Fan & Yan 2010), panels in countries with generally low survey response rates would have a higher propensity to use them than the panels already achieving target response without offering money in exchange for panel registration.

We also have to acknowledge certain limitations of this study. Based on the information we have collected for those panels, we included ORR for more than 80% of all recruitment events in online panels based in countries outside Sweden or the US. On the other hand, we were not able to include any ORR for 21% of all US/Sweden-based online panels and we had limited time-series for additional 36% of panels based in those two countries with more competitive online panel data collection market. Including ORR or longer ORR time-series for those panels would most probably result in a decreased pooled effect size, as North American panels stood out with lower ORR, *ceteris paribus*. However, with a very high heterogeneity of effect sizes, average ORR should not be the primary focus of a meta-analysis on recruitment rates. Instead, extended time-series of recruitment events from the existing panels would offer new opportunities a few years in the future (e.g., in 2025). While our sample of response rates was comparable to the studies of Manfreda et al. (2008) and Daikeler et al. (2020), our findings could be more robust with additional ORR; also, from the panels we could not find/access ORR for 2007-2019 (see Table 3.3 in the Appendix 3). The main contribution of a larger sample size would be an increased statistical power to help identify more statistically significant methodological

predictors of ORR in meta-regression and explain more variability. This would have to be combined with more detailed methodological reports on recruitment strategies, which would help with identification of additional moderators as predictors, such as sending advanced letters or invitation content. In that case, even the existing coding and selection of moderators could be substantially improved. Certain moderators we used in meta-regressions, like *Multiple-mode contact/recruitment*, could also be split into individual predictors. Lastly, to fully exploit the panel nature of the data, carrying out multi-level meta-regression to control for unobserved heterogeneity should be tested in the future. With the existing data, that could not be done as that type of regression analysis would require more within-panel variability of predictors. Conducting more survey recruitment experiments in this space (and making the results publicly available, similarly to Blom et al. 2015; DiSogra et al. 2009; Rao et al. 2010) would not only improve their panel recruitment strategies but also offer a better-quality meta-analysis data from the modeling perspective. However, the existing evidence already provides very valuable insight into how recruitment to probability-based panels is done in practice, and what methodological solutions work better than some others in a worldwide context.

## 3.6 References


AARP. (n.d.). *AARP Research*. Retrieved December 15, 2020, from https://www.aarp.org/research/

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

Bilgen, I., Dennis, J. M., & Ganesh, N. (2018). *Nonresponse Follow-up Impact on AmeriSpeak Panel Sample Composition and Representativeness.* NORC.

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

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

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

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look at the future. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 1-22). John Wiley & Sons.

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

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and psychological measurement*, *60*(6), 821-836.

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology, 43*, 121-145.

Čehovin, G., Bosnjak, M., & Lozar Manfreda, K. (2018). Meta-Analyses in Survey Methodology: A Systematic Review. *Public Opinion Quarterly, 82*(4), 641-660.

Daikeler, J., Bosnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513-539.

DiSogra, C., Callegaro, M., & Hendarwan, E. (2009). Recruiting probability-based web panel members using an address-based sample frame: Results from a pilot study conducted by knowledge networks. *Joint Statistical Meetings, Survey Research Methods*, 5270-5283.

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in human behavior*, *26*(2), 132-139.

Greaves, L. M. (2017). *An Investigation into the Feasibility of an Online National Probability Panel Study in New Zealand.* University of Auckland.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167-189.

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

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

Lugtig, P., Das, M., & Scherpenzeel, A. (2014). Nonresponse and attrition in a probability-based online panel for the general population. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 135-153). John Wiley & Sons.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79-104.

Mavletova, A., & Couper, M. P. (2015). A Meta-Analysis of Breakoff Rates in Mobile Web Surveys. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies* (pp. 81–98). Ubiquity Press. http://dx.doi.org/10.5334/bar.f

Montgomery, R., Dennis, J. M., & Ganesh, N. (2016). *Response rate calculation methodology for recruitment of a two-phase probability-based panel: the case of AmeriSpeak.* NORC.

Rao, K., Kaminska, O., & McCutcheon, A. L. (2010). Recruiting probability samples for a multi-mode research panel with internet and mail components. *Public Opinion Quarterly*, *74*(1), 68-84.

Scherpenzeel, A., & Toepoel, V. (2012). Recruiting a probability sample for an online panel: Effects of contact mode, incentives, and information. *Public opinion quarterly*, *76*(3), 470-490.

Shih, T. H., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field methods*, *20*(3), 249-271.

Singapore Management University. (n.d.). *About the Singapore Life Panel*. Retrieved December 15, 2020, from https://rosa.smu.edu.sg/singapore-monthly-panel

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The end of the (research) world as we know it? Understanding and coping with declining response rates to mail surveys. *Society & Natural Resources*, *32*(10), 1139-1154.

Struminskaya, B., Kaczmirek, L., Schaurer, I., & Bandilla, W. (2014). Assessing representativeness of a probability-based online panel in Germany. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 61-85). John Wiley & Sons.

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

USC Dornsife Center for Economic and Social Research. (n.d.). *Understanding America Study.* Retreived December 15 2020, from https://uasdata.usc.edu/index.php

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://www.jstatsoft.org/v36/i03/.

WebSM. (n.d.). *Probability based panels – WebSM*. Retrieved December 15, 2020, from http://www.websm.org/db/42/17863/Highlights/Probability_based_panels/

## Appendix 3

Table 3.3: A list of identified probability-based online panels (1/2)

| No. | Name of panel | Organization | Country | Established (if known) | Active | Source of overall recruitment rates | Reasons for missing or incomplete recruitment time series |
|---|---|---|---|---|---|---|---|
| 1 | Probit | EKOS | Canada | | yes | panel organization | |
| 2 | LISS panel | CentERdata | Netherlands | 2007 | yes | publicly available documentation (article) | could not provide the latest overall recruitment rates |
| 3 | GESIS panel | GESIS | Germany | 2013 | yes | publicly available documentation (reports) | |
| 4 | German Internet panel | University of Mannheim | Germany | 2012 | yes | publicly available documentation (website) | couldn't share information on the latest recruitment (2018) until a journal article is published |
| 5 | Knowledge Panel | GfK/Ipsos | USA | 1999 | yes | publicly available documentation (article) | limited information from the documentation; discussion in person, no information sent |
| 6 | Gallup panel | Gallup | USA | 2004 | yes | publicly available documentation (article) | limited information from the documentation, couldn't share more rates in the required form |
| 7 | FFRISP panel | Stanford University and Abt SRB | USA | 2007 | no | publicly available documentation (report) | |
| 8 | AmeriSpeak panel | NORC at the University of Chicago | USA | 2014 | yes | | discussion in person, not enough information provided after |
| 9 | American Trends Panel | Pew Research Center | USA | 2014 | yes | panel organization | |
| 10 | American Life Panel | RAND Corporation | USA | 2003 | yes | publicly available documentation (report) | limited information from the documentation; couldn't share information, will be released in a report |
| 11 | American National Election Study Panel | Stanford University and the University of Michigan | USA | 2008 | no | publicly available documentation (reports) | |
| 12 | ELIPSS panel | CDSP/DIME-SHS | France | 2012 | yes | panel organization, publicly available documentation (reports) | |
| 13 | SSRI panel | Social Science Research Institute | Iceland | 2010 | yes | panel organization | |
| 14 | NatCen Opinion Panel | NatCen Social Research | UK | 2015 | yes | publicly available documentation (reports) | |
| 15 | Australian Health and Social Science Survey | Institute for Health and Social Science Research at CQU | Australia | 2009 | no | | They did not have access to the relevant information after a few years since the panel was deactivated |
| 16 | Life in Australia™ | Social Research Centre | Australia | 2016 | yes | panel organization | |

Table 3.3: A list of identified probability-based online panels (2/2)

| No. | Name of panel | Organization | Country | Established (if known) | Active | Source of overall recruitment rates | Reasons for missing or incomplete recruitment time series |
|---|---|---|---|---|---|---|---|
| 17 | Demoskop Panel | Demoskop | Sweden | | yes | | no response to any of our emails |
| 18 | Novus Panel | Novus | Sweden | 2008 | yes | panel organization | provided rates for the most recent recruitment only |
| 19 | Sifo Panel | Kantar Sifo | Sweden | | yes | | no response to any of our emails |
| 20 | Cross-National Online Survey (CRONOS) | European Social Survey | EST, SVN, UK | 2016 | no | publicly available documentation (reports) | |
| 21 | Norwegian Citizen Panel | University of Bergen, Uni Research Rokkan Center | Norway | 2013 | yes | publicly available documentation (reports) | |
| 22 | Understanding America Study | University of Southern California | USA | 2014 | yes | publicly available documentation (website) | |
| 23 | The American Panel Survey | Washington University in St. Louis | USA | 2011 | no | publicly available documentation (reports) | |
| 24 | KAMOS | The Center for Asian Public Opinion Research & Collaboration Initiative | Korea | 2016 | yes | publicly available documentation (conference presentation) | responded to our first email, never provided more information after several reminders |
| 25 | SSRS Opinion Panel | SSRS | USA | 2015 | yes | panel organization | provided rates for the most recent recruitment period only |
| 26 | IranPoll Online panel | People Analytics | Iran | | yes | | no response to any of our emails |
| 27 | The Citizen Panel | Laboratory of Opinion Research | Sweden | 2010 | yes | panel organization | |
| 28 | Taking Part | Ipsos Mori | UK | 2016 | yes | publicly available documentation (reports) | |

Table 3.4: Meta-regression models, unweighted – random effect models 3 and 4 with ORR as the outcome variable and moderators as predictors (n=93 recruitment waves/years)

| **Moderators** | **Model 3** (Unweighted, NO interactions) | | | **Model 4** (Unweighted, WITH interactions) | | |
|---|---|---|---|---|---|---|
| | Coef. | Standard error | p value | Coef. | Standard error | p value |
| Year of recruitment | -0.004 | 0.003 | 0.130 | -0.005 | 0.003 | 0.093† |
| Type of incentives (multiple answer variable) | | | | | | |
| No incentives | 0.176 | 0.045 | <.0001*** | | | |
| Lottery/prize draw | 0.201 | 0.057 | <.0001*** | 0.022 | 0.035 | 0.523 |
| Conditional monetary incentives | -0.010 | 0.062 | 0.867 | | | |
| Unconditional monetary incentives | 0.143 | 0.046 | 0.002** | | | |
| *Both unconditional and conditional incentives* | | | | *0.175* | *0.045* | *<.0001**** |
| *Conditional monetary incentives only* | | | | *-0.194* | *0.050* | *<.0001**** |
| *Unconditional monetary incentives only* | | | | *-0.041* | *0.040* | *0.307* |
| Guaranteed recruitment incentives in USD | 0.006 | 0.004 | 0.121 | 0.007 | 0.004 | 0.088† |
| Recruitment mode (multiple answer variable) | | | | | | |
| Interactive voice response (IVR) | -0.131 | 0.047 | 0.005** | -0.159 | 0.060 | 0.008** |
| Face-to-face | 0.018 | 0.038 | 0.630 | 0.001 | 0.044 | 0.981 |
| Telephone | 0.054 | 0.025 | 0.029* | | | |
| Mail/postal | -0.090 | 0.026 | <.0001*** | | | |
| *Both telephone and mail/postal* | | | | *0.041* | *0.054* | *0.446* |
| *Telephone only* | | | | *0.022* | *0.049* | *0.649* |
| *Mail/postal only* | | | | *-0.124* | *0.052* | *0.016** |
| Multiple-mode contact/recruitment | -0.039 | 0.028 | 0.167 | -0.043 | 0.029 | 0.138 |
| Piggybacking recruitment | 0.026 | 0.037 | 0.478 | 0.015 | 0.040 | 0.719 |
| North American panel | -0.092 | 0.027 | <.0001*** | -0.095 | 0.027 | <.0001*** |
| Intercept | 0.091 | 0.056 | 0.103 | 0.307 | 0.062 | <.0001*** |
| $T^2$ | 0.005 | | | 0.005 | | |
| $I^2$ | 1.000 | | | 1.000 | | |
| $R^2$ | 0.614 | | | 0.612 | | |
| Test for Residual Heterogeneity | QE(df = 79) = 19607.9, p < .0001 | | | QE(df = 78) = 19590.3, p < .0001 | | |
| Test of Moderators | QM(df = 13) = 158.4, p< .0001 | | | QM(df = 14) = 158.2, p< .0001 | | |

***p<0.001, **p<0.01, *p<0.05, †p<0.10; Coef. – regression coefficient

# Chapter 4 Social-psychological aspects of probability-based online panel participation

## 4.1 Introduction

With survey response rates declining to single-digit numbers in certain survey modes (Keeter et al. 2017), survey researchers and their organizations, including those that establish and maintain online panels, need to have a better understanding of why certain people participate in particular types of surveys and what are the psychological differences between respondents and nonrespondents. Also, survey methodology as a research discipline is predominantly based on quantitative evidence about particular methods, and much less on qualitative evidence – since it is fundamentally a quantitative research approach. For example, an experiment of the leverage-saliency theory (LST) of survey participation (Groves et al. 2000) and an investigation into social-psychological theories explaining response in web surveys (Keusch 2015), were both studies based on quantitative evidence. Lastly, there is limited evidence on the effect of personality on survey nonresponse, especially in panel survey research (e.g., Hansson et al. 2018).

This chapter is based on the assumption that, in order to understand survey respondents' behavior and their motivation for, or reasons for not participating in (panel) survey research, more qualitative evidence on the social-psychological aspects of survey participation is needed. The existing psychological research in the field of survey methodology seems to have been more focused on the topics like the psychology of item response and associated measurement error (e.g., Krosnick 1991; Tourangeau et al. 2000; Ward & Meade 2018) than on the psychology of survey participation and associated nonresponse error (e.g., Groves et al. 1992; Holmberg et al. 2008; Keusch 2015), especially in the case of panel studies (e.g., Haunberger 2011). Dillman (2020) argues that theories of response behavior tend to be dated, ignore certain techniques for improving response, are limited to single-mode applications, and ignore how the design of each survey contact, as well as any associated invitation materials, should be guided by the theory.

General population probability-based online panels are, from a social-psychological perspective, a special case for a couple of reasons. First, they often use mixed-mode protocols, with offline recruitment and reminders (Kaczmirek et al. 2019). Second, the panel nature of survey participation exhibited in regular, often monthly or even more frequent, requests for completion of questionnaires (see Chapter 3). These differences have important implications for the research into social-psychological aspects of survey participation.

Although most studies in this space have been based on quantitative evidence on (non)response in cross-sectional surveys, but not on survey response in longitudinal or panel settings, we will present mostly qualitative evidence to answer the following research questions:

1. *What are the main motivational factors and barriers which influence panellists' panel participation behavior (recruitment outcomes, survey completion, voluntary attrition)?*
2. *Do the main motivational factors and barries differ between panellists with different participation patterns (frequent-respondents, infrequent respondents, and nonrespondents), and is their panel behavior associated with their personality traits?*
3. *How can their panel participation behavior, i.e., survey completion over time, be explained with social-psychological theories on survey participation?*

## 4.2 Social-psychological theories of survey participation

There are various contemporary social psychology theories that may be relevant to explain survey participation, i.e., why people participate in surveys. The most common theoretical frameworks discussed in the literature are (a) social exchange theory, (b) cognitive dissonance, (c) self-perception, (d) commitment and/or involvement, (e) theory on planned behavior, (f) compliance heuristics, and (g) leverage-salience theory (cf. Albaum & Smith 2012; Keusch 2015), while economic exchange theory can explain the role of incentives (Lavrakas 2008) and other behavioral change theories could explain the process of voluntary attrition (such as the reasoned action approach of Fishbein & Ajzen 2011). We will review these frameworks in more detail, posit the link between the frameworks and motivational factors and barries for survey participation, and review the literature on the association between personality and survey response.

1) Social exchange theory

Social exchange theory (SET) is based on the premise that people's feelings about interactions and relationships depend on the associated outcomes. Previous rewards for participation in activity, how often activity is rewarded, and the value of the activity/rewards, can all have an effect on participation in the activity. The perception of costs and benefits accompanying activity determine the evaluation of the activity, and the perceived rewards should be at least equal to perceived costs for a person to continue with the activity (Homans 1961). Social exchange is based on *unspecified* obligations with no contract and no exact price/costs (Blau 1964).

In survey methodology, SET is most frequently used to explain respondents' decisions whether to participate or not participate in surveys (Dillman 1978; Goyder & Boyer 2008). While rewards in the

cost-benefit model are associated with personal interests of sampled members (e.g., satisfaction with participation), examples of costs are shortage of time, uncertainty about unfamiliar situations, or privacy issues concerns (Keusch 2015).

2) Economic exchange theory

The main difference between social exchange and economic exchange is that the obligations incurred are specified in economic exchange. Also, the types of exchange differ in the trust required for the exchange and promoted by the exchange. In economic transactions, a formal contract exists, which specifies exact quantities to be exchanged (Blau 1964).

Economic exchange theory explains, similarly to SET, that perceived costs of participation should not exceed perceived benefits. It also provides an explanation for why incentives that are perceived as payment or compensation for time/effort can in practice increase response, in contrast to perceptions of incentives as a token of appreciation, as in case of social exchange (Biner & Kidd 1994; Lavrakas 2008).

3) Self-perception theory

Self-perception theory explains the relationship between behavior, attitudes and beliefs. Individuals own attitudes can be based on their perception of their own overt behavior. Not only that, their attitudes and beliefs can be influenced by their participation in role-playing and through self-observation of that behavior – a phenomenon known as self-persuasion (Bem 1967).

In surveys, self-perception theory explains why previous survey response (or a related activity) can affect respondents' propensity to respond to a subsequent survey invitation. For example, Trussell and Lavrakas (2004) showed how those who agreed to participate in the first stage of a mixed-mode survey were more likely to participate in the second stage (with no rewards) than those who were never contacted or refused to participate in the first stage, but had been offered $10 in the second stage. Furthermore, a shorter questionnaire followed by a request to a longer questionnaire (so-called a foot-in-the-door technique) is considered an effective response maximization approach; other approaches based on this theory are labelling potential respondents as regular survey respondents, or communicating their previous survey behavior to them (Keusch 2015).

4) Cognitive dissonance theory

This theory addresses a motivational state known as cognitive dissonance, which results from a person holding two (or more) inconsistent cognitions. This creates dissonance for the person, which is motivational because it is psychologically uncomfortable. This causes the person to seek to remove or

alter one of those two cognitions. Cognitive dissonance is based on consistency among our cognitions, which need to be restored (Bem 1967; Festinger 1957).

As it relates to survey research, the literature explains that there are different conditions that create dissonance among survey participants, thereby leading them to be more willing to respond. The most common is the use of unconditional/noncontingent incentives that are given prior to survey completion. Such incentives are consistently reported to be effective in increasing response from a respondent because accepting the incentive without earning it, is posited to create dissonance (Boulianne 2008). Thus, the respondent is motivated to eliminate the dissonance by competing the survey task. Keusch (2015) also lists reminder messages as a means to motivate respondents because such messages create dissonance for the respondent.

5) Reasoned action approach

The reasoned action approach is a theory of behavioral prediction, and is the most current form of theories of reasoned action and of planned behavior. The approach explains how beliefs associated with a given behavior guide the decision to perform or not perform that behavior. There are three types of beliefs: (a) behavioral, (b) injunctive and descriptive normative beliefs, and (c) control beliefs. These determine an individual's attitudes, perceived norms, and perceived behavioral control, which if combined lead to behavioral intention. The stronger the intention and the more actual control a person has over the performance of the behavior, the more likely the behavior is performed (Fishbein & Ajzen 2011).

Bosnjak et al. (2005) applied an extended version of theory of planned behavior to web-based panel surveys. They found that intention to participate in the survey is predicted best by one's own perception of control over her/his decision to participate in the survey, the attitude towards survey participation, and subjective norms. Moreover, response behavior can be predicted by the intention to participate in the survey, but not directly by perceived control over survey participation.

6) Leverage-salience theory

The leverage-salience theory of survey participation helps explain survey nonresponse through various survey attributes affecting the response/nonresponse outcome. According to the theory, each survey attribute (e.g., mode of contact, number of contacts, survey topic, survey purpose, identity of the survey sponsor, use of incentives, questionnaire length/burden, etc. (Groves et al. 2000; Groves et al. 2009; Seifert 2008)) will have different leverage for different persons on their decision to participate in a given survey, while the activation of the leverage is dependent upon whether that attribute is made salient to the potential respondent (Groves et al. 2000). Survey attributes can be

envisioned as weights on a scale, which tip the balance of the scale in either a response or nonresponse direction. Thus, the same survey request will not have the same outcome with different persons or even the same person at different times, as few people will always react the same way (agree to participate or refuse to) to survey requests with different survey attributes.

7) Compliance heuristics

While people frequently decide about performing an activity based on the attractiveness of its features, other social-psychological factors may play a role in the decision, and a survey request situation predominantly favors a heuristics approach to decision-making – given the brief amount of cognitive time and energy that goes into the decision (cf. Groves et al. 1992). Cialdini (1987) lists the following compliance heuristics principles: (a) reciprocation (repaying for what someone has provided us), (b) social proof (acting like others are acting, in accord with social evidence), (c) liking (either liking those from whom the request comes, with whom one is familiar with, or are being liked by), (d) authority (a recognized authority or those with superior knowledge and judgement), (e) scarcity (the idea of losing something that is short in supply), and (f) commitment and consistency (being consistent with previous actions/activities and/or something we have committed to). In the case of surveys, Albaum and Smith (2012) introduced commitment and involvement as main compliance heuristic principles positively affecting consistency in responding to survey requests.

This review of socio-psychological theories also serves as a theoretical background for studying nonresponse error/bias (and their sources) in some other chapters of this thesis. For example, social exchange, economic exchange and cognitive dissonance theories can be applied to explain the role of monetary rewards in increasing survey response, either conditional or unconditional incentives – this is studied in Chapter 3 with a meta-analytical approach using recruitment data from 23 worldwide probability-based panels.

## 4.3 Association between personality and survey participation

The literature on the association between personality and survey participation is still limited and findings are not consistent. Porter and Whitcomb (2005), who studied the relationship between nonresponse and personality types in student surveys by using Holland's (1966, 1985) personality measures (investigative, artistic, social, and enterprising), concluded that those with investigative personalities and those with less enterprising personalities were more likely to comply with survey requests.

However, most other studies on personalities and nonresponse have used the Big 5 Personality traits test (Goldberg 1992). Hansson et al. (2018), who studied attrition in a general population longitudinal

study, found that higher neuroticism and higher extraversion, as well as lower agreeableness, were all individually associated with a higher propensity for attrition. Lugtig (2014), who studied attrition in a probability-based online panel sample, reported that early attritors (those who opt-out in the first 12 months) scored higher on extraversion, but those who attrited in the first 6 months also scored higher on agreeableness and lower on conscientiousness. Online panellists responding infrequently scored lower on conscientiousness and higher on extraversion than the most loyal panellists.

Of note, while DiSC is a personality-based assessment (Marston 1928) often applied in practice to measure surface traits and explain how these traits lead to behavior (Jones & Hartley 2013), to the best of our knowledge it has not been used in research on nonresponse or attrition in cross-sectional or panel studies. The DiSC was chosen because a significant correlation exists with more commonly used Big 5 Personality traits test, but it is primarily focused on identification of behavior styles/preferred behavior and not on personality traits. As such it is more commonly used by industry than academia (Jones & Hartley 2013), and it might have a potential to explain other types of non-organizational behavior (such as survey participation).

## 4.4 Methods

### 4.4.1 Data

Three kinds of data were gathered from panellists of Life in Australia™, which was a mixed-mode probability-based online panel with approximately 4,000 active panellists in 2021.[31]

There are three main components of this study. First, all Life in Australia™ panellists responding in the June 2019 survey (Wave 28) were asked an open-ended question about their experience participating in the panel:

> *We would like to understand why you chose to be part of Life in Australia™ and what, if anything, you value about being part of it.*
>
> *What does being part of Life in Australia™ mean to you or what motivates you to participate in the surveys?*

[31] The Life in Australia™ is a probability-based online panel established in 2016 (using dual-frame RDD sampling, telephone recruitment), replenished in 2018 (using RDD mobile sampling and telephone recruitment), and expanded in 2019 (using address-based sampling and postal recruitment). The data from panellists have mostly been collected monthly, all panellists have been invited to participate in the majority of surveys, and the telephone mode has been used to collect data from the offline population. Panellists can either receive incentives for their survey completion or can donate to charities (Kaczmirek et al. 2019). As of May 2021, 50 waves of data collection have been carried out.

The question was asked in a regular Life in Australia™ monthly panel survey. For our study we coded 1,557 verbatim responses in order to generate the quantitative data that we analyzed and which we are reporting here. Second, qualitative in-depth interviews (IDIs) were conducted with 15 panellists who answered the open-ended question. They were questioned about their motivations and barriers related to their panel participation, with a focus on the social-psychological drivers for joining the panel and completing monthly surveys. Third, the same 15 respondents completed the DiSC personality trait assessment (Marston 1928) soon after their qualitative interview. For that reason, an online DiSC test was programmed. A summary of data sources used in this study is available in Table 4.1 below. All 15 panellists completed their IDI and all but one completed the DiSC inventory.

Table 4.1: Data used in the study on psychological aspects of online panel participation

| Data | Type of data | Sample size | Data collection period |
|---|---|---|---|
| Verbatim responses to an open-ended survey question | Qualitative | n=1,557 | June 2019 |
| In-depth interviews | Qualitative | n=15 | March 2020 – February 2021 |
| DiSC test | Quantitative | n=14 | March 2020 – February 2021 |

### 4.4.2 Data collection

The survey data were collected in June 2019. Out of 2,000 panellists completing the questionnaire (74.7% completion rate and 8.4% cumulative response rate in Wave 28), 1,715 (84.4%) provided a valid open-ended answer to the question on why they chose to join the panel (i.e., their motivations to participate in Life in Australia™). Of those 1,715, 1,557 panellists or 77.9% of all those who completed that wave gave consent to the use of their data for research purposes. Hence, verbatims from 1,557 panellists were used in this study.

The 15 in-depth interview participants were recruited and semi-structured interviews conducted between March 2020 and February 2021. Data were collected by two qualitative interviewers; one from the Social Research Centre (SRC) and one from Australian National University (ANU)[32]. Due to the COVID pandemic, interviews were conducted virtually or via the phone. The participants were asked questions about their survey participation in general, recruitment to Life in Australia™,

[32] The SRC interviewer was trained and monitored by the qualitative team leader from the Social Research Centre. The ANU interviewer was trained and monitored by the primary investigator of this study from the Australian National University. As part of monitoring, the primary investigator listened to recordings during the IDI data collection period and later discussed any issues with the interviewers. Both of them used the same list of questions for a semi-structured interview and discussion guide (see Appendix 4). It was concluded that there was no reason to believe that any so-called interviewer effects substantially affected participants' answers to questions.

motivation and barriers to panel participation, and, if applicable, the reasons for not completing wave questionnaires and/or voluntary attrition. The full list of questions for a semi-structured IDIs is in the Appendix 4.

Following the in-depth interview, the 15 participants received an email with an invitation to an online DiSC assessment. It was programmed in Qualtrics and consisted of 28 closed-ended questions, each written as groupings of four statements, from which the respondents had select one of the four statement that described her/him the most and one that described her/him the least (see Appendix 4). Each group of four statements included one statement associated with one of DiSC personality traits, Dominance, Influence, Steadiness and Conscientiousness.[33]

### 4.4.3 Panel participation behavior classification

To study different respondent profiles in the second and third component of the study, the Life in Australia™ sample was divided into three distinctive groups (see Table 4.2 for more information). From each of these groups, between four and six respondents were recruited and qualitative data was gathered from each. While the groups were internally homogeneous, they were externally heterogeneous compared to each other as we aimed to study psychological aspects of online panel participation within particular groups of respondents with very different panel behavior over time.

Table 4.2 presents the target groups of panellists from Life in Australia™ that were chosen, sampling criteria, sample sizes, as well as the main panel-management objective for those types of panellists in practice. Frequent-respondents were those panellists who participated in every survey to which they had been invited. The second group were so-called 'stopped-responding' panellists, and in practice we recruited those who had not responded in the previous four consecutive waves at the time of recruitment. The nonrespondent group included both fast voluntary attritors, i.e., those who opted-out in the first 6 months after recruitment in 2019, as well as those who had been recruited (and profiled), but never participated in a monthly Life in Australia™ survey/wave. For our complete classification of probability-based online panel members in practice, please see Table 4.5 in the Appendix 4.

[33] We used a 28-item version first known as Personal Profile System 2800 Series (PPS 2800) and then as DiSC Classic. Dominance can be used to describe people who are outspoken with their opinions, direct, and forceful; Influence can be used to describe people who are lively, enthusiastic and outgoing, enthusiastic, Steadiness can be used to describe people who are accommodating, gentle, and patient with others' mistakes; Conscientiousness can be used to describe people who are reserved, precise, and analytical (Everything DiSC, n.d.).

Table 4.2: Panellist classification, groups of panellists interviewed in this study (qualitative)

| Group | Participation behavior (trend) in Life in Australia™ surveys | Main panel management objective | Sample size[34] |
|---|---|---|---|
| Frequent-respondents (100% or 'gold star' respondents) | Completion rate 100% prior to recruitment to IDIs | Maintaining high participation | 6 |
| Stopped-responding ('dozers') | 4 consecutive waves of unit nonresponse prior to recruitment to IDIs | Reactivation | 5 |
| Nonrespondents ('backouts', 'fast attritors') | Never responded after recruitment, or opted-out in late-2019 in the first 6 months after their recruitment | Initial activation, delaying or avoiding voluntary attrition | 4 |

### 4.4.4 Coding and data analysis

**Processing the Life in Australia™ Open-ended Survey Data.** As noted, written answers (also known as verbatims) to the open-ended question about motivations to join the panel were provided by 1,557 panellists. Those verbatims were transformed into quantitative data for our analyses through the following processes:

- The two authors independently drew random samples of 100 verbatims and created their own preliminary set of categorical codes to characterize/capture different motivations that the panellists had expressed for joining the panel. The authors did not discuss their own creation of categories with each other while they were doing this.
- The two authors then compared their respective categories and found that they had substantial agreement between them. In the few instances where the categories did not initially corresponded, they discussed the variations and readily reached agreement on how to accommodate the differences.
- A final categorization scheme using 16 substantive motivations, plus a "Misc. Other" category, and a "Non-responsive" category (all displayed in Tables 4.3 and 4.6), was agreed.

[34] To recruit participants, an expression of interest (EOI) request email was sent to Life in Australia™ panellists who were selected from the whole panel based on their participation behavior described in Table 4.2. Recruitment of backouts/voluntary attritors was considerably more challenging than recruitment of dozers and gold-star panellists. To recruit gold star panellists, 133 EOIs were emailed out, 36 responded positively, and 13 had to be called in the end for six to be successfully recruited. To recruit five dozers, 24 panellists were called, including those who did not respond positively to the EOIs. To recruit nonrespondents, all 64 ('fast attritors') and 74 ('backouts') panellists, who were initially sent an EOI email, had to be contacted over the phone for a total of four (2+2) to be recruited and interviewed.

- From extensive past experience with the coding of open-ended verbatims provided by survey respondents, the authors agreed to code as many as three different motivations from each panellist's verbatim, but no fewer than one.
- The set of 1500+ verbatims was split between the two authors, and each coded those independently of the other. For 47% of the 1,557 panellists, one motivation was coded from their verbatim; for another 44%, two motivations were coded; and for 8%, three motivations were coded.
- After the coding was complete, the frequencies for the use of each of the 16 substantive categories by the two authors was compared. A very similar level of frequency was found for 14 of the 16 categories. There was not as much correspondence for the following two motivational categories: "To be informed about topical issues" and "Contributing to the survey/study/research/science."
- The authors discussed their respective use of these two categories and developed an understanding of why there had been a difference in the use of each. Based on this understanding, the assignment of each of these two categories was reviewed and where needed the coding of those categories was revised so as to achieve high consistency between the two coders.

Using the data generated by the coding of the verbatims, as many as three new motivational variables were created for each panellist. Then, using the data in those three variables, 16 new variables were created – one for each motivation. These dichotomous variables indicated whether a panellist was coded to have mentioned a particular motivation in her/his verbatim answer.

**Processing the In-depth Interviews.** Data from qualitative IDIs were transcribed and analyzed in NVivo. The same 16 substantive codes used for coding responses to open-ended questions were used to identify motivational factors for joining the panel and completing monthly questionnaires. Separate coding frames were prepared for (1) barriers, and (2) indicators of social-psychological theories explaining panel participation (based on theoretical review of survey participation theories) (see coding frames in the Appendix 4).

**Processing the Personality Test Data**. DiSC assessment data were processed in SPSS. For each DiSC personality trait (Dominance, Influence, Steadiness and Conscientiousness), indexes were calculated as scores from 28 closed-ended questions. The maximum score for each personality trait was 28, which means that for all 28 groups of statements, the statement associated with a particular personality trait would have to be selected as "describes me most". The minimum score was -28, with statements associated with a particular personality trait selected as "describes me least" in all 28

groups of statements. In the end, we compared individual scores for DiSC personality traits, personality types which stood out each panellist, as well as the average panel behavior group scores (see Table 4.11 in the Appendix 4 for descriptive statistics).

## 4.5 Results

### 4.5.1 Coded motivations reported by panellists in the open-ended data

To answer the first research question on the main motivational factors which influence panellists' panel participation behavior, we analyzed survey data (an open-ended question). Table 4.3 presents the frequencies for the proportion of panellists who mentioned each of 18 categories into which the open-ended data were coded (16 substantive motivational categories, a Misc.-Other category, and a Nonresponsive-Answer category). Of note, the rows in the table are organized into motivational "themes" or "groupings" using qualitative research "sense making" by the authors (cf. Roller & Lavrakas 2015). We generally ordered them by their frequency while listing together motivations combined in components as broader motivations.

A Principal Components Analyses (PCA) with Varimax rotation of the 16 motivational variables was conducted. As shown in Table 4.9 in the Appendix 4, eight components were identified as having an eigenvalue of 1.00 or greater. These accounted for 57% of the total variance among the 16 motivational variables. These eight components were subjected to a Varimax rotation and the rotated Component Matrix is shown in Table 4.10 in the Appendix 4. The loadings in Table 4.10 are for the highest absolute value in each row of the table. Those with an absolute value of 0.400 or greater were used to define each component. Out of eight components with eigenvalue of 1.00 or greater, five components were with two or more motivations with an absolute loading of 0.400 or greater, and three components were with mutually exclusive motivations (see further explanation in the Appendix 4). Thus, we identified the following broader motivations: (a) *being motivated by an intellectual attraction,* (b) *being "gifted/valued" for one's participation*, (c) *being motivated to "give/contribute" something as a panel member,* (d) *doing something interesting with one's time*, and (e) *enjoying being part of this particular panel managed by a respected organization* (see Table 4.3).

As shown in Table 4.3, four-fifths of the panellists (80%) mentioned at least one motivation that referred to them being pleased that they were able to share their views and that they believed that someone was "listening"; motivation codes 1, 2 and 4. One-fifth (21%) of panellists expressed at least one motivation that referred to their being attracted intellectually to participate in the panel; motivation coded 7, 8, and 11. In the next largest group, 18% of the panellists of the panellists mentioned at least one motivation that referred to being gifted/valued for their input; motivation

coded 5 and 9. One in nine (11%) of the panellists mentioned at least one motivation of wanting to give/contribute to something they were part of, and one in eleven (9%) mentioned at least one motivation referring to doing something interesting with their time. Other motivations, such as enjoying being part of this particular panel managed by a respected organization or sharing views/opinions in a non-judgemental platform, were less prevalent.

Table 4.3: Motivational factors (coded open-ended question answers)

| **Broader motivation (components, PCA analysis)[35]** | **Code/motivation** | **%** |
|---|---|---|
| / | (4) Self-actualization, allows my voice to be heard | 35.8 |
| | (1) Sharing views/opinions to make a difference or influence change | 32.2 |
| | (2) Sharing views/opinions to represent others like me/population subgroups/minorities | 12.0 |
| *Being motivated by an intellectual attraction* | (7) Thought-provoking to participate in the panel | 11.6 |
| | (11) Enjoying surveys/participating in research | 5.5 |
| | (8) To be informed about topical issues | 3.9 |
| *Being "gifted/valued" for participation* | (5) My opinions are valued, appreciated, taken into account | 10.3 |
| | (9) Receiving incentives | 7.8 |
| *Being motivated to "give/contribute" something* | (14) Contributing to the survey/study/research/science | 5.8 |
| | (10) Donating to charity | 5.1 |
| *Doing something interesting with one's time* | (12) Interesting topics | 8.2 |
| | (16) Have the time/something to do | 1.0 |
| / | (17) Other motivation | 5.0 |
| | (15) Positive sentiment towards Australia | 2.9 |
| *Enjoying being part of this particular panel managed by a respected organization* | (6) Like being part of something, part of a team | 2.3 |
| | (13) Social Research Centre/Life in Australia™ are reliable enterprises | 1.3 |
| / | (3) Sharing views/opinions in a non-judgemental platform | 1.8 |
| | (18) Non-responsive answer | 1.7 |

### 4.5.2 Comparing recruitment strategies with self-reported motivations to join and participate in a panel

Moreover, we expanded our answers to the first research question by comparing reported motivations to Life in Australia™ recruitment communication. Table 4.4 compares (a) the persuasive statements that the SRC used during their recruitment of the Life in Australia™ panel, with (b) the coded self-reports of why panellist joined and remained active in the panel. (As previously noted, the authors gathered all of the panel recruitment scripts and other materials that were used for panel recruitment/maintenance and did a content analysis of them to identify the motivations that they contained.)

[35] See Tables 4.9 and 4.10 in the Appendix 4 for more information.

In understanding this comparison, it is important to recall that all panellists had been exposed to the persuasive statements prior to being asked to provide their open-ended verbatims of motivational influences.[36] Thus, one possibility is that when asked what motivated them to be part of the panel, the panellists' answers were influenced by what they recalled hearing or reading from the SRC about why they should join and stay active in the panel.

As shown in Table 4.4, all but one of the motivational reasons (left column) that were used by the SRC were mentioned by panellists in their verbatims (right column). The one persuasive statement that the SRC used to recruit panellists that was not explicitly mentioned by the panellists with enough frequency to justify forming a unique code for it was "Your participation is voluntary and you can drop out at any time."

Among the persuasive statements used by the SRC in recruitment that matched up most frequently with the motivations that panellists mentioned in their verbatims (as shown in the right-hand column of Table 4.4) were:

- #1 – Sharing views/opinions to make a difference or influence change
- #5 – My opinions are valued, appreciated, taken into account
- #14 – Contributing to the survey/study/research/science.

However, the majority of panellist (55%) did not mention any of these three reasons for why they joined and stayed active in the panel. Whereas, 42% mentioned one of these three reasons and another 3% mentioned two of the reasons; no panellist mentioned all three reasons.

Our interpretation of the results presented in Table 4.4, and our sense-making of them, is that while some of the panellists may have been influenced by their recall of the SRC's recruitment strategies when they were asked to explain their motivations for being part of the panel, the preponderance of the motivations in the verbatims that the panellist provided to the open-ended question were indicative of the panellists' own independent thinking at the moment they answered the question, and were not affected by the recruitment-related persuasive communications they had previously received from the SRC.

After completing this work, the two authors decided to further investigate why the panellists may have expressed these particular 16 motivations. This led the authors to content-analyze all the written and spoken communications from the SRC and its interviewers to panellists at the time(s) that the

[36] This is not to imply that all panellists were exposed to all of the recruitment strategies that the SRC utilized. Rather, that all panellists were exposed to at least some of these strategies in the contact that they had with SRC interviewers and via other communications that the SRC had with panellists.

panellists were being recruited into Life in Australia™ or being encouraged to keep completing the monthly questionnaires.

Table 4.4: Mapping reasons in recruitment communications and self-reported motivation

| No | Reasons in Recruitment communications to join Life in Australia™ | Self-reported coded motivation to participate in Life in Australia™ from verbatims |
|---|---|---|
| 1 | You can influence Australian researchers, policy-makers, and academics | *1 - Sharing views/opinions to make a difference or influence change* |
| | | *14 - Contributing to the survey/study/research/science* |
| 2 | You will help others better understand Australia and Australians | *1 - Sharing views/opinions to make a difference or influence change* |
| | | *5 - My opinions are valued, appreciated, taken into account* |
| | | *14 - Contributing to the survey/study/research/science* |
| 3 | Your views will be heard | *4 - Self-actualization, allows my voice to be heard* |
| | | *5 - My opinions are valued, appreciated, taken into account* |
| 4 | It's an opportunity to share your views | *4 - Self-actualization, allows my voice to be heard* |
| | | *1 - Sharing views/opinions to make a difference or influence change* |
| | | *2 - Sharing views/opinions to represent others like me/population subgroups/minorities* |
| | | *3 - Sharing views/opinions in a non-judgemental platform* |
| 5 | Your views will be represented | *2 - Sharing views/opinions to represent others like me/population subgroups/minorities* |
| | | *5 - My opinions are valued, appreciated, taken into account* |
| 6 | You will gain incentives/ rewards/make donations | *9 - Receiving incentives* |
| | | *10 - Donating to charity* |
| 7 | Your input will be appreciated | *5 - My opinions are valued, appreciated, taken into account* |
| | | *14 - Contributing to the survey/study/research/science* |
| 8 | Your input is valuable | *5 - My opinions are valued, appreciated, taken into account* |
| | | *14 - Contributing to the survey/study/research/science* |
| | | *1 - Sharing views/opinions to make a difference or influence change* |
| 9 | You will be participating in important national research | *14 - Contributing to the survey/study/research/science* |
| | | *1 - Sharing views/opinions to make a difference or influence change* |
| 10 | You are special - relatively few Australians get this invitation-only chance to join | *5 - My opinions are valued, appreciated, taken into account* |
| | | *2 - Sharing views/opinions to represent others like me/population subgroups/minorities* |
| 11 | This is a chance to participate in something innovative/novel/unique | *14 - Contributing to the survey/study/research/science* |
| 12 | Your data are protected by Aussie privacy laws and kept confidential | *13 - Social Research Centre/Life in Australia™ are reliable enterprises* |
| 13 | Your participation is voluntary and you can drop out at any time | */* |
| 14 | You will be helping ANU | *14 - Contributing to the survey/study/research/science* |
| | | *13 - Social Research Centre/Life in Australia™ are reliable enterprises* |
| 15 | You will be affiliated with something ANU is doing | *13 - Social Research Centre/Life in Australia™ are reliable enterprises* |

Table 4.4 presents a comparison between the motivations that were identified in the 1,557 verbatims by the authors and the themes in the recruitment materials that the SRC used, as identified by the authors. As shown in Table 4.4, there is a relatively close correspondence between (a) what a panellist was exposed to in recruitment (and in subsequent SRC follow-up communications to a panellist that

may have occurred), and (b) the motivations for joining and staying in Life in Australia™ that the panellists expressed in their verbatims. However, the correspondence was not perfect, as the panellists expressed six motivations that were not part of what the SRC had told them during recruitment.[37] All in all, we believe that much of what the SRC used in their recruitment protocols struck chords that resonated with many of the panellists who joined Life in Australia™. And, we believe that that was the primary reason for the similarity between what the SRC communicated to panellists during recruitment to persuade them to join Life in Australia™ and what the panellists later reported in their verbatims.[38]

### 4.5.3 Motivation and barriers to participation

To further answer the first research question on the main motivational factors as well as barries which influence panellists' panel participation behavior, we analyzed IDI data. Further to the evidence from survey data (see Section 4.5.1), we will now present the qualitative evidence on motivation and barriers (i.e., barriers are reasons for nonparticipation). As discussed earlier, LST posits that essentially all persons invited to participate in a research panel will have reasons to join/stay active and reasons not to join/stay active.

Furthermore, while motivation to join already has been addressed with our survey data (i.e., the verbatims to the open-ended question), we now will further discuss motivational factors for staying active in the panel across different phases of panel lifecycle, i.e., the recruitment, survey participation, and survey nonparticipation phases using our qualitative IDI data.

#### 4.5.3.1 Motivation across different stages of the panel lifecycle

The results on the motivation from the IDIs did not differ substantially from the survey results based on verbatim responses to an open-ended question in the survey. The most notable difference was that, throughout the IDIs (most of which lasted between 30 and 45 minutes), participants mentioned several different motivational factors, while in a survey and responding to the open-ended question, they rarely wrote about more than two motivations. This confirms the value of the qualitative data in better understanding panel participation motivation and barriers.

[37] These six motivations were: (1) Like being part of something, part of a team; (2) Thought-provoking to participate in the panel; (3) To be informed about topical issues; (4) Enjoy surveys, participating in research; (5) Interesting topics; and (6) Positive sentiments towards Australia.

[38] It is possible that the SRC should consider adding those motives that panellists expressed for joining Life in Australia™, but were not mentioned previously in recruitment for Life in Australia™, to some of their future recruitment efforts; or at least train their interviewers to consider mentioning these reasons for why someone might join Life in Australia™.

Most commonly successful recruitment was a result of motivational factors associated with (1) *the topic of research*, for about two-thirds of all qualitative interview participants. Some of them liked or found important either social or political topics; for example, one participant was interested in what the survey was about, and one found it interesting to participate. The second most common factor convincing participants to join the panel was (2) *the sponsor* of the research, and most of them mentioned the (Australian National) University. Most participants reported that they would much rather participate in social or academic surveys than commercial surveys, with just a couple of exceptions. The other common reasons for joining the panel, each of them mentioned by about one fifth of participants, were (3) *to have a say*, (4) *to contribute to research*, and (5) *to receive incentives*. Two participants also mentioned that they wanted to (6) *represent their opinions or particular population subgroups.* Only one IDI participant did not remember anything about how s/he joined the panel.

**Examples of motivations for joining the panel**

*1) "[Topics] aligned with issues or questions that I would have myself about particular things."*

*2) "…the idea of providing academic data to an institute I trust [Australian National University] and have high regard for, make sense to me."*

*3) "Social things… that I thought those results might influence… there is often a disconnect between what politicians think…"*

*4) "…wanted to contribute some valuable information around the topics that are discussed…"*

*5) "…probably the reward at the end, it's not just the surveys, you know, lots of people like to get something out of it."*

*6) "I want people to know that there are other opinions out there and it's a way to try to balance, perhaps, what you're hearing."*

Evidence on intrinsic motivation for panel participation and survey completion is fairly consistent with the previously reported results based on survey data. About two-thirds of all participants reported (1) *contribution-focused* motivation, with *sharing views/opinions to make a difference, contributing to research,* and *sharing views to represent opinions* as the most notable motivational factors. Three of the IDI participants were academics/researchers, and they reported scientific or research-supporting motivation to participate. On the other hand, for two participants, we identified (2) *survey-focused motivation*, with *interesting/relevant topics/questions*, *being informed about topical issues,* and *surveys being*

**Examples of Intrinsic motivations for panel participation**

*1) "the fact that it felt like I was providing some beautiful information for an academic service and my bread and butter at work is academic."*

*"The main thing is that I am able to have a say and, sort of, have an input into what is actually happening…"*

*2) "I was very interested to see what was, sort of, topically being researched."*

*"…helps me to, sort of, define what I think about these issues perhaps… I always like doing that, try to think of that"*

*though-provoking* as the most commonly mentioned motivators. One participant reported (3) *self-expression-focused motivation*; and *having his/her voice heard* was generally the most common answer among those motivational factors. For the remaining couple of participants, we observed *incentives-focused* motivation.[39]

*3) "…being able to give my feedback, on different issues. We're having a voice to express our feedback on issues."*

*"Well, I think it's nice to have someone listening to your input."*

The evidence for the importance of offering rewards in return for panel participation is quite mixed. We identified receiving incentives at the main motivational factor for only one-fifth of participants. The other four-fifths provided different interpretations of the value of being offered something in return for participation. In terms of their perception of the importance of material rewards, we could classify our IDI participants into four distinctive groups of fairly equal sizes: (1) monetary rewards are quite important for their participation, (2) monetary rewards are not very important for their participation and they would likely participate without receiving them, (3) monetary rewards are not very important for their participation but they believe they are more important to other respondents, and (4) they exclusively donate their rewards to charities. Generally speaking, intrinsic motivation seems to be much more prevailing in explaining panel participation than extrinsic motivation.

**Extrinsic motivation for panel participation (the importance of rewards)**

*1) "…to get paid every survey, so that was like an incentive to me. You know, stay at home mom."*

*2) "…now I am committed to it, I would be disappointed [if rewards were not offered anymore], but I would probably continue to do it anyway."*

*"I collected all the vouchers and I never used any and I guess that was a bit of a waste, really… I would just forget them and not use them."*

*3) "The money incentive is, I guess, really helpful for a lot of people."*

*4) "I suppose it was just lovely to think that the donation was going to some charity."*

These results, also combined with the results from section 4.5.1, indicate that social exchange theory has a better potential to explain the role of incentives in a probability online panel context than economic exchange theory. Most panellists perceive monetary rewards as one of motivational factors, and not as a payment for filling out Life in Australia™ questionnaires (see further discussion in section 4.5.5).

#### 4.5.3.2 Barriers across different stages of panel lifecycle

[39] The classification of motivational factors into four main groups of motivations, *contribution-focused, survey-focused, self-expression-focused,* and *incentives-focused*, is fairly consistent with the results of PCA analysis (see Tables 9 and 10 in the Appendix 4). For example, *survey-focused motivation* is captured in Components 1 and 6.

The qualitative in-depth interview participants were also asked a number of questions about their barriers for participation (see Table 4.7 in the Appendix 4). The most commonly reported barriers or negative aspects of participation in Life in Australia™ surveys (or even surveys in general) were: (a) lack of time/being busy, (b) lack of information provided about the survey, (c) issues related to survey questions (format, repetition, content), (d) major life changes, (e) unappealing survey topics, and (f) participation in research that is either commercial or market research or not perceived as "legitimate". We will now present the identified barriers towards participation at different panel lifecycle stages: (i) in the recruitment stage, (ii) affecting survey questionnaire participation, and (iii) reasons for complete attrition (dropping out) from the panel.

An acknowledged limitation of our study is that we could not directly investigate barriers or reasons for not joining the panel directly, as we did not have the resources to sample and interview original nonrespondents from the time of the panel's formation. That said, we were able to identify potential barriers in the recruitment stage (see Table 4.7), which provided insight into how sampled respondents made their decision whether to join an online panel or not. The most common barriers or concerns in the recruitment stage identified by the participants were the (1) *the type of research* – about one-half of respondents explained that they would not (or less likely) participate in commercial and market research, or research not considered legitimate. On the other hand, one participant would rather participate in cross-sectional commercial surveys ("to stay out of politics"). The other commonly reported barriers were (2) *the lack of information provided about the questionnaire/panel*, and (3) *not knowing where the data were going*, including *privacy issues*.[40] The other less commonly reported barriers were (4) *not appealing/not interesting questionnaire topics*, and (5) *low motivation*, also due to *not having a very positive attitude towards surveys in general*.

**Examples of potential barriers in the recruitment stage**

*1) "… but if it changed, probably not [continue participating] if it was more commercial…"*

*"… and a lot of those commercial surveys are very manipulative in the way they're phrased."*

*2) "… maybe saying, like, this will go for another six months, then we'll change, giving people a bit more information upfront."*

*"I don't recall that [there was a voucher]… I would have printed that off and used it because I love to shop."*

*3) "Sometimes I get a little bit wary of anything online… you don't really know who you're talking to and where the information is going."*

*4) "I'd probably go, kind of, go more to commercial side than policy or politics… I tend to stay out of politics."*

*5) "It has to really interest me… to even want to do a survey, to be honest."*

[40] Recall that privacy and confidentiality concerns were essentially unmentioned in the survey verbatims. That is, essentially no one wrote anything to the effect that they "joined the panel because they were confident that their privacy would not be violated" and/or "that their confidentiality would be maintained."

Regarding barriers affecting questionnaire completion in the ongoing data collection waves, participants mentioned different negative aspects of Life in Australia™ survey participation, including the barriers that lead to not completing all questionnaires. The most common barrier was (1) *being busy or lack of time*, which was identified by about two-thirds of participants. It often resulted in panellists not participating in questionnaires until receiving a reminder(s), and it did not always lead to unit nonresponse. The other common barriers to questionnaire completion were directly related to *the questionnaires themselves*: (2*) the format, content or repetition of questions*, and (3) *not appealing questionnaire topics*, which were reported by about one-half and one-third of participants, respectively. The other barrier directly related to the questionnaires, albeit less common, was (4) *the length of surveys*. Again, barriers such as *not knowing where the data are going* and *lack of information provided about the questionnaire*, were mentioned as concerns in the questionnaire completion stage.

**Examples of barriers affecting questionnaire/wave participation**

*1) "It's all time, time, depending on how busy I am, or if I get to go to my email…"*

*"It was just a busy time and it was always a little challenging to get on and do it."*

*2) "… some of the questions were just too black and white. And it did require more context to answer."*

*"…[questions] the last few times were a bit difficult… but I thought they were very similar surveys... almost identical."*

*3) "Technology is not my field anyways, so I was, sort of, just putting in any answer I could think of.*

*"I think when it comes to voting or political, that sort of thing, I don't find that at all interesting."*

*4) "… last few surveys I did, took me a bit longer than usual."*

The decision to voluntary attrit (i.e., leave the panel), to not respond to any Life in Australia™ surveys after joining the panel (i.e., so-called "backouts"), or to stop responding, were mostly made based on a combination of barriers. The most commonly identified barrier was (1) *insufficient information provided about the panel*; two participants did not know it was a continuing study, one participant did not know they could skip a wave of data collection without opting out of the panel, and one participant did not know that there was an option to opt-out. For those who stopped responding to questionnaires after a period of participation, but did not opt-out, the most common factor was (2) *major life changes*, such as health issues, moving house, or other family/household changes. However, those major life events seemed to be the last straw while there were additional aspects of panel

**Examples of barriers leading to voluntary attrition**

*1) "…probably would like to know how often if would be, the commitment, if it was once a month, or…"*

*"I can't remember if you have to keep doing it or if you could've just gone: 'I am not doing that one', I can't remember."*

*2) "I actually had a heart attack last year."*

*"We had our house on the market, and we've since moved from New South Wales to Queensland."*

*"The kids now live with a partner [and incentives are no longer that important]."*

membership that participants were not satisfied with even before. Some other notable barriers leading to voluntary attrition or "backing-out" were: (3) *incentives related barriers*, (4) *content of questions*, and (5) *low motivation and/or unwillingness to commit* to Life in Australia™ surveys. One participant stopped responding due to technical issues after switching her/his device for providing data.

*3) "The awards, I don't believe anybody gets them."*

*4) "I feel like I am being steered in the response that I'm going to give… I didn't feel it [the third survey] was an honest and open a survey as the previous two…"*

*5) "I didn't want to commit to something long term… I've done the thing that I was going to do and I'm out."*

#### 4.5.3.3 Differences between panel participation groups in motivation and barriers

To answer the second research question on the differences between panellists with different participation pattern, we compared their main motivational factors (IDI), as well as their personalities (DiSC). As discussed previously, most IDI participants identified a number of different motivational factors (up to 10 of those listed in Table 4.6 in the Appendix 4), and the differences among panel participant groups in their motivation to join the panel and to respond to questionnaires were relatively minor. On average and to no surprise, frequent-respondents identified more motivational factors than those panellists who stopped responding and nonrespondents. At the same time, five of the six frequent-respondents reported predominantly *contribution-focused* motivation while the stopped-responding panellists and backout nonrespondents reported more evenly distributed types of motivation: *contribution-focused*, *survey-focused* and *incentives-focused* motivation. In terms of reporting individual motivational factors, only two of nine stopped-responding panellists and nonrespondents mentioned how they *share opinions to make a difference or influence change* (in contrast to five of the six frequent-respondents), while no nonrespondents specifically mentioned that they participate to *contribute to survey or research*.

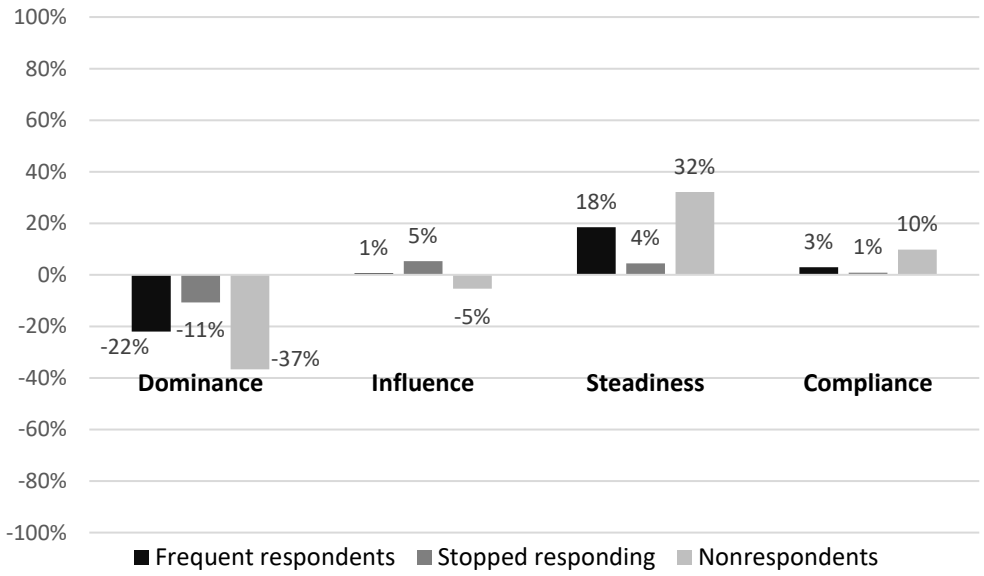
In contrast to reporting motivational factors, the distinction between the analyzed groups of panellists in their barriers for participation was clearer. First, frequent-respondents did not identify as many barriers and negative aspects of participation compared to stopped-responding panellists and nonrespondents. On the other hand, five of the six explained that they would not or were less likely to participate in *commercial/market/not legitimate research.* Second, stopped-responding panellists were different to the other two groups in the initial barriers, and also in the final barrier that influenced them to stop responding. They were less satisfied with *survey topics*, and *format, content or repetition of questions*, but the main reason why they stopped responding was *a major life event/change* (4 of 5) and *issues with technology* (associated with moving house). On the other hand, nonrespondents had very different reasons for their attrition. Besides *not having sufficient*

*information* about the panel and the questionnaires, they were either less motivated to participate in continuing research, did not want to commit, did not remember recruitment to the panel, or did not understand the difference between Life in Australia™ and market research surveys.

### 4.5.4 Differences between panel participation groups in their personality traits

After carrying out DiSC personality traits testing, we also identified minor differences in personalities between frequent-respondents, stopped-responding panellists, and nonrespondents (backouts and attritors).[41] Nonrespondents on average scored lower on Dominance and higher on Steadiness than frequent-respondents and stopped-responding panellists (see Figure 4.1). The two traits are diametrically opposite, and the theory explains that Dominance is associated with enjoying challenges, and Steadiness with getting restless and bored when involved in routine and repetitive activities. However, while the nonrespondent group was quite homogeneous in their personality traits, the other groups were less so. For example, there were three participants from the frequent-respondent group with high Steadiness and Conscientiousness and low Dominance trait scores, but two with high Influence and low Conscientiousness scores. However, the investigation into DiSC personality traits as predictors of panel participation seems to be an interesting topic for future research.

Figure 4.1: DiSC assessment, average scores for groups (adjusted, range -100% to 100%[42])




[41] We have to acknowledge the fact that those groups were very small in size, i.e., 5 frequent-respondents, 5 stopped-responding and 4 nonrespondents (with missing data for 1 frequent-respondent). Consequently, the observed differences are only indicative and should be interpreted in the context of a qualitative study. As personality testing was an exploratory component of this study, future research with larger (sub)samples (and proper statistical testing) is required to determine whether DiSC personalities are associated with panel response patterns.

[42] Value -100% is the lowest possible value (if all 28 statements associated with a particular personality trait would be selected as "describes me least") and value 100% is the highest possible value (if all 28 statements associated with a particular personality trait would be selected as "describes me most"). See Table 4.11 in the Appendix 4 for more information.

### 4.5.5 Social-psychological theories explaining panel participation

As addressed later, there are a number of social-psychological theories explaining survey participation. The results from our 15 qualitative interviews offer valuable evidence providing answers to the third research question on the ability of social-psychological theories explaining panellists' panel participation behavior. Instead of trying to define the theory which explains the recruitment outcome and a particular response trend of each interviewee the best, we will present collated evidence on each theory separately with quotes from our interviews. Coding frame presented in Table 4.8 in the Appendix 4 was used in qualitative data analysis.

1) Social exchange theory

In the interviews, the majority of participants listed a number of motivational factors as (non-monetary) rewards/benefits of their participation, while not many explicitly stated costs of participation, such as investing time and effort, or privacy concerns. However, Frequent-respondent 5 nicely outlined the cost-benefit balance explained by SET.

In a longitudinal format, the results from qualitative interviews provided evidence that SET can explain a change of behavior as a result of costs exceeding rewards after a period of time, since there were newly introduced barriers as "costs" of participation or the existing perceived costs increased. Nonrespondent 4 outlined the increase of the total perceived costs from the first two Life in Australia™ surveys to the third wave, after which the panellist opted-out.

**Frequent-respondent 5**

On rewards as an important component: *"Probably not the number one component, but it's good to know that if I'm giving the time [as costs], that I could probably do two things: assisting the survey and have a voice and potentially assist people maybe in greater need than I am [as rewards]."*

**Nonrespondent 4 (a voluntary attritor):**

On motivation: *"… a lot of social things that I thought those results might influence [politicians]"*
*"…also, the thing that you can donate money."*

On experience: "*I quite enjoyed doing them ... and I was quite looking forward to seeing the answers.*"

On new barriers: "*The third [wave] had a lot of political stuff… I felt like the questions were targeted at providing a negative answer*."

On costs exceeding rewards: "*… and I was really torn because I wanted to do it for that reason [donating to charities], but I just also started feeling: 'Oh, I feel like I am being steered in the response I was going to give…'."*

The role of incentives as a token of appreciation (similar to the perception of Frequent-respondent 5: *"Probably not the number one component…"*), which is how the effectiveness of rewards is explained by the social exchange theory, is later studied in Chapter 10. The respondents from the RDD SMS Web-push study were offered small conditional incentives (i.e., $5 supermarket coupons), which were hardly sufficient to be perceived as a payment for survey participation, which is how the use of incentives can be explained by the economic exchange theory.

2) Economic exchange theory

In contrast to SET, participants who viewed survey participation as economic exchange emphasized the importance of monetary incentives/coupons and saw receiving them as a form of payment (see the examples in the box). However, based on other motivational factors identified by respondents, it is hard to argue that survey participation has been strictly just a mechanism to earn money for any of the interviewed panellists in exchange for their participation. The majority of them reported either contribution-based or survey-based motivation, or perceived incentives as a token of appreciation and/or only as one of several motivational factors of which the majority was intrinsic.

**Frequent-respondent 3**

On previous participation in questionnaires: *"…in some cases you get bonus points for something that goes towards like a frequent flyer program, a store program."*

On importance of incentives: *"Some of them take a little longer [questionnaires], when I should be prioritizing other things. Well, at least I am getting something out of it at the end."*

**Stopped-responding 1**

On motivation at recruitment: *"… probably the reward at the end… lots of people like to get something out of it… getting a little pay… getting that little bit of extra money makes you want to do the [questionnaire]."*

**Nonrespondent 1 (a 'backout')**

On motivation to participate: *"Getting paid to do a survey."*

In practice, economic exchange theory has a better potential to be applied to participation in nonprobability online panels, especially those with a large proportion of so-called professional respondents (similarly to Nonrespondent 1 who reported *"getting paid to do a survey"* as the main motivational factor). In Chapter 9, we compare the accuracy of probability samples and nonprobability online panel samples. The theory (e.g., Baker et al. 2010) explains that nonprobability online panels do not produce as accurate survey estimates as those based on probabilistic sample selection. This in practice indicates that many panellists, whose survey participation could be primarily explained with the economic exchange theory, are fundamentally different to those whose survey participation could be primarily explained with the social exchange theory (or some other social-psychological theories explaining intrinsic motivation).

3) Self-perception theory

As the theory explains, people come to understand their own attitudes and interests by making inferences based on their (previous) behavior, which in the case of a survey questionnaire can explain how previous questionnaire completion affects future questionnaire completion. The results from qualitative interviews show how almost all participants reported previous questionnaire completion, from irregular to more regular, but only one participant reported previous panel participation. After regularly participating in different kinds of questionnaires, some Life in Australia™ panellists saw themselves as survey participants or as respondents supporting survey research for various reasons (after developing better understanding of the challenges of survey data collection). Also, after participating in a panel for some time, they can start seeing themselves as panel survey participants, which can prevent them from response inactivity or opting-out, no matter how infrequent their previous participation was.

**Frequent-respondent 2**

On previous participation in surveys: *"Yes, I did a few… the last I did was six months, six months or so."*

**Stopped-responding 3**

On experience completing questionnaires: *"My PhD was around interviews so, and I've put together [questionnaires]s myself for various things. Probably a moderate amount [participating in surveys]."*

**Nonrespondent 3 (a voluntary attritor)**

On frequency of previous participation: *"Not that often, no, just randomly… wouldn't be able to tell you how many times."*

**Frequent-respondent 1**

Motivation to respond: *"Oh, just being a part of the survey…"*

**Frequent-respondent 4**

On payment incentive: *"Now I am sort of, like, entrenched… I would probably continue to do it anyway [without rewards]."*

**Stopped-responding 4**

On questionnaire completion participation: *"I just built in into my routine…"*

Moreover, self-perception theory has substantial value in explaining the effectiveness of particular methodological solutions. In Chapter 5, we study the added value of so-called online panel paradata (i.e., a history of panellists' survey completion) in predicting future panel participation. This important topic can be, besides on some other theories investigated in this chapter, built on self-perception theory. The relationship between past participation and future participation could be a result of panellists seeing themselves as panel participants; for example, Frequent-respondent 1 explained that s/he was motivated by *"… just being a part of the survey [Life in Australia]"*, which resulted in continuing participation of that panellist with 100% completion rate at the time of recruitment.

4) Cognitive dissonance theory

In the interviews, not many participants explicitly expressed dissonance for not completing a questionnaire. However, while noncontingent/unconditional incentives have not been used in Life in Australia™, reminders were identified as a measure creating some form of dissonance in a longitudinal/panel design. They were reported to be effective to encourage frequent-respondents (i.e., completing all questionnaires) to not be skipping waves, as nonresponse would be a deviation from their panel behavior and their contributions to the panel. One stopped-responding panellist even indicated how reminders, which they were still receiving months after they stopped responding, created dissonance (stress/tension).

**Frequent-respondent 4**

On reminders: *"… and when you say that's nearing to the end [the questionnaire], I'm like, OK, I'm going to do it now. I'll just force myself. But that's probably how I work. I like to move things to the last minute."*

**Frequent-respondent 6**

On reminders: *"Phone calls are always pleasant, but they always feel a little bit unnecessary."*

**Stopped-responding 3**

On still receiving reminders: *"Yes, I am. I don't like getting those [reminders], but when all this COVID [is over] …"*

In contrast to this study, the value of unconditional incentives as a response maximization approach could be studied with data from worldwide probability online panels (Chapter 3). Comparing the value of unconditional and conditional incentives is, indirectly at least, comparing the application of cognitive dissonance theory opposite cost-benefit theories (social and economic exchange theories).

5) Reasoned action approach

To explain panel participation with the reasoned action approach, we tried to match participants' answers with as many concepts which the theory is based on as possible, including attitudes, norms, behavior control and behavioral intention/readiness. While no participant's verbatim response covered the whole model by identifying each of the concepts as applying to them, about one-half of participants reported intention to (continue) participate in Life in Australia™ waves, and about the same proportion of participants reported positive attitudes towards Life in Australia™ or surveys in general. That was clearly associated with continuing participation behavior, as frequent-respondents were more likely to report those intentions and positive attitudes towards survey research.

**Frequent-respondent 1**

Attitude, intention: *"Over the years I've done [questionnaires] off and on since I was in high school."*

*"Oh sure, yeah [will continue to respond in the future]."*

**Frequent-respondent 6**

Attitude, intention: *"My experience for the [panel] has been entirely positive."*

*"I'm fascinated by data and fascinated by social research data."*

*"It's not on my radar to stop doing it."*

**Frequent-respondent 3**

Control: *"So might not be able to do it that day or the following day… have to wait for that reminder to come through, just depending of what's happening with life."*

Some participants also reported control over participation – actual control and perceived behavioral control. The reasoned action approach also can explain a behavioral change as a result of changed beliefs, attitudes, norms, or intentions. This is especially relevant for panel and other longitudinal studies which can change data collection or panel characteristics over time. For example, Nonrespondent 4 (see sidebar quotes) who opted out after their third wave initially had a very positive attitude towards Life in Australia™. However, the third questionnaire completely changed her/his attitude and beliefs about the direction of the panel survey, and the intention to continue participating in the panel, which resulted in voluntary attrition.

**Nonrespondent 3 (a voluntary attritor)**

Control: *"…knowledge that you could stop [participating in panel] whenever you wanted to, was reassurance enough."*

**Nonrespondent 4 (a voluntary attritor)**

Change of attitudes, beliefs and intention: "*I didn't feel it [the third survey] was an honest and open a survey as the previous two…"*

*"…I just also started feeling: 'Oh, I feel like I am being steered in the response I was going to give…'"*

*"I revisited… I closed and restarted it [the third questionnaire]... and that second time gave me a sense I was right the first time."*

*"…and I just went: 'No, I am not doing it.'"*

Similar to self-perception theory, reasoned action approach is another theory which can be applied to explain future participation based on past panel participation trends. As defined by the theory and confirmed with our IDI data (e.g., for Nonrespondent 4), a change in panel response behavior could indicate a change of intentions or even attitudes towards the panel/survey, which are predictive of future panel participation. We use a number of behavioral change indicators, which were derived from online panel paradata, as predictors of future nonresponse and voluntary attrition in regression models in Chapter 5.

6) Leverage-salience theory

In the IDIs, we could identify the principles of the LST in both recruitment and wave-by-wave data collection stages of the panel lifecycle. When being asked to join the panel, respondents often reported different panel survey attributes that tipped the scale in 'successful-recruitment' direction. For example, in the case of Frequent-respondent 6, the survey sponsor/authority, topics and content, and proper use of data, were the attributes that convinced them to join the panel.

**Frequent-respondent 6**

Recruitment: *"The idea of providing academic data to an institute that I trust… contents of the surveys… good balance between a relevant, and weighty [questionnaires] … the opinion is used in a valuable way."*

*"I am very sensitive to misuse of sensitive data."*

**Stopped-responding 1**

Questionnaire participation (on why stop responding): *"Last few [questionnaires] I did, took me a bit longer than usual… it's just some of the questions… would be good if some of the questions gave 'unsure' [answer option]… that last few I did were a bit difficult [questions]… I thought they were quite similar [questionnaires]…"*

**Nonrespondent 4 (a voluntary attritor)**

Recruitment: "*…because it was across a range of topics… I thought those results might influence… also the thing we could donate money…"*

Panel participation (on why opting-out): *"The questions [in the third wave] were targeted at providing a negative answer. And I felt very strongly about it… I remember thinking there was no response that was representative. And I went quite a way through [the questionnaire]... 'I don't like that question, the way it's worded, I don't like potential answers'…"*

In a longitudinal format, introducing new recruitment/ maintenance attributes with high leverage but with negative disposition could lead to nonresponse and attrition later in the panel lifecycle. The evidence from the IDI with Stopped-responding panellist 1 reveals other attributes affecting the decision to participate in future waves, such as question format, question difficulty, or repetitiveness of questionnaires. The type of questions, their content and answer options were also the reason why Nonrespondent 4 opted-out of the survey, while initially enjoying the topics and participation in the first two waves, as well as appreciating a chance to donate to charities. However, they felt strongly about the inadequacy of survey questions (high leverage) which could not be balanced out by other attributes with positive disposition.

Having evidence that the leverage-salience theory could be applied in a longitudinal context, as well as in self-administered surveys, it is prudent to test the effectiveness of communicating different survey attributes with a positive disposition in the survey recruitment stage. While it is not related to longitudinal panel participation, the study on response in the RDD SMS Web-push survey is somewhat associated with the LST. In Chapter 10, I tested the effect of different survey attributes (i.e., types of incentives, topic, benefits, research sponsor), as well as their magnitude, on recruitment outcomes.

7) Compliance heuristics

With the qualitative interviews, we did not find evidence that participants invested a lot of time and cognitive effort into making a decision to join the panel or participate in surveys. In contrast, we observed a number of compliance heuristic principles responsible for successful recruitment and minimizing survey nonresponse. We can argue that all principles except for Reciprocation played an important role in recruitment, but some also in wave completion. The most commonly identified heuristic principle in the recruitment stage was Authority – more than one-half of participants reported that their decision to join the panel was positively affected by the authority/research organization (i.e., "ANU/Australian National University", "university", "academic institution"). Moreover, Commitment seemed to be the most important principle applied in the questionnaire completion stage – two-thirds of all participants and all frequent-respondents expressed some form of commitment (previous, current, future) and consistency in participation. On the other hand, Liking, Social Proof and Scarcity were less common principles that lead to positive recruitment (or survey participation) outcome.

**Frequent-respondent 1**

Authority: "*Because it came through from the ANU [Australian National University], it was a part of that survey, I thought, it's a university-based survey not a marketing survey…*"

**Frequent-respondent 6**

Liking: *"…persuaded by a very pleasant, very articulate, very nice individual…"*

**Nonrespondent 4 (a voluntary attritor)**

Social proof: "*I am trying to remember if someone told me about it as well… I've got this funny feeling that one of my friends is doing it… that might have been how I got on, actually… I am pretty sure it might have been a word of mouth.*"

**Stopped-responding 3**

Scarcity: *"We're living in a time where we don't really get that, that opportunity to put out thoughts forward or voice forward and I thought this was an opportunity to actually have some say in, in changing things, the way the things are shaped at the moment."*

**Stopped-responding 2**

Commitment and consistency: *"Yeah, yeah [will continue participating]… I don't understand why people would leave the survey."*

Commitment as one of compliance heuristics principles could be another social-psychological theory used to explain the association between past and future panel participation. As reported by one of the panellists: *"…I don't understand why people would leave the survey…"*, commitment could result in continuing participation in panels; or better to say, longer interruptions in participation in panel surveys could indicate a lack of commitment and may result in continuing non-completion or voluntary attrition. Several indicators which identify panel behavior patterns that signal lower levels of commitment are used as predictors of nonresponse and attrition in regression models in Chapter 5.

The evidence from IDIs on social-psychological theories explaining recruitment outcomes, wave participation, and panel behavior trends reveals how different theories can be applied to justify decision-making of the same panellist. This is similar to how the same methodological solution (e.g.,

predicting future participation with past panel behavior/online panel paradata) could be based on multiple theories. We identified notable overlap between the theories in practice, as the same motivational factors are in fact elements of different theories; for example, authority/sponsor acts as a motivational attribute in leverage-salience theory and as a principle in compliance heuristics. Thus, the recruitment outcome of Frequent-respondent 6 can be explained with both *leverage-salience theory* (balancing survey/panel attributes such as sponsor, topics, proper data use), and *compliance heuristics* (authority and liking), whereas their wave participation can be explained with the *reasoned action approach* (positive attitude, intention to participate, actual 100% completion rate), among others. On the other hand, attrition or nonresponse of particular participants could also be explained with different theories. A good example of that is Nonrespondent 4 whose decision to opt-out could be explained with *social-exchange theory* (having a say and donations as rewards, newly introduced costs exceeding rewards in the form of the topic and particular question format), *reasoned action approach* (changed beliefs, attitudes, intentions to participate for the same reason), and *leverage-salience theory* (survey attribute with high leverage and negative disposition, i.e., topic and question format, tipping over the scale despite of having other interesting topics and donations at the other end of the scale). This should be considered good news since each theory offers different recruitment and panel management solutions, and for some respondents, at least in theory, there are several effective strategies available to increase response or mitigate attrition.

## 4.6 Discussion

We believe that the mixed-method study presented in this chapter addresses important knowledge gaps in empirical research on the social-psychological theoretical aspects of survey participation. Building on similar studies such as those from Groves et al. (2000), Albaum and Smith (2012), and Keusch (2015), this study is the first to our knowledge to attempt to apply relevant theoretical and conceptual response behavioral frameworks to probability-based online panels. Using survey data (i.e., coded verbatims to an open-ended question), quantitative results from a personality test, and qualitative data (i.e., from IDIs), this research helps to identify both motivations and barriers to survey participation in a longitudinal context, and shows how they are linked to various social-psychological theories.

The evidence from this research confirms how several social-psychological theories, including LST, should not be limited only to explaining cross-sectional recruitment outcomes. As we have demonstrated, these theories can (and should) also be applied to explaining panel/longitudinal survey participation. What is necessary for social-psychological theories in a longitudinal context, is that they should be flexible enough to explain trends and changes in survey response behavior over time. Based

on our results, both traditional behavioral change theories (e.g., the theory of planned behavior/reasoned action approach; Fishbein & Ajzen 2011) and other investigated theories (e.g., SET, LST) have been shown to be sufficiently robust to help understand the time dimension of survey participation. And that finding has important practical implications for survey practitioners. For example, online panel organizations have to consider continuously offering rewards which are at least equal to costs of participation (SET; Blau 1964). They also have to consider that newly-introduced survey characteristics (e.g., sensitive topics) could unsettle the cost-benefit balance, discourage panellists from completing a wave, or even result in opting-out (Groves et al. 2000).

One of the most important implications from the analysis of IDI data is that it is difficult to rule out any theory when trying to explain behavioral decisions about the participation of an individual respondent/panellist. When analyzing qualitative interview data, it is much more applicable to identify elements, principles, or concepts of different theories in participants' answers. If an IDI participant does not provide answers which could be linked to a particular theory, it does not necessarily mean that that particular theory could not explain some participation outcome, trend or change. Inability to identify all theories explaining a participant's panel response behaviors can also be perceived as a limitation of qualitative interviews, in which a limited amount of information can be provided. We also found that it was challenging to identify the theory that would best explain participation of an individual panellist, which is consistent with what Keusch (2015) reported. The reason for that was that multiple theoretical elements, principles, or concepts were identified in most of our qualitative IDIs. This suggests to us that effective online panel management solutions might well be based on essentially any of the social-psychological theories studied in this paper, or even more accurately, that applying a combination of theories should be considered. For that reason, several of the studies in this thesis investigate survey/panel recruitment and panel management solutions that are more or less based on the theory of survey participation, which has origins in social-psychological theory.

Moreover, our findings related to panel participation motivation are less consistent with economic exchange theory (or potentially SET) than we expected (cf. Dillman et al. 2014), especially when in comparison to the role of incentives in cross-sectional probability-based surveys (see Chapter 3). Given that only one in four respondents reported incentives-based motivation to respond to questionnaires, online panel managers should consider adjusting incentive protocols so as to redirect the money saved into other panel management solutions. For example, using lottery/prize draw incentives may be a much less expensive response maximization option for a panel, given that about seven out of eight panellists that were studied in a survey did not report being primarily motivated by receiving supermarket coupons or donating the same amount to charity. Thus, online panels could offer their frequent-respondents a chance to participate in prize-drawings instead (i.e., providing the

value of the prize and the likelihood of winning it were appealing to the frequent-respondents). On the other hand, we identified what appear to be more influential contribution-based motivations and survey-based motivations, such as “having a say” (i.e., letting their voice be heard) or interesting survey topics. Thus, information about these potential motivations, should be properly communicated by panel managers to panellists at various stages of the panel lifecycle, including at the time of recruitment.

Our detailed and methodical analysis of online panel recruitment materials revealed interesting phenomena. As we noted, our findings are possibly confounding in relation to the correspondence between (a) what the online panel organization told sampled respondents during recruitment in trying to convince them to join the panel, and (b) what the panellist reported in their verbatims about why they joined the panel, even though for the vast majority of panellists this occurred a few years apart. However, we believe it is highly unlikely that this played a major or direct role in influencing what the panellist reported in their verbatims. That is, we do not think it is likely that the panellists merely “parroted back” what they recalled from the communications with a recruiter or from other communications from the survey organization as to why they might want to join the panel. Instead, we believe it is much more likely that certain persuasive statements used by the recruiting organization to gain cooperation resonated with pre-existing attitudes and beliefs within those who joined the Life in Australia™ panel. Thus, those sampled people for whom these motivations were inherently important were more likely to join the panel than were those for whom these motivations were unimportant. This reasoning closely follows from LST (Groves et al. 2000).

Furthermore, we identified a number of new motivations which could be communicated to sampled respondents at future recruitment waves/occasions. The majority of them were those which we grouped into survey-focused motivation. While we found elements of contribution-focused, incentives-focused and self-expression-focused based motivations in Life in Australia™ recruitment, their recruitment communication did not contain messages that panellists will enjoy surveys or participating in research on interesting topics, that participation with be thought-provoking, and how they will get informed about topical issues. As survey-focused motivation was the second most important broader group of motivational factors (after contribution-focused motivations), our findings offer opportunities for probability-based online panels to include those motivations in their recruitment communication with an aim to increase their recruitment rates.

Besides motivations, one of the most important contributions of our study is identification of barriers to continuing panel survey completion in a longitudinal design. We realized that online panel organizations often have little control over barriers that can lead to continuous nonresponse or voluntary attrition, such as major life changes including moving house/city/state, divorce, or illness.

On the other hand, panels have more control over the amount and quality of information provided to panellists, including survey and data collection characteristics such as survey topics, question format, and questionnaire length. We would recommend probability-based online panels providing more information about their panel, continuing survey completion, and results from studies their panellists participated in (more effective communication), as well as collecting data on survey characteristic preferences to achieve better long-term engagement of their panellists (efficient use of provided feedback).

We have to note that the findings would be more robust if we, similarly to the study from Brüggen et al. (2011), studied motivations and barriers by collecting additional closed-ended survey data on those two topics. In this study, we introduced structured lists of potential motivational factors and barriers to participation in probability-based online panel research. These could be used to design structured survey questions which would be asked the whole online panel. Besides conducting statistical testing, we could associate motivations and barriers with panellists' socio-demographic characteristics. In addition, as we noted previously, an important limitation of our research is that we were unable to gather equivalent data about barriers (or motivations) to panel membership/participation from a sample of those who were sampled originally to join the Life in Australia™ panel but failed to do so. Although our IDIs with frequent-respondents, the stopped-responding cohort, and the nonresponding backouts and attritors provided insights into what they viewed as barriers to initial membership and continued participation in the panel, it remains uncertain whether those who were sampled but never agreed to join the panel may have held similar or different views about the pluses and minuses associated with panel membership and participation. Past research about nonresponse in panels (Lavrakas et al. 2012) identified inadequate incentives and privacy concerns as prominent reasons that people do not join or maintain their participation in panels. In light of these barriers essentially not being mentioned by our panellists, we can assume that their own motivations and barriers differ in nonignorable ways from the cohort of persons who never joined the panel in the first place. This is a very important area for future research. However, follow-up studies of nonrespondents are both quite expensive and may fail to gain cooperation from a representative sample of the original nonrespondents (cf. Lavrakas et al. 2021) possible negating the value of the follow-up study.

Lastly, throughout the review of survey participation literature, analysis of the data, and the review of recruitment materials, we did not observe a clear and consistent link between social-psychological theories, and recruitment and response maximization strategies. Thus, we agree with Dillman (2020), at least to some extent, who speculated that theories of response behavior are dated and survey invitations in the 21st Century are not properly guided by the theory. Therefore, future research on the effectiveness of recruitment strategies which are more based on survey participation theories are

needed, and this general recommendation should not be limited to probability-based online panel research. Moreover, knowing more about the association between motivations to participate in panel surveys, barriers, and socio-demographics characteristic of panellists (as discussed previously) would uncover both new opportunities for targeted recruitment, as well as potential treatments for representation bias. While this study already provides valuable evidence for online panel organizations to understand their panellists' response behavior better, it also identifies a number of potential practical solutions worth exploring further.

## 4.7 References


Albaum, G., & Smith, S. M. (2012). Why people agree to participate in surveys. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 179-193). Springer.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychology Review*, *74*, 183-200.

Biner, P. M., & Kidd, H. J. (1994). The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates. *Psychology & Marketing*, *11*(5), 483-492.

Blau, P. M. (1964). *Exchange and power in social life.* John Wiley.

Bosnjak, M., Tuten, T. L., & Wittmann, W. W. (2005). Unit (non) response in web-based access panel surveys: An extended planned-behavior approach. *Psychology & Marketing*, *22*(6), 489-505.

Boulianne, S. (2008). Incentives. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 328-331). Sage.

Brüggen, E., Wetzels, M., De Ruyter, K., & Schillewaert, N. (2011). Individual differences in motivation to participate in online panels: the effect on response rate and response quality perceptions. *International Journal of Market Research*, *53*(3), 369-390.

Cialdini, R. B. (1987). *Influence* (Vol. 3). A. Michel.

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

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons.

Dillman, D. A. (2020). Towards Survey Response Rate Theories That No Longer Pass Each Other Like Strangers in the Night. In P. S. Brenner (Ed.), *Understanding Survey Methodology* (pp. 15-44). Springer.

Everything DiSC. (n.d.). *About Everything DiSC: Theory and Research*. Retrieved May 1, 2021, from https://www.everythingdisc.com/EverythingDiSC/media/SiteFiles/Assets/History/Everything-DiSC-resources-aboutdisc.pdf

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fishbein, M., & Ajzen, I. (2011). *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological assessment*, *4*(1), 26-42.

Goyder, J., & Boyer, L. (2008). Social exchange theory. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 827-829). Sage.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475-495.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *The Public Opinion Quarterly, 64*(3), 299-308.

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

Hansson, I., Berg, A. I., & Thorvaldsson, V. (2018). Can personality predict longitudinal study attrition? Evidence from a population-based sample of older adults. *Journal of Research in Personality*, *77*, 133-136.

Haunberger, S. (2011). Explaining Unit Nonresponse in Online Panel Surveys: An Application of the Extended Theory of Planned Behavior. *Journal of Applied Social Psychology*, *41*(12), 2999-3025.

Holland, J. L. (1966). *The Psychology of Vocational Choice*. Blaisdell.

Holland, J. L. (1985). *Making Vocational Choices*. Prentice-Hall.

Holmberg, A., Lorenc, B., & Sweden, S. (2008). Understanding the Decision to Participate in a Survey and the Choice of the Response Mode. *Proceedings of Q2008 European Conference on Quality in Official Statistics*, 1-9.

Homans, G. C. (1991). Social behavior in elementary forms. Harcourt, Brace & World.

Jones, C. S., & Hartley, N. T. (2013). Comparing correlations between four-quadrant and five-factor personality assessments. *American Journal of Business Education*, *6*(4), 459-470.

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

Keeter, S., Hatley, N., Kennedy, C., & Lau, A. (2017). *What low response rates mean for telephone surveys.* Pew Research Center.

Keusch, F. (2015). Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, *65*(3), 183-216.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*(3), 213-236.

Lavrakas, P. J. (2008). Economic exchange theory. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 219-220). Sage.

Lavrakas, P. J., Dennis, J. M., Peugh, J., Shan-Lubber, J., Lee, E., & Charlebois, O. (2012, May 17-20). *Investigating Nonresponse Bias in a Nonresponse Bias Study* [Conference presentation]. 67th Annual Conference of the American Association for Public Opinion Research, Orlando, United States of America.

Lavrakas, P. J., Zuckerberg, A., & Megra, M. (2021, May 11-14). *Investigating Nonresponse and Nonresponse Biases in a Nonresponse Follow-up (NRFU) Study* [Conference presentation]. 76th Annual Conference of the American Association for Public Opinion Research, Virtual.

Lugtig, P. (2014). Panel attrition: separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research, 43*(4), 699-723.

Marston, W. M. (1928). *Emotions of normal people.* Harcourt, Brace and Co.

Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement and personality. *Research in higher education*, *46*(2), 127-152.

Roller, M. R., & Lavrakas, P. J. (2015). *Applied qualitative research design: a total quality framework approach*. Guildford Press.

Seifert, T. (2008). Leverage-Salience Theory. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 423-425). Sage.

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

Trussell, N., & Lavrakas, P. J. (2004). The influence of incremental increases in token cash incentives on mail survey response: Is there an optimal amount?. *Public Opinion Quarterly*, *68*(3), 349-367.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*(2), 231-263.

## Appendix 4

### Classification of panellists

Table 4.5: General classification of probability-based online panel respondents (partially based on literature review)

| **Level 1** | **Level 2** | **Level 3** | **Level 4** | **Participation behavior (trend)** | **Main panel management objective** |
|---|---|---|---|---|---|
| Recruited and profiled online panellists (registered panellists) | Participants | Stayers | 100% respondents ('gold star' respondents) | Completion rate (COMR) 100% | Maintaining high participation |
| | | | Frequent respondents | COMR 67 - 99% | |
| | | Lurkers | Less frequent respondents | COMR 33 - 66% | Increasing participation |
| | | | Infrequent respondents | COMR <33% | |
| | Nonparticipants | Stopped responding without opting out | Sleepers | 3 consecutive nonresponding waves | Reactivation |
| | | | Dozers | 4 consecutive nonresponding waves | |
| | | | Comatose | 5+ consecutive nonresponding waves | |
| | | Attritors | Never responded after recruitment ('backouts') | Never opted out | Initial activation |
| | | | Fast voluntary attritors | Opting out in 6 months after recruitment | Delaying or preventing opting-out |
| | | | Gradual voluntary attritors | Opted out in >6 months after recruitment | |
| *Recruited but not profiled respondents* | | | | | |
| *Not recruited respondents* | | | | | |

## Semi-structured qualitative interview - list of questions

**1. Introduction**

1.1 Can you tell me a bit about yourself? *Probe for any background info i.e., age, employment status, studying, etc.*

**2. General background on survey participation**

2.1 How often do you participate in surveys? prior to joining Life in Australia™?

2.2 What kinds of surveys have you participated in? Prompt: telephone, in person, mail, longitudinal, opt-in panels

2.3 What attracted you to the surveys you have participated in? How do you choose which surveys to participate in?

2.4 Have/why have you declined to participate in certain surveys?

2.5 Are you a member of any other online panels? *(Similar to Life in Australia™) (If YES, how do you generally find this experience?) If needed clarify that an 'online panel' means that participants are asked to respond to multiple surveys over time and can sometimes measure change over time.*

2.6 Generally speaking, what would you say are the mains aims of surveys?

**3. Initial experiences engaging with Life in Australia™**

3.1 When were you first contacted by the Social Research Centre about the Life in Australia™ study? How were you first contacted?

3.2 How did you feel about being contacted about the study? Is there anything about that first contact you would have like to be different?

3.3 What initially appealed to you about being part of the Life in Australia™ surveys? Did you have any initial hesitations?

3.4 What appealed to you in particular about being part of Life in Australia™?

3.5 What were your initial expectations after joining the panel? Did you feel like you had received enough information about the survey?

**4. Participation in Life in Australia™ surveys (for Frequent respondents, Stopped-responding, Attritors)**

4.1 How many Life in Australia™ surveys do you recall taking part in? Which topics?

4.2 How did you usually undertake the survey? Which device? Mobile phone or computer?

4.3 Would you usually complete the survey in one go or come back to it?

4.4 How would you describe the Life in Australia™ surveys overall? The topics? The length of surveys? Relevance? Impact?

4.5 While you were a panellist, what did you enjoy about the Life in Australia™ survey?

4.6 Which aspects of the surveys did you not enjoy? *Probe: Could you tell me a bit more about what you didn't like about it? THEN If you could, how would you change <unenjoyable aspect>?*

4.7 Once notified about an upcoming monthly survey, how would you decide whether to participate or not? Why? Amount of time?

4.8 How did you feel about receiving reminders to complete the surveys? *If needed probe for: Were there too many/not enough? Wrong mode/wrong tone?*

4.9 What kinds of things motivate you to respond to each survey? (for Frequent respondents) *Probe: timing, availability, topics, interest, user experience, social desirability/compliance.*

4.10 What barriers were there to responding to each survey?

4.11 How important is the voucher provided to thank you for completing each survey? What is it that makes this <important/not important> to you? How would you normally use this voucher? Why?

**5. Voluntary Attrition (only for those who opted-out)**

5.1 When did you decide to opt-out of/stop responding to Life in Australia™?

5.2 How long had you been considering opting-out?

5.3 Why did you decide to stop doing the surveys or being part of Life in Australia™? Had something changed over the course of your participation?

5.4 Did the survey differ in any way from your initial expectations? When you joined the study, were you expecting to be responding to surveys on a monthly basis?

5.5 Would you consider responding to a Life in Australia™ survey in the future?

**6. No participation after recruitment (only for those who never responded after recruitment)**

6.1 When did you decide to not participate in Life in Australia™ surveys?

6.2 Why did you decide to not participate in the surveys? Had something changed after you agreed to join Life in Australia™?

6.3 When you joined the study, were you expecting to be invited to surveys on a monthly basis?

6.4 Would you consider responding to a Life in Australia™ survey in the future?

**7. Wrap up (those who never responded after recruitment are only asked 7.1- 7.2, 7.6-7.7)**

7.1 Overall, do you think the Life in Australia™ survey is or is not a worthwhile initiative? *Probe: What makes you think it is <worthwhile/not worthwhile>?*

7.2 How do you think the Life in Australia™ survey data are used?

7.3 Has the experience of participating in the panel met your expectations?

7.4 Did you feel like you were a part of something, being involved in a national survey like this?

7.5 How to you feel about the feedback you receive about Life in Australia™ survey results? Would you like to see more or less of this?

7.6 Is there anything you and think of that might encourage people to participate or to stop people from leaving the survey?

7.7 Any other thoughts that we haven't covered?

## Qualitative data analysis – coding

Table 4.6: Coding frame for motivation with quotes from verbatims (an open-ended question)

| No | Code | Examples, quotes |
|---|---|---|
| 1 | Sharing views/opinions to make a difference or influence change | "I have currently become an Australian citizen, so would like to have a say."<br>"A chance to contribute towards social change. I have a unique view…"<br>"…if we all share our views it can only help form a better Australia for us…"<br>"Being able to voice interests and influence change positively" |
| 2 | Sharing views/opinions to represent others like me/population subgroups/minorities | "Adding my views increases the diversity and range of answers…"<br>"…to voice an opinion that may be different to those in the main capital cities."<br>"to have my views included in a survey… good balance of other demographics."<br>"to inform others on what may otherwise be an unheard opinion" |
| 3 | Sharing views/opinions in a non-judgemental platform | "…i never feel judged or made to feel my answers are wrong"<br>"I am having my say honestly, without being judged, questioned or someone wanting to argue…"<br>"being somewhat anonymous helps me open up and share my true perspectives"<br>"I get to express exactly how I feel … without fear of being labelled racist" |
| 4 | Self-actualization, allows my voice to be heard | "A chance to air my views"<br>"Allowing my opinions to be heard"<br>"…always good to be asked your opinion"<br>"An opportunity to say your point of view" |
| 5 | My opinions are valued, appreciated, taken into account | "My opinion is valued."<br>"…and also a sense of being valued."<br>"Hope that my input is valued"<br>"your take on things is taken into account" |
| 6 | Like being part of something, part of a team | "gives me a sense of participation in expressing my views and opinion"<br>"Being able to be a part of a long-term survey which can measure views…"<br>"Being a part of Life in Australia™ makes me feel like a valued member of society"<br>"…and gives a sense of involvement in my country." |
| 7 | Thought-provoking to participate in the panel | "makes me think about things that I probably would not consider"<br>"It helps me focus on what i think"<br>"…challenged to take a position on an issue I may not necessarily have considered"<br>"I appreciate the way i am made to think about how i feel" |
| 8 | To be informed about topical issues | "I managed to get a glimpse of issues which matter to Australians"<br>"Curious about type of questions and the research being done"<br>"am interested to see which topics are deemed to be of interest"<br>"…provide me an insight into the narratives permeating through our communities" |

| | | |
|---|---|---|
| 9 | Receiving incentives | “Besides the financial motivation”<br>“Cash rewards”<br>“Coles voucher”<br>“Small payments help!” |
| 10 | Donating to charity | “a bit of money for a charity without too much effort”<br>“something that just costs me time but contributes to a charity each month”<br>“Mainly the charity donations.”<br>“…and nice to donate money to good cause” |
| 11 | Enjoying surveys/participating in research | “I find it interesting…”<br>“I think it is absolutely very interesting to participate in the surveys”<br>“I enjoy partaking in these surveys,” |
| 12 | Interesting topics | “I enjoy the questions a good test for my memory”<br>“I enjoy reading and answering the questions about different topics”<br>“I find the questionnaires and topics questioned on interesting.” |
| 13 | Social Research Centre/Life in Australia™ are reliable enterprises | “the researching institutions are reliable and do good work”<br>“I generally find Life in Australia™ to be well conducted.”<br>“The Life in Australia™ surveys cover a wide variety of subjects and are associated with reputable organisations and institutions…” |
| 14 | Contributing to the survey/study/research/science | “A chance to provide input for surveys that are done by experts for research”<br>“As a scientist I appreciate the value of data, and I am happy to provide it.”<br>“Contribute to research and the evidence base”<br>“Data is king. Topics seem relevant.” |
| 15 | Positive sentiment towards Australia | “Because i am proud and happy to be in Australia”<br>“It’s a good country to live and work and retire.”<br>“Being able to represent my country”<br>“Because I love this country, I want to keep it as Australia,” |
| 16 | Have the time/something to do | “Being retired I have the time to participate”<br>“Something to do…”<br>“I am retired and have time available to take these surveys.” |
| 17 | Other motivation | “Because it takes the pulse of life in Australia”<br>“Interest in public issues.” |
| 18 | No answer on the motivation | “area and style of questions methodology some loaded”<br>“At this point not much improvement in the Australian e-commerce”<br>“I am glad to be part of this study”<br>“It feels right” |

Table 4.7: Coding frame for barriers

| No | Code |
|---|---|
| 1 | Negative attitude towards commercial/market surveys |
| 2 | Negative attitude towards most surveys in general |
| 3 | Cannot distinguish between Life in Australia™ and market surveys |
| 4 | Lack of information provided about the survey |
| 5 | Not knowing where the data is going, privacy concerns |
| 6 | Major life events/changes |
| 7 | Topics they don't like/enjoy/have no opinion on |
| 8 | Difficult, repetitive, biased, hard to answer questions |
| 9 | Length of surveys |
| 10 | Being busy, lack of time, other activities more important |
| 11 | Low motivation to participate |
| 12 | Lost track of the survey |
| 13 | Not willing to commit |

Table 4.8: Coding frame for indicators of social-psychological theories

| No | Final code | Theory |
|---|---|---|
| 1 | Listing rewards and costs of panel participation (as exchange) | SET |
| 2 | Mentioning a new negative aspect of panel participation (resulting in change) | SET |
| 3 | Mentioning a new positive aspect of panel participation (resulting in change) | SET |
| 4 | Perceiving incentives as payment for participation or the central benefit/motivator | EET |
| 5 | Incentives do not seem to be high enough to participate in panel surveys | EET |
| 6 | Reporting past survey participation behavior | SPT |
| 7 | Seeing themselves as survey respondents/panel members | SPT |
| 8 | Positive attitude towards panel/surveys/research | RAA |
| 9 | Expressed perceived control over survey/panel participation | RAA |
| 10 | Expressed behavioral intention to participate in (panel) surveys | RAA |
| 11 | Expressed actual control/ability to participate in (panel) surveys | RAA |
| 12 | Changed behavior due to a change of beliefs, attitudes, and/or intentions | RAA |
| 13 | Reported dissonance for not completing a survey/participating in panel surveys | CDT |
| 14 | Had to be reminded until participating | CDT |
| 15 | Changes in response due to (a) particular survey attribute(s) | LST |
| 16 | Evaluation of survey/panel attributes with positive and negative dispositions | LST |
| 17 | Identifying an authority as a motivational factor for joining the panel/participating | CH (A) |
| 18 | Reported commitment to the panel, consistency in participation, involvement | CH (C) |
| 19 | Reported liking those who survey/panel request came from as a motivational factor | CH (L) |
| 20 | Participation as a repayment for something they were provided | CH (R) |
| 21 | Interpreting joining the panel/participation as a (rare) opportunity | CH (S) |
| 22 | Survey/panel participation as a result of acting like others are/social evidence | CH (SP) |

SET - Social-exchange theory, EET - Economic-exchange theory, SPT - Self-perception theory, CDT - Cognitive dissonance theory, RAA - Reasoned action approach, LST - Leverage Salience Theory, CH - Compliance Heuristics, A - Authority, C - Commitment, consistency, involvement, L - Liking, S – Scarcity, SP – Social proof, R – Reciprocation

## Principal Component Analysis

Table 4.9: Principal Component analysis, components and eigenvalues

| Component | Eigenvalue* | % of Variance | Cumulative % |
|---|---|---|---|
| 1 | 1.333 | 8.328 | 8.328 |
| 2 | 1.248 | 7.799 | 16.127 |
| 3 | 1.167 | 7.297 | 23.424 |
| 4 | 1.153 | 7.205 | 30.628 |
| 5 | 1.115 | 6.968 | 37.597 |
| 6 | 1.081 | 6.757 | 44.353 |
| 7 | 1.046 | 6.535 | 50.888 |
| 8 | 1.009 | 6.306 | 57.194 |
| 9 | 0.996 | 6.223 | 63.417 |
| 10 | 0.982 | 6.139 | 69.556 |
| 11 | 0.968 | 6.049 | 75.606 |
| 12 | 0.937 | 5.857 | 81.463 |
| 13 | 0.905 | 5.653 | 87.117 |
| 14 | 0.854 | 5.338 | 92.455 |
| 15 | 0.841 | 5.259 | 97.713 |
| 16 | 0.366 | 2.287 | 100 |

*we selected components with eigenvalues > 1

Table 4.10: Principal Component analysis, rotated solution

| Motivational Variables | Component | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Motivation 1: Sharing views/opinions to make a difference or influence change | | 0.574 | | 0.448 | | | | |
| Motivation 2: To represent others like me/population subgroups/minorities | | | | -0.900 | | | | |
| Motivation 3: Sharing views/opinions in a non-judgmental platform | | | | | | | -0.691 | |
| Motivation 4: Self-actualization, allows my voice to be heard | | -0.859 | | | | | | |
| Motivation 5: My opinions are valued, appreciated, taken into account | | | 0.650 | | | | | |
| Motivation 6: Like being part of something, part of a team | | | | | | | 0.442 | |
| Motivation 7: Thought-provoking to participate in the panel | 0.685 | | | | | | | |
| Motivation 8: To be informed about topical issues | 0.464 | | | | | | | |
| Motivation 9: Receiving incentives | | | 0.549 | | | | | |
| Motivation 10: Donating to charity | | | | | 0.607 | | | |
| Motivation 11: Enjoying surveys/participating in research | 0.598 | | | | | | | |
| Motivation 12: Interesting topics | | | | | | 0.701 | | |
| Motivation 13: Social Research Centre/Life in Australia™ are reliable enterprises | | | | | | | 0.456 | -0.404 |
| Motivation 14: Contributing to the survey/study/research/science | | | | | 0.764 | | | |
| Motivation 15: Positive sentiments towards Australia | | | | | | | | 0.858 |
| Motivation 16: Have the time/something to do | | | | | | 0.449 | | |

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization
Loadings >=0.4 (absolute value) are displayed

- Five components were identified including 2+ loadings with positive values greater than 0.400. Component 1 captures *being motivated by an intellectual attraction* to what the panel experience provides to the panellist. Component 3 refers to a motivation of *being "gifted/valued" for one's participation*. Component 5 is about *being motivated to "give/contribute" something* as a panel member, in an altruistic sense. Component 6 seems related to *doing something interesting with one's time*. Component 7 is about *enjoying being part of this particular panel managed by a respected organization*.

- Three other components were identified, including one loading with a positive value greater than 0.400 and one loading with a negative value below -0.400 – Components 2, 4, and 8. For motivations as part of these components, we can argue that they are mutually exclusive. In practice, it means, for example, that reporting *sharing views/opinions to make a difference or influence change* resulted in a decreased propensity for panellists to also report *represent others like me/population subgroups/minorities* or *self-actualization* (Components 2 and 4 ).

**DiSC assessment**

We would like to thank you for participating in our study.

We will present you with 28 groups of four statements for non-judgmental personality and behavioral assessment. Please answer honestly and spontaneously.

Sometimes it may be difficult to decide which description to select. Remember there are no right or wrong answers, so just make the best decision you can.

Please read each of the four statements below. Then select the one that describes you the most and the one that describes you the least.

| | Describes me **most** | | Describes me **least** |
|---|---|---|---|
| Q1 | o | People look up to me | o |
| | o | I tend to be a kind person | o |
| | o | I accept life as it comes | o |
| | o | People say I have a strong personality | o |
| | Describes me **most** | | Describes me **least** |
| Q2 | o | I find it difficult to relax | o |
| | o | I have a very wide circle of friends | o |
| | o | I am always ready to help others | o |
| | o | I like to behave correctly | o |
| | Describes me **most** | | Describes me **least** |
| Q3 | o | I tend to do what I am told | o |
| | o | I like things to be very neat and tidy | o |
| | o | People can't put me down | o |
| | o | I enjoy having fun | o |
| | Describes me **most** | | Describes me **least** |
| Q4 | o | I respect my elders and those in authority | o |
| | o | I am always willing to do new things – to take a risk | o |
| | o | I believe things will go well | o |
| | o | I am always willing to help | o |
| | Describes me **most** | | Describes me **least** |
| Q5 | o | I am a neat and orderly person | o |
| | o | I am very active, both at work and play | o |
| | o | I am a very calm and placid person | o |
| | o | I generally get my own way | o |
| | Describes me **most** | | Describes me **least** |
| Q6 | o | I am very contented with life | o |
| | o | I tend to trust people | o |
| | o | I like peace and quiet | o |
| | o | I have a very positive attitude | o |
| | Describes me **most** | | Describes me **least** |
| Q7 | o | I have a great deal of will power | o |
| | o | I always take notice of what other people say | o |
| | o | I try to be obliging | o |

| | | | |
|---|---|---|---|
| | o | I am always cheerful | o |
| | Describes me **most** | | Describes me **least** |
| Q8 | o | I am self-confident | o |
| | o | People say I am a sympathetic type | o |
| | o | I have a tolerant attitude towards life | o |
| | o | I am an assertive person | o |
| | Describes me **most** | | Describes me **least** |
| Q9 | o | I never lose my temper | o |
| | o | I like things to be precise and correct | o |
| | o | I am very sure of myself | o |
| | o | I enjoy having a laugh and a joke | o |
| | Describes me **most** | | Describes me **least** |
| Q10 | o | My behavior is well disciplined | o |
| | o | People see me as being helpful | o |
| | o | I am always on the move | o |
| | o | I persevere until I get what I want | o |
| | Describes me **most** | | Describes me **least** |
| Q11 | o | I enjoy competition | o |
| | o | I do not treat life too seriously | o |
| | o | I always consider others | o |
| | o | I am an agreeable type | o |
| | Describes me **most** | | Describes me **least** |
| Q12 | o | I am very persuasive | o |
| | o | I see myself as a gentle person | o |
| | o | I am a very modest type | o |
| | o | I often come up with original ideas | o |
| | Describes me **most** | | Describes me **least** |
| Q13 | o | I am very helpful towards others | o |
| | o | I don't like tempting fate | o |
| | o | I don't give up easily | o |
| | o | People like my company | o |
| | Describes me **most** | | Describes me **least** |
| Q14 | o | I tend to be a cautious person | o |
| | o | I am a very determined person | o |
| | o | I am good at convincing people | o |
| | o | I tend to be a friendly person | o |
| | Describes me **most** | | Describes me **least** |
| Q15 | o | I don't scare easily | o |
| | o | People find my company stimulating | o |
| | o | I am always willing to follow orders | o |
| | o | I am a rather shy person | o |
| | Describes me **most** | | Describes me **least** |
| Q16 | o | I am very willing to change my opinion | o |
| | o | I like a good argument | o |
| | o | I tend to be an easy going type | o |
| | o | I always look on the bright side of life | o |

| | Describes me **most** | | Describes me **least** |
|---|---|---|---|
| Q17 | o | I am a very social sort of person | o |
| | o | I am very patient | o |
| | o | I am a very self-sufficient sort of person | o |
| | o | I rarely raise my voice | o |
| | Describes me **most** | | Describes me **least** |
| Q18 | o | I am always ready and willing | o |
| | o | I am always keen to try new things | o |
| | o | I don't like arguments | o |
| | o | People describe me as high spirited | o |
| | Describes me **most** | | Describes me **least** |
| Q19 | o | I enjoy taking a chance | o |
| | o | I tend to be very receptive to other people's ideas | o |
| | o | I am always polite and courteous | o |
| | o | I am a moderate rather than an extreme person | o |
| | Describes me **most** | | Describes me **least** |
| Q20 | o | I tend to be a forgiving type | o |
| | o | I am a sensitive person | o |
| | o | I have a lot of energy and vigour | o |
| | o | I can mix with anybody | o |
| | Describes me **most** | | Describes me **least** |
| Q21 | o | I enjoy chatting with people | o |
| | o | I control my emotions | o |
| | o | I am very conventional in my outlook | o |
| | o | I make decisions quickly | o |
| | Describes me **most** | | Describes me **least** |
| Q22 | o | I tend to keep my feelings to myself | o |
| | o | Accuracy is very important to me | o |
| | o | I like to speak my mind | o |
| | o | I am very friendly | o |
| | Describes me **most** | | Describes me **least** |
| Q23 | o | I like to handle things with diplomacy | o |
| | o | I am very daring | o |
| | o | Most people find me acceptable | o |
| | o | I feel satisfied with life | o |
| | Describes me **most** | | Describes me **least** |
| Q24 | o | I am obedient | o |
| | o | I am always willing to have a go | o |
| | o | Loyalty is one of my strengths | o |
| | o | I have a good deal of charm | o |
| | Describes me **most** | | Describes me **least** |
| Q25 | o | I tend to be an aggressive type | o |
| | o | I am good fun and have a lot of personality | o |
| | o | People tend to see me as an ‘easy touch’ | o |
| | o | I tend to be rather timid | o |
| | Describes me **most** | | Describes me **least** |

| | | | |
|---|---|---|---|
| Q26 | o | I am good at motivating people | o |
| | o | Patience is one of my major strengths | o |
| | o | I am careful to say the right thing | o |
| | o | I have a strong desire to win | o |
| | Describes me **most** | | Describes me **least** |
| Q27 | o | People find me easy to get on with | o |
| | o | I get a lot of satisfaction from helping others | o |
| | o | I always think things through | o |
| | o | I prefer to get things down now rather than later | o |
| | Describes me **most** | | Describes me **least** |
| Q28 | o | I am good at analysing situations | o |
| | o | I get restless quickly | o |
| | o | I think about how my decisions might affect others | o |
| | o | People see me as relaxed and easy going | o |

Table 4.11: Descriptive statistics for DiSC assessment traits scores (qualitative sample)

| | **Dominance** | **Influence** | **Steadiness** | **Compliance** |
|---|---|---|---|---|
| Number of participants | 14 | 14 | 14 | 14 |
| Mean | -6.4 | 0.1 | 5.1 | 1.2 |
| Median | -6.0 | -0.5 | 5.0 | 0.5 |
| Std. Deviation | 8.3 | 5.5 | 5.7 | 6.8 |
| Minimum | -20 | -8 | -4 | -10 |
| Maximum | 6 | 11 | 15 | 11 |

# Chapter 5 The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design

## 5.1 Introduction

Panels as online survey methods are now routinely used for collecting data and have been increasing in number. The online panel survey mode has introduced new sources of survey errors, even compared to traditional longitudinal surveys/non-online panels, such as birth cohort studies or life-cycle studies. There are at least two important elements related to these survey errors, which are specific to longitudinal surveys and panels collecting cross-sectional survey data: panel conditioning and attrition. In online panel studies, attrition is predominantly considered as permanent nonresponse from a particular data collection wave onwards (Kocar 2020). Besides attrition, unit nonresponse/survey non-completion is another potential source of representation bias (Groves et al. 2009), although clearly not specific to online panel surveys. While response rates alone are not a reliable indication of error, and it has been reported that the association between response rate and bias is weak at best (Groves & Peytcheva 2008), a high unit nonresponse typically signals a higher likelihood of nonresponse bias (Baker et al. 2010). Further, the respondents who opt-out should at some point be replaced on the panel with new respondents to preserve adequate sample size – particularly for certain population sub-groups – which increases the costs of panel management and data collection (Kruse et al. 2009). Online panels should be considered a form of hybrid between traditional longitudinal studies and web surveys since they predominantly use the online survey mode for collecting data from panel members, but track individuals over time (even if longitudinal outcomes aren't always the focus of the data collection). That often includes collection of paradata specific to online panels and storing the entire history of each member's panel behavior. Since this class of paradata have been a less explored topic (Callegaro 2013), and psychological theory explains that past behavior predicts future behavior fairly well (e.g., see Ouellette & Wood 1998), in this study we firstly review differential nonresponse and attrition, and then investigate the predictive power of online panel paradata to mitigate the problem of nonparticipation in probability-based online panel research. This could ultimately lead to a reduction of nonresponse error as defined in Total Survey Error framework (Groves et al. 2009).

## 5.2 Literature review

### 5.2.1 Attrition and unit nonresponse in online panels

We can distinguish between two types of attrition in panel studies: forced and normal. While forced attrition is managed by the data collector and occurs systematically at the end of eligibility, normal attrition is not managed and is a form of nonresponse; it occurs when panel members do not reach the end of their eligibility and leave the panel earlier for a variety of reasons, such as opting out, not participating fully, or falsifying interviews (Baker et al. 2010). In this study, we will use a classification by Callegaro and DiSogra (2008), who introduced slightly different online panel attrition outcomes: voluntary attrition, involuntary attrition, and mortality, with the focus on voluntary/opting-out attrition. It has been previously reported that voluntary attrition not only decreases the online panel sample size; selective attrition may introduce additional biases on top of that due to recruitment (Lugtig 2014).

Another source of representation bias in online panel surveys is unit nonresponse or survey non-completion, such as non-contact, refusal, or break-off (for definitions of survey outcomes see The American Association for Public Opinion Research 2016), especially with respect to the demographic or attitudinal characteristics of panel members. Both unit nonresponse and attrition may be considered sources of non-random survey errors (Cheng et al. 2016) in case of differential nonparticipation (nonresponse and voluntary attrition). Once both sources of nonparticipation are combined, the representation bias tend to increase, and nonignorable nonresponse can be the reason why even refreshment samples cannot fully correct for attrition bias (Schifeling et al. 2015).

In web surveys, response rates are significantly influenced by numerous factors, such as the questionnaire topic, length, sequencing, formatting, sampling method, whether participation is by invitation or not, pre-notification, and reminders (Fan & Yan 2010), as well as by demographic characteristics such as age, race/ethnicity, and education (Couper et al. 2007). In longitudinal studies, there are several predictors of attrition and response and some of these are specific to the panel format. Watson and Wooden (2009) concluded that it could not be assumed that experience with nonresponse in cross-sectional surveys is always relevant for predicting response and attrition in longitudinal surveys. They reported that there was a large random component to survey nonresponse in longitudinal studies in Australia, yet there are observable characteristics in the interview process and the respondents that are predictive of nonresponse. For example, respondents' perception of the survey in the preceding longitudinal study wave might influence cooperation in future waves (Watson & Wooden 2009). Moreover, Kruse et al. (2009) reported different levels of attrition in an opt-in panel in subsamples with different demographic and attitudinal characteristics. Using incentives, especially

from the second wave onwards, should increase response rates in certain panels (Castiglioni et al. 2008); however, Frankel and Hillygus (2014) argued that an individual's initial motivation to participate in a study might also be related to attrition probability, with those motivated strictly by monetary incentives having a higher probability of attrition. In probability-based online panels, attrition might also be predicted by including self-reported measures of personality, such as conscientiousness and openness to experience (Cheng et al. 2016), and panellists can be classified according to response type and attrition group, such as "stayers", "late-comers", "fast attritors", and "lurkers" to help understand their future participation. While stayers participate in almost all waves, lurkers are infrequent respondents, attritors opt-out of the panel at some point, and fast attritors leave even earlier (Lugtig 2014).

### 5.2.2 Paradata and their use in online panels

Paradata in surveys can be defined as additional data captured during the process of generating survey statistics and can be collected at different stages with different levels of detail (Kreuter 2013). Hence, different classifications, types and possible applications of paradata exist. In web surveys, paradata may be categorized into (1) device-type paradata (e.g., device, browser, and operating system (OS) used), and (2) questionnaire navigation paradata (e.g., mouse clicks, order of answering, last question answered before breaking off, and time spent per question). In addition to those for cross-sectional web surveys, there is a separate class of paradata – online panel paradata, which includes survey invitations received, surveys completed, attrition, and survey topics (Callegaro 2013). Web survey paradata can be collected in different phases: prior survey phase, recruitment phase, access phase, and response phase (McClain et al. 2019), and can be used for examining total survey errors (McClain et al. 2019; Olson & Parkhurst 2013), nonresponse (Lynn 2017), panel attrition (Lugtig & Blom 2018; Roßmann & Gummer 2016) or calculating propensity score weights adjusting for attrition (Roßmann & Gummer 2016). Lugtig and Blom (2018) concluded that paradata-identified behavior largely predicts nonresponse and Roßmann and Gummer (2016) reported an improvement in the fit of the nonresponse model after adding respondents' participation history, while both studies used a limited number of variables from paradata specific to online panels (e.g., participation in the previous wave). However, as Callegaro (2013) concluded, paradata for online panels are still a little explored topic, especially in a longitudinal design which takes advantage of the ability to derive longitudinal types of predictors. Also, longitudinal/panel data analysis methods controlling for unobserved heterogeneity can be used.

### 5.2.3 Statistical methods to study panel participation with panel paradata

To study panel participation, "static" statistical methods, such as survival analysis (Kruse et al. 2009), logistic regression (Castiglioni et al. 2008; Roßmann & Gummer 2016), multiple linear regression (Cheng et al. 2016), classification and regression trees (Lugtig & Blom 2018), and other tree-based machine learning methods such as boosting methods (Kern et al. 2019) have generally been used in previous studies. On the other hand, there are several advantages of analyzing paradata in a panel form using dynamic panel data modeling techniques. Analyzing panel data offers more accurate inference of panel parameters, greater capacity to capture complex behavior (including controlling the impact of omitted variables and generating more accurate predictions), and simplifying computation and statistical inference while involving at least two dimensions: a cross-sectional one and a time-series one (Hsiao 2007, pp. 3-6). In case of binary outcome variables (such as survey response in a wave, 1=yes, 0=no), binary logistic panel data analysis should be used instead of more traditional linear panel data models. Dynamic logit (or probit) models were previously adopted to allow for the use of binary panel data to disentangle true state dependence from the propensity to experience outcomes in all periods. For subject *i* at occasion *t*, the basic assumption *(i=1, …, n, t=1, …, T)* is presented in Equation 5.1:

$$\log \frac{p(y_{it}=1|\alpha_i, x_{it})}{p(y_{it}=0|\alpha_i, x_{it})} = \alpha_i + x'_{it}\beta + y_{i,t-1}\gamma \qquad (5.1)$$

where n is the sample size, T is the total number of occasions, $y_{i,t}$ is the binary response variable, x is a vector of exogenous covariates, $\alpha_i$ are individual-specific parameters for the unobserved heterogeneity and β and γ are structural parameters (Bartolucci & Nigro 2010). The challenge of any panel data analysis, in order to obtain valid inference on structural parameters, is to control the impact of unobserved heterogeneity, which effects can either be assumed as random variables (random effects model), as fixed parameters (fixed effects model), or both (mixed effects model) (Hsiao 2007, p. 8). An alternative to that is using pooled data analysis which is fundamentally applying classical regression (e.g., linear or logit) to pooled data. While this type of regression obtains minimum variance estimates of covariates under certain conditions, fixed-effect and random-effect models would often minimize variance better while accommodating a greater variety of covariates and sample sizes (Ward & Leigh 1993).

### 5.2.4 Outline of the study

This study investigates the differential nonparticipation in probability-based online panels and the power of online panel paradata predictors of nonparticipation rates, i.e., nonresponse and voluntary attrition rates. In contrast to similar research in the field, longitudinal panel participation data, i.e.,

survey outcome statuses, are explored in detail. The longitudinal nature of paradata enable the derivation of a number variables measuring panel response behavior over time. Panel data analysis will be carried out, which include the dimension of time in the models so as to improve the accuracy of the predictions. This study aims to answer the following research questions (RQs):

*RQ1: What is the extent of differential nonresponse and differential voluntary attrition in probability-based online panel surveys?*

The theory on nonresponse in longitudinal and online panel studies suggests that there are a number of socio-demographic characteristics associated with nonparticipation in surveys (e.g., Kruse et al. 2009; Watson & Wooden 2009). By answering this question, we will also determine if the available socio-demographic predictors should be included in dynamic logistic regression models to improve the accuracy of prediction with online panel paradata (this will also contribute to answering the RQ2).

*RQ2: What is the predictive power of online panel paradata with or without socio-demographics?*

Assuming we will be able to identify some level of differential nonresponse in online panels, we will compare the predictive power of online panel paradata with and without socio-demographics. This approach is similar to behavioral research in psychology where personality traits and past behavior as predictors of future behavior are being compared (e.g., Harris et al. 2016).

*RQ3: To what extent do dynamic logistic regression models, i.e., random- and fixed-effects models, improve the accuracy of prediction in comparison to pooled "static" regression models, if at all?*

Moreover, we will show if using the advantages of the panel structure of the data, which is to create dynamic regression models, can increase the accuracy of prediction of nonresponse in probability-based panels in contrast to pooled estimation with panel data reported in the literature (e.g., Castiglioni et al. 2008; Cheng et al. 2016; Kruse et al. 2009; Roßmann & Gummer 2016). We will start off by comparing regression coefficients between pooled, random- and fixed-effects models with the same predictors.

*RQ4: How many waves of online panel participation data are needed to predict nonresponse and voluntary attrition with desirable accuracy?*

Since our time-series is much longer in comparison to studies carried out by Lugtig and Blom (2018) and Roßmann and Gummer (2016), we will provide insight into how much data are required for fairly accurate prediction.

*RQ5: How do we determine the right balance between "costs" and "benefits" when identifying nonrespondents for further treatment?*

Identifying potential nonrespondents itself would have little value for an online panel organization without following with some kind of a treatment to increase response and decrease attrition (e.g., Lugtig 2014). We will show how identification as the first step in improving participation becomes inefficient and cost-ineffective at some level, and discuss practical solutions to that.

## 5.3 Methods

### 5.3.1 Data

The dataset used in this research was all members of the Life in Australia™, the only mixed-mode probability-based online panel in Australia. The Life in Australia™ dataset used in this study did not consist of substantive survey data, but of panel response, attrition, incentives, and other characteristics of the panel members for waves 1-30 (data collection period: December 2016 and August 2019). There was a substantial panel refresh after this time period, plus the combination of the 2019/20 Black Summer bushfire season and the COVID-19 pandemic which, combined, introduced the strong potential for a structural break in the dataset and is therefore a period best suited to a focused research program.

It was possible to use the dataset to study survey participation, including nonresponse and panel attrition, and included information for 3,322 panel members whose demographic information had been collected at the end of 2016 (Kaczmirek et al. 2019). The relatively small top-up sample from May 2018 is not included in this study. For each of the 30 waves of subsequent data collection, the dataset included all relevant information about the activity of panel members. If a panel member became inactive (excluding vacations or public holidays) due to voluntary (panel opt-out) or involuntary (retired) attrition, or due to mortality (death), participation data were no longer collected for that respondent from the successive wave as attrited units are no longer relevant for analysis (cannot rejoin and re-attrite, hence no variability in response). These missing data make the panel an unbalanced panel in panel data analysis.

### 5.3.2 Population and sampling

The population in this research was defined as “Australian residents aged 18 years or older”. The recruitment rate for the establishment of the Life in Australia™ panel was 21.1% and the profile rate was 77.7%. For the recruitment process, a dual-frame random digit dialing (RDD) sample design was employed, with a 40:60 (pilot) and 30:70 (the main recruitment effort) split between landline and cell phone sample frames. The offline population, so-called offliners, completed surveys by telephone (Kaczmirek et al. 2019). All members of the sample were invited to participate in the majority of surveys between December 2016 and August 2019, except for waves 5(a), 8, 13, and 20. All variable

values for all units were, nevertheless, included in the analysis, since the increased time gap between survey invitations could well prove to be one of the predictors of survey participation.

### 5.3.3 Data analysis, statistical modeling and derived covariates

To analyze the data and to answer the research questions, multivariate statistical analysis was used, including multiple linear regression, and panel data analysis, including logit random- and fixed-effect models. All of these models were created to study attrition and nonresponse using paradata and not for substantive analysis using substantive survey items. Since most of the dependent variables in the models were binary and consisted of a set of independent variables including controls, binary logistic regression was used in the majority of the models. In addition to multiple linear regression analysis and logistic regression (aggregated participation variables), this study used logistic regression analysis for the binary panel data (dynamic logit model) in the main models, since the added value would be consideration of the longitudinal dimensions of panel participation, as measured by the online panel paradata. The selected longitudinal or panel data consisted of repeated observations of the same units at different points in time, enabling control for unobserved heterogeneity. We also considered using probit models, but we did not observe non-constant error variances, and the results would have been fairly similar.

To answer the research questions, this article will present four separate but to some extent related multivariate models: (1) multiple linear regression to investigate the socio-demographic aspects of panel response rates (the dependent variable: *participation percentage*); (2) binary logistic regression to investigate socio-demographic features of panel attrition (the dependent binary variable: *attritor*); (3) dynamic logit regression to investigate panel response using past response patterns and data collection characteristics (the dependent variable: *nonresponse in a particular wave*); and (4) dynamic logit regression to investigate panel attrition using past response patterns and data collection characteristics (the dependent variable: *voluntary attrition in a particular wave*)[43].

The derived variables as exogenous covariates/predictors of panel participation in fixed- and random-effect models were based on the AAPOR categorisation of the survey outcome rates (see The American Association for Public Opinion Research 2016): participation rate prior to wave, non-contact rate prior to wave, refusal rate prior to wave, non-refusal rate prior to wave, donation-to-charity rate[44] prior to wave, consecutive waves without participation, consecutive waves with nonresponse, consecutive waves with non-contact, consecutive waves with refusal, consecutive waves with non-

[43] To not exclude any cases, we later imputed missing demographic information using multiple imputations for models 3 and 4.

[44] Charity rate is a special type of rate and is not one of standard survey outcome rates. Yet, it is associated with motivation to participate in online panel surveys and could be treated as a type of panel behavior measured with online panel paradata.

refusal, consecutive participating waves of donating to charity, and changes in the survey outcomes in the preceding two waves. For each type of survey outcome, the rates prior to a wave of data collection were calculated for each respondent in the panel. For example, the response rate at 5b was the total survey completion rates in waves 1, 2, 3, 4, and 5a for that respondent. The difference between the response and participation rates was the denominator – the response rate was calculated as the number of interviews divided by the number of invitations while the participation rate was calculated as the number of interviews divided by the number of all waves (invited or not)[45]. The charity rate was calculated as the number of donations to charity divided by the number of survey completions. Moreover, as with the panel survey outcome rates, consecutive occurrences of a particular survey outcome prior to a wave of data collection were calculated for each respondent in the panel.[46] Changes between survey outcomes were possible to calculate from wave 3 on, since two consecutive waves of data were required to identify changes in respondents' participation behavior prior to a wave. Changes from interview to other outcomes (including non-contact, non-refusal, physical or mental inability/incompetence, but excluding refusal) in consecutive waves, were the less severe changes of survey response outcomes, while any other survey outcome (including interview) to refusal should be considered as more severe change.

## 5.4 Results

### 5.4.1 Descriptive statistics

The results of the descriptive analysis for all independent variables in the first two models (as well as the covariates for random effects) and the dependent variables of *survey response percentage* and *attritor* will be presented. This is the introductory analysis into more in depth analysis in the following results subsection to answer research questions 1–5 (Subsections 5.2–5.6).

The differences in the panel participation rates in the first 30 waves of the Life in Australia™ data collection between the socio-demographic groups of analytic importance can be seen in Table 5.1. Of all Life in Australia™ panellists recruited in 2016 (n=3,322), only those who were once active, i.e., responded in at least one wave out of 30, were included (n=2,990). The groups with the lowest average response rates (RR)/survey completion rates were the youngest (18–24 years of age at recruitment – RR 58.44%, and 25–34 years of age – RR 67.75%), respondents who spoke a language other than

[45] *Participation rate (prior to wave$_n$)* = 'number of completed questionnaire by wave$_n$' / 'total number of waves by wave$_n$'; for example, if a panellist completed questionnaires in waves 1-3, was not invited in wave 4 and refused to participate in wave 5, the *participation rate* before wave 6 was 0.6 or 60% (3 waves out of 5).

[46] For example: *Consecutive refusal (prior to wave$_n$)* = consecutive waves prior to wave$_n$ with refusal survey outcome (waves invited to only); if a panellist completed questionnaire in wave 1, refused to participate in wave 2, was not invited in wave 3, and again refused to participate in waves 4 and 5, the *consecutive refusal* before wave 6 was 3 (three consecutive waves invited to, i.e., 2, 4 and 5).

English at home (65.69%) and those who were Indigenous (69.00%). The groups with the highest overall response rates were 65 years of age or above (65–74 years – 84.83%, 75+ years – 82.20%), better educated (with a Bachelor or higher – 78.86%), and carers (81.11%). Response rates tended to increase with age and were higher in the Australian-born respondent group and amongst females.

On the other hand, the attrition statistics were quite different. The groups most likely to opt-out of the panel were the least educated (up to Year 11 or the lower 20.31%), the offline population (22.40%), and the elderly (75+ years of age – 24.05%), and voluntary attrition generally tended to increase with age. Those least likely to attrite were the youngest (18–24 years of age – 6.69%, 25–34 years of age – 8.93%), those who were Indigenous (7.81%), and those who speak language other than English at home (10.86%).

Table 5.1: Survey response percentage and attritor sample statistics (n=2,990)

| | | Survey response % | | Voluntary attritor (in any wave, in %) | |
|---|---|---|---|---|---|
| | n | Mean | SD | No | Yes |
| **Gender** | | | | | |
| Female | 1,576 | 76.60 | 29.52 | 86.42 | 13.58 |
| Male | 1,403 | 74.55 | 31.12 | 86.60 | 13.40 |
| **Education** | | | | | |
| Bachelor or higher | 1,127 | 78.86 | 28.91 | 88.11 | 11.89 |
| Certificate/diploma/trade | 1,062 | 73.59 | 30.83 | 87.01 | 12.99 |
| Year 12 or equivalent | 343 | 72.65 | 31.41 | 88.63 | 11.37 |
| Year 11 or less | 458 | 74.33 | 30.86 | 79.69 | 20.31 |
| **Capital city in state** | | | | | |
| No | 999 | 76.93 | 29.34 | 86.79 | 13.21 |
| Yes | 1,966 | 75.59 | 30.21 | 86.52 | 13.48 |
| **Born in Australia** | | | | | |
| No | 820 | 72.75 | 31.81 | 86.10 | 13.90 |
| Yes | 2,160 | 76.72 | 29.62 | 86.71 | 13.29 |
| **Only English spoken at home** | | | | | |
| No | 442 | 65.69 | 35.11 | 89.14 | 10.86 |
| Yes | 2,547 | 77.32 | 29.03 | 86.06 | 13.94 |
| **Indigenous status** | | | | | |
| No | 2,921 | 75.74 | 30.18 | 86.41 | 13.59 |
| Yes | 64 | 69.00 | 34.61 | 92.19 | 7.81 |
| **Other healthcare card** | | | | | |
| No | 1,965 | 74.72 | 30.61 | 87.33 | 12.67 |
| Yes | 992 | 78.11 | 29.05 | 85.08 | 14.92 |
| **Carer status** | | | | | |
| No | 2,400 | 74.37 | 30.97 | 85.58 | 14.42 |
| Yes | 582 | 81.11 | 26.32 | 90.38 | 9.62 |
| **Population**[47] | | | | | |
| Offline | 433 | 72.53 | 28.45 | 77.60 | 22.40 |
| Online | 2,557 | 76.10 | 30.56 | 87.99 | 12.01 |
| **Age group** | | | | | |
| 18-24 years | 239 | 58.44 | 35.33 | 93.31 | 6.69 |
| 25-34 years | 403 | 67.75 | 33.61 | 91.07 | 8.93 |
| 35-44 years | 418 | 71.10 | 32.76 | 89.71 | 10.29 |
| 45-54 years | 518 | 75.56 | 29.42 | 87.07 | 12.93 |
| 55-64 years | 636 | 79.62 | 28.53 | 85.53 | 14.47 |
| 65-74 years | 532 | 84.83 | 23.35 | 82.71 | 17.29 |
| 75 or more years | 237 | 82.20 | 22.28 | 75.95 | 24.05 |
| **Socio-Economic Indexes for Areas** | | | | | |
| Quartile 1 | 417 | 76.69 | 30.08 | 88.73 | 11.27 |
| Quartile 2 | 520 | 76.76 | 29.88 | 85.00 | 15.00 |
| Quartile 3 | 570 | 76.55 | 29.58 | 88.95 | 11.05 |
| Quartile 4 | 635 | 75.28 | 30.67 | 84.88 | 15.12 |
| Quartile 5 | 822 | 75.50 | 29.59 | 86.25 | 13.75 |
| **Age, mean with SD** | 2,949 | | | 50.31 (17.18) | 56.89 (16.62) |
| **Survey response %, mean with SD** | 2,990 | | | 78.65 (29.61) | 55.93 (27.07) |

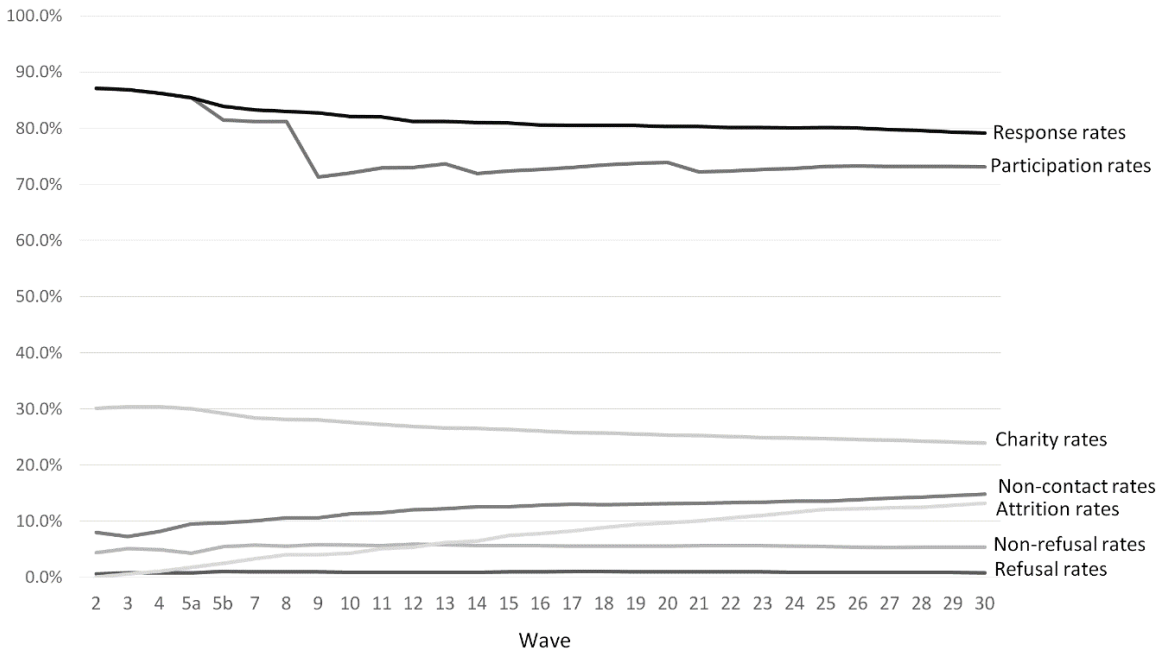
The negative association between the survey response rate and attrition (attritors: 55.96% average completion rate, non-attritors: 78.65% average completion rates, statistically significant at $p<0.01$) can be explained by the fact that attritors responded with a lower propensity than non-attritors even

[47] At profile survey (before first wave for panellist).

before opting out of the sample. If we have a closer look at the age variable, we can observe an interesting phenomenon – survey response (i.e., completion rates not taking retention rate into account) increased with age while voluntary attrition rates also increase with older panellists. It seems that higher attrition age groups (e.g., 65–74 years) were left with people who were more likely to respond whilst lower attrition age groups (e.g., 18–24 years) still contained respondents who were less motivated but who would not attrite, also known as infrequent respondents (lurkers).

Secondly, this article will present the averages for the differently derived panel participation rates as independent variables in the models predicting survey nonresponse and panel attrition. The differences in panel response behavior over time, calculated at the individual level and presented as aggregated statistics, can be seen in Figure 5.1. Besides survey completion rates, we are adding attrition rate which represents the cumulative percentage of respondents who actually opted out. It can be seen that individual response rates slowly decreased over time – the average response rate prior to a wave decreased from almost 90% to just above 80% and the curve flattens at wave 15. Participation rates decreased at a higher rate since not all respondents were invited in waves 5a, 8, 13 and 20; it flattened in wave 9 at about 72% for all remaining active respondents. Non-refusal and refusal rates remained fairly constant while non-contact and cumulative attrition rates increased over time, and charity rates gradually decreased after wave 5a. It was concluded that the principal reasons for survey nonresponse were non-contact and panel attrition but not refusals.

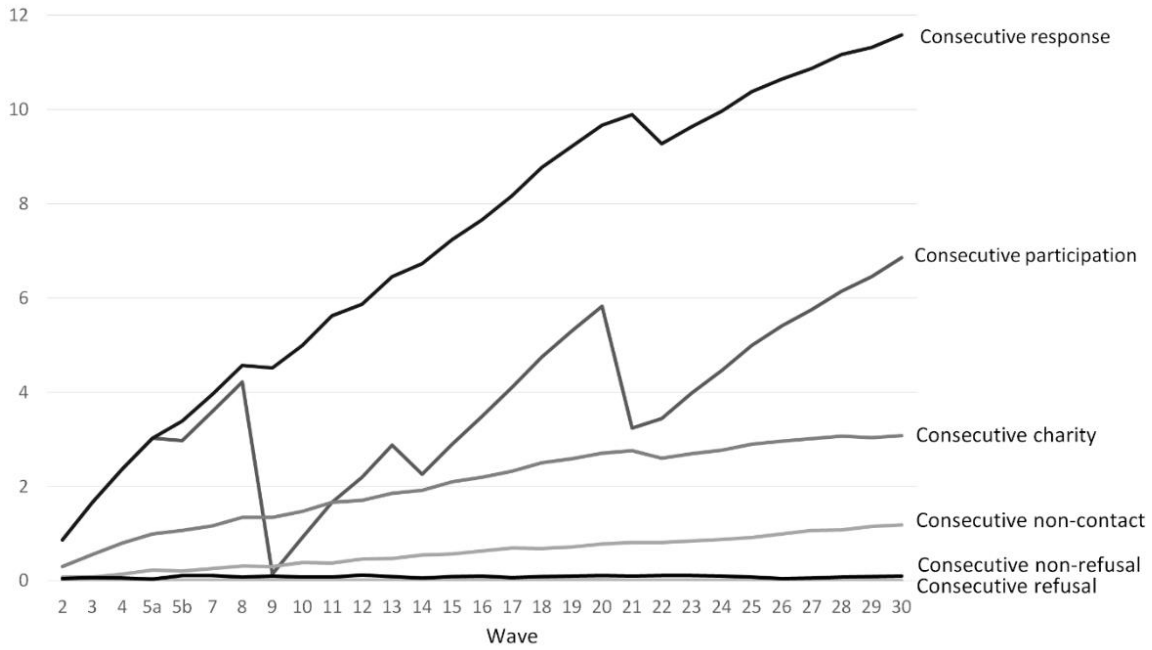
Figure 5.1: Different average panel survey outcome rates prior to waves 2–30 (n=2,990)




The differences in consecutive survey outcomes over time, calculated at the individual level and presented as aggregated statistics, can be seen in Figure 5.2. It is evident that consecutive response

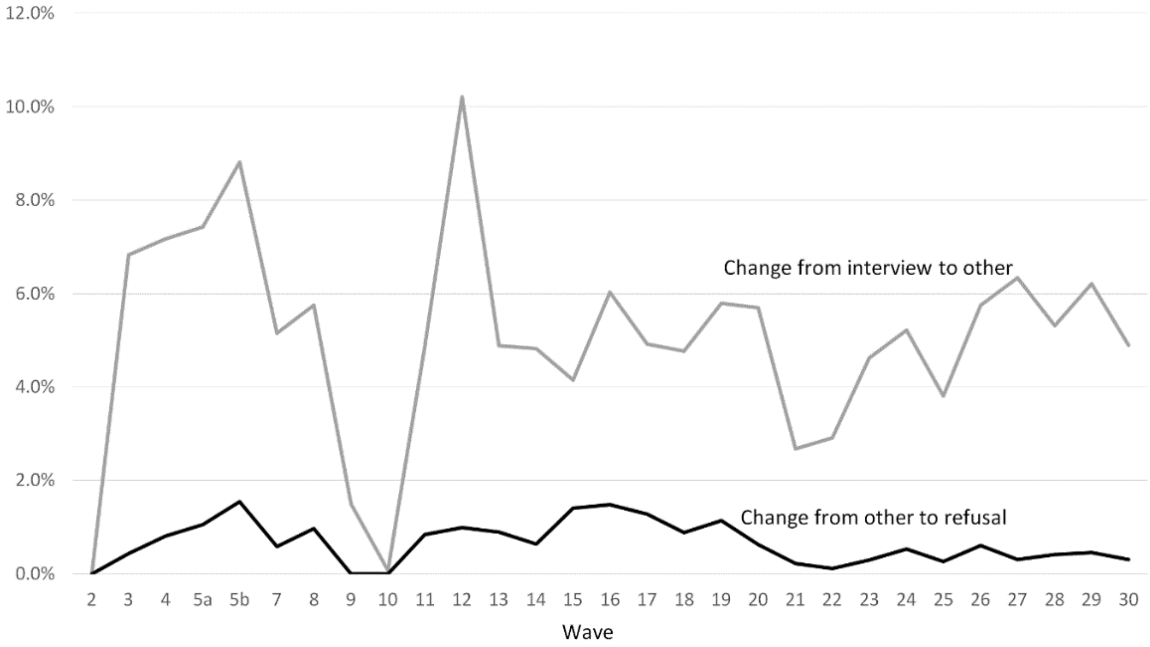
rates increased over time, which can also be attributed to the fact that almost one-third of the Life in Australia™ sample recruited in 2016 responded to every invitation. Consecutive participation followed the same distributions as consecutive response only until wave 5a. After that, not all respondents were invited in each wave and only a very small subsample was asked to participate in wave 8; therefore, the consecutive participation rate had decreased to close to zero prior to wave 9. Consecutive charity and non-contact rates increased slowly in a linear fashion over time. Consecutive non-refusal and refusal rates were much lower compared to those in other consecutive survey outcomes and did not alter significantly (at the aggregated level) over time.

Figure 5.2: Different average consecutive survey outcomes prior to waves 2–30 (n=2,990)




The differences in survey outcomes over time, identified at the individual level and presented as aggregated statistics, can be seen in Figure 5.3. Changes from interview to other outcomes were much more common than changes from any outcome to refusal, as the more severe changes of panel participation behavior. Prior to waves 9 and 10, the rates decreased to almost zero since only a small subsample was invited to participate in wave 8. The highest interview-to-other change rates were observed prior to wave 12 (i.e., in wave 10 survey outcome=interview, and wave 11 survey outcome=other) with one out of ten active respondents changing their survey outcome. Conversely, the highest other-to-refusal change rates were observed prior to waves 5b and 16 with approximately 1.5% of respondents changing their survey outcome to refusal.

Figure 5.3: Changes in survey outcomes prior to waves 2–30 (n=2,990)




### 5.4.2 Socio-demographic predictors of panel nonresponse and attrition (RQ1)

The first multiple linear regression model demonstrated the effects of the characteristics of the online panel respondents (as the independent variables) on the nonresponse rate (as the dependent variable). The results from Table 5.2 helped answer the first research question regarding differential nonresponse and the socio-demographic predictors of panel participation (RQ1).

The effects of different independent variables on the continuous variable in the model *survey response rate* can be seen in Table 5.2. The results of the regression analysis showed that the overall individual response rate for all waves was positively associated with higher education (the coefficients for certificate/diploma/trade and Year 12 or equivalent, Year 11 and lower were all below -3, with $p<0.05$), only English spoken at home (with a 95% Confidence Interval (CI) 2.92 to 9.73, and $p<0.001$), and being older (age, a continuous variable, with a 95% CI 0.38 to 0.52, and $p<0.001$). The online population tended to produce a higher response rate than the offline respondents (with a 95% CI 5.69 to 12.23, and $p<0.001$) and the Socio-Economic Indexes for Areas (SEIFA) Quartile 5 group tended to respond less frequently (with a 95% CI -7.12 to -0.54), ceteris paribus.

The strongest socio-demographic predictors of survey response rates in the online panel surveys were age, online-offline status, level of education, and whether languages other than English were spoken at home. The adjusted R-Squared value equaled 0.085, meaning that the model explained 8.5% of the variability in the response data. While that indicates that differential nonresponse is present, it does not seem to be severe (RQ1).

Table 5.2: Multiple linear regression results, the effect of socio-demographic characteristics on overall survey completion rate in waves 1-30, 2872 persons

| | Coef | L 95% CI | U 95% CI | p value |
|---|---|---|---|---|
| **Gender** | | | | |
| Female | 0 | | | |
| Male | -1.70 | -3.81 | 0.40 | 0.113 |
| **Education** | | | | |
| Bachelor or higher | 0 | | | |
| Certificate/diploma/trade | -6.96 | -9.51 | -4.42 | <0.001** |
| Year 12 or equivalent | -3.86 | -7.46 | -0.26 | 0.036* |
| Year 11 or less | -9.29 | -12.72 | -5.85 | <0.001** |
| **Capital city in state** | | | | |
| No | 0 | | | |
| Yes | 1.43 | -1.08 | 3.95 | 0.263 |
| **Born in Australia** | | | | |
| No | 0 | | | |
| Yes | 1.95 | -0.64 | 4.54 | 0.141 |
| **Only English spoken at home** | | | | |
| No | 0 | | | |
| Yes | 6.32 | 2.92 | 9.73 | <0.001** |
| **Indigenous status** | | | | |
| No | 0 | | | |
| Yes | -3.38 | -10.65 | 3.89 | 0.362 |
| **Other healthcare card** | | | | |
| No | 0 | | | |
| Yes | -0.36 | -2.86 | 2.14 | 0.779 |
| **Carer status** | | | | |
| No | 0 | | | |
| Yes | 4.27 | 1.58 | 6.97 | 0.002** |
| **Population** | | | | |
| Offline | 0 | | | |
| Online | 8.96 | 5.69 | 12.23 | <0.001** |
| **SEIFA** | | | | |
| Quartile 1 | 0.12 | -3.56 | 3.81 | 0.947 |
| Quartile 2 | -1.15 | -4.61 | 2.30 | 0.513 |
| Quartile 3 | 0.00 | | | |
| Quartile 4 | -1.62 | -4.94 | 1.69 | 0.337 |
| Quartile 5 | -3.83 | -7.12 | -0.54 | 0.023* |
| **Age** | 0.45 | 0.38 | 0.52 | <0.001** |
| **Constant** | 43.99 | 37.54 | 50.44 | <0.001** |
| **Adjusted R-Squared** | 0.085 | | | |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level

The second regression model demonstrated the effects of the characteristics of the online panel respondents (as the independent variables) on attrition (as the dependent variable). The results

helped answer the first research question regarding the differential attrition and its socio-demographic predictors.

The effects of different independent variables on the binary dependent variable in the logit regression model *voluntary attritor* can be seen in Table 5.3. The same independent variables as in the first multiple regression model were used. The results showed that panel opt-out attrition (0=no, 1=yes) was positively associated with participants speaking language other than English at home (95% CI 0.10 to 0.92, and p=0.016), those not holding other healthcare cards (those holding other healthcare care with a 95% CI -0.56 to -0.01, and p=0.040), not being a carer (carers with a 95% CI -0.75 to -0.09, and p=0.012), being younger (age, continuous, with a 95% CI 0.03 to 0.05, and p<0.001), being an "offliner" (with a 95% CI -0.70 to -0.07, and p=0.015) and with a lower response rate (with a 95% CI -0.03 to -0.02, and p<0.001).

Table 5.3: Logistic (logit) regression results, the effect of socio-demographic characteristics on voluntary attrition in waves 1-30, 2872 persons

| | Coef | L 95% CI | U 95% CI | p value |
|---|---|---|---|---|
| **Gender** | | | | |
| Female | 0 | | | |
| Male | -0.09 | -0.33 | 0.14 | 0.431 |
| **Education** | | | | |
| Bachelor or higher | 0 | | | |
| Certificate/diploma/trade | -0.07 | -0.36 | 0.22 | 0.648 |
| Year 12 or equivalent | -0.20 | -0.63 | 0.23 | 0.355 |
| Year 11 or less | 0.07 | -0.29 | 0.43 | 0.697 |
| **Capital city in state** | | | | |
| No | 0 | | | |
| Yes | 0.10 | -0.18 | 0.38 | 0.488 |
| **Born in Australia** | | | | |
| No | 0 | | | |
| Yes | -0.05 | -0.34 | 0.23 | 0.710 |
| **Only English spoken at home** | | | | |
| No | 0 | | | |
| Yes | 0.51 | 0.10 | 0.92 | 0.016* |
| **Indigenous status** | | | | |
| No | 0 | | | |
| Yes | -0.50 | -1.47 | 0.48 | 0.317 |
| **Other healthcare card** | | | | |
| No | 0 | | | |
| Yes | -0.29 | -0.56 | -0.01 | 0.040* |
| **Carer status** | | | | |
| No | 0 | | | |
| Yes | -0.42 | -0.75 | -0.09 | 0.012* |
| **Population** | | | | |
| Offline | 0 | | | |
| Online | -0.39 | -0.70 | -0.07 | 0.015* |
| **SEIFA** | | | | |
| Quartile 1 | 0.05 | -0.39 | 0.48 | 0.836 |
| Quartile 2 | 0.24 | -0.15 | 0.64 | 0.224 |
| Quartile 3 | 0 | | | |
| Quartile 4 | 0.39 | 0.00 | 0.77 | 0.047* |
| Quartile 5 | 0.17 | -0.21 | 0.56 | 0.382 |
| **Age** | 0.04 | 0.03 | 0.05 | <0.001** |
| **Survey response %** | -0.03 | -0.03 | -0.02 | <0.001** |
| **Constant** | -2.08 | -2.82 | -1.34 | <0.001** |
| **Pseudo R-Squared** | 0.145 | | | |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level

### 5.4.3 Online panel paradata predictors of panel nonresponse: statistical modeling to address RQ2, RQ3 and RQ4

Panel nonresponse, which was predicted using the paradata and socio-demographic characteristics of the online panellists, was defined as any survey non-completion outcome. To investigate those aspects of respondent behavior affecting nonresponse and to later answer research question RQ2 (predictor choice) and RQ4 (required length of time-series), a number of longitudinal variables as independent variables/covariates in the dynamic panel regression models were derived (see Subsection 5.3.3 and Figures 5.1, 5.2, and 5.3). Response rate was excluded since it is highly correlated with the participation rate and was a linear combination of the other survey outcome rates.

Panel nonresponse without subsequent attrition was investigated with three different models: (1) the pooled logit regression model (non-longitudinal); (2) the random-effect dynamic logistic model; and (3) the fixed-effect dynamic logistic model, to later answer RQ3 (statistical modeling choice). In the first static logistic regression model, every observation (respondent in a wave) was independent, and allowed for individual effects. The results of this regression were compared to those of the dynamic logistic regression. The difference between the two dynamic models was that the results of the random-effect model could be generalized to the population from which the Life in Australia™ sample was drawn, while the fixed-effect within-person model controlled for time-invariant person-related characteristics and could be considered as slightly more accurate for this specific sample. The fixed-effect model did not include any controls, since the demographics were collected at recruitment only and the within-group variance equaled zero.

The results of the logit regression analysis can be seen in Table 5.4. The coefficients for the pooled and random-effect models were fairly similar: participation rates were negatively associated with nonresponse (the coefficients in the pooled model -2.73, and -3.34 in the random-effect model; both significant at the 0.01 level) while the charity rate was positively associated with the dependent variable (coefficient, pooled: 0.40, random-effect: 0.68; significant at the 0.01 level). Moreover, consecutive participation, non-contact, refusal and non-refusal all increased the probability of nonresponse in both models (pooled model coefficients between 0.03 [participation] and 1.10 [refusal] and for random-effect between 0.04 [participation] and 0.98 [refusal]; all significant at the 0.01 level), while consecutive response decreased the probability of nonresponse (pooled mode coefficient -0.12, random-effect model coefficient -0.09, both significant at the 0.01 level). In both models, the effect of a change from interview to other (coefficients, pooled regression: 0.09, $p=0.028$, random effect: 0.15, $p<0.001$) and a change from other to refusal (coefficients, pooled regression: 0.51, $p<0.001$, random-effect: 0.45, $p=0.008$) were significant.

The fixed-effect model produced slightly different results. While all the consecutive survey outcomes prior to a wave and changes in survey outcomes were similarly positively associated with nonresponse, non-contact rate (coefficient -3.84, $p<0.001$), refusal rate (coefficient -3.10, $p<0.001$), and non-refusal rate (coefficient -4.45, $p<0.001$), were negatively associated with nonresponse in contrast to pooled and random-effect regression. One of the reasons for these differences was that fixed-effect models require some variability in the response variable, thus respondents with 100% total survey completion rates, about one-third of the whole sample, were automatically excluded from the analysis.

The strongest online panel paradata predictors of nonresponse after controlling for socio-demographic characteristics were: (1) all consecutive waves with a particular survey outcome, (2) changes in survey outcome, and (3) survey outcome rates prior to a certain wave, such as participation, non-refusal rates, or charity rates. In general, variables derived from online panel paradata (predominantly survey completion rates) showed to be good predictors of nonresponse and should perform well in predicting nonrespondents, presented in the next paragraphs.

Table 5.4: Logit regression, random-effect and fixed-effect within-person logistic regression results, the effect of previous response trends on **nonresponse** in certain wave, 2,990 persons, waves 1-30

| | Logit regression model (a pooled model) | | | | Random-effect within-person logistic regression model | | | | Fixed-effect within-person logistic regression model | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| Participation rate | -2.73 | -3.40 | -2.06 | <0.001** | -3.34 | -4.13 | -2.55 | <0.001** | -4.98 | -5.85 | -4.10 | <0.001** |
| Non-contact rate | 0.49 | -0.17 | 1.15 | 0.145 | -0.18 | -0.99 | 0.62 | 0.657 | -3.84 | -4.75 | -2.92 | <0.001** |
| Refusal rate | 0.49 | -0.37 | 1.34 | 0.264 | -0.15 | -1.23 | 0.92 | 0.781 | -3.10 | -4.36 | -1.85 | <0.001** |
| Non-refusal rate | 1.05 | 0.37 | 1.73 | 0.003** | -0.11 | -0.95 | 0.74 | 0.807 | -4.45 | -5.41 | -3.50 | <0.001** |
| Charity rate | 0.40 | 0.30 | 0.49 | <0.001** | 0.68 | 0.53 | 0.83 | <0.001** | 0.50 | 0.19 | 0.80 | 0.002** |
| Consecutive participation | 0.03 | 0.01 | 0.05 | <0.001** | 0.04 | 0.02 | 0.05 | <0.001** | 0.00 | -0.02 | 0.02 | 0.962 |
| Consecutive response | -0.12 | -0.13 | -0.11 | <0.001** | -0.09 | -0.10 | -0.07 | <0.001** | 0.02 | 0.01 | 0.04 | 0.001** |
| Consecutive non-contact | 0.55 | 0.52 | 0.58 | <0.001** | 0.45 | 0.41 | 0.48 | <0.001** | 0.49 | 0.45 | 0.52 | <0.001** |
| Consecutive refusal | 1.10 | 0.87 | 1.32 | <0.001** | 0.98 | 0.74 | 1.23 | <0.001** | 1.02 | 0.77 | 1.27 | <0.001** |
| Consecutive non-refusal | 0.33 | 0.27 | 0.39 | <0.001** | 0.21 | 0.15 | 0.28 | <0.001** | 0.32 | 0.26 | 0.39 | <0.001** |
| Consecutive charity donations | 0.01 | 0.00 | 0.03 | 0.024* | 0.02 | 0.00 | 0.03 | 0.015* | 0.03 | 0.01 | 0.05 | <0.001** |
| Change from interview to other | 0.09 | 0.01 | 0.17 | 0.028* | 0.15 | 0.06 | 0.23 | 0.001** | 0.24 | 0.15 | 0.32 | <0.001** |
| Change from other to refusal | 0.51 | 0.20 | 0.82 | 0.001** | 0.45 | 0.12 | 0.78 | 0.008** | 0.43 | 0.09 | 0.76 | 0.012* |
| Constant | 1.27 | 0.65 | 1.90 | <0.001** | | | | | | | | |
| **Pseudo R-Squared** | 0.415 | | | | | | | | | | | |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level; pooled logit regression and random-effect models include the following controls: gender, age, education, capital, born in Australia, English only spoken at home, Indigenous status, another health card, carer status, online/offline population and SEIFA

### 5.4.4 Online panel paradata predictors of panel attrition: statistical modeling to address RQ2, RQ3 and RQ4

Panel attrition, explored using panel paradata and socio-demographic characteristics, was a binary variable in these models with "0" representing non-attrition and "1" representing panel "opt-out" attrition. As with nonresponse, attrition was investigated with pooled logit, random-effect logistic and fixed-effect within-person logistic regression models and the same derived predictors. The fixed-effect model did not include any controls, since the demographic information was collected at recruitment only and there was no within-group variance for those variables. For the same reason, not all respondents but only attritors were included in the fixed-effect regression analysis (n=404).

As with the nonresponse models in Table 5.4, the pooled and random-effect models predicting attrition (which occurred only once) were reasonably similar (see the coefficients in Table 5.5). As such, only the dynamic models will be discussed and compared. First of all, it became apparent that fewer dimensions of survey participation behavior in previous waves had had an effect on attrition in the random-effect model. However, all of the average survey outcome rates prior to a wave and consecutive participation (a coefficient of 0.35, $p<0.001$) had a statistically significant effect in the fixed-effect model only (coefficients of -19.87 [participation], -14.27 [non-contact], -15.65 [non-refusal], and -9.62 [refusal] with $p=<0.001$), while charity rates had an effect in the random-effect model only (coefficient 0.94, $p<0.001$). However, consecutive response (negative), consecutive refusal, consecutive charity, and change from other to refusal (all positive) had a significant effect on attrition in both models.

Table 5.5: Logit regression, random effect and fixed effect within-person logistic regression results, online and offline samples, the effect of previous response trends on **voluntary panel attrition** in certain wave, 2,990 persons, waves 1-30

| | Logit regression model (a pooled model) | | | | Random-effect within-person logistic regression model | | | | Fixed-effect within-person logistic regression model | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| Participation rate | -2.37 | -4.94 | 0.21 | 0.072 | -2.60 | -5.26 | 0.06 | 0.056 | -19.87 | -24.57 | -15.17 | <0.001** |
| Non-contact rate | -1.32 | -3.84 | 1.21 | 0.307 | -1.45 | -4.03 | 1.12 | 0.269 | -14.27 | -18.95 | -9.60 | <0.001** |
| Refusal rate | 1.14 | -1.39 | 3.67 | 0.377 | 1.17 | -1.40 | 3.75 | 0.371 | -9.62 | -15.30 | -3.93 | 0.001** |
| Non-refusal rate | -0.32 | -2.88 | 2.23 | 0.805 | -0.46 | -3.07 | 2.14 | 0.728 | -15.65 | -20.49 | -10.80 | <0.001** |
| Charity rate | 0.91 | 0.50 | 1.31 | <0.001** | 0.94 | 0.51 | 1.36 | <0.001** | -0.97 | -2.84 | 0.90 | 0.311 |
| Consecutive participation | -0.01 | -0.09 | 0.06 | 0.714 | -0.01 | -0.09 | 0.07 | 0.843 | 0.35 | 0.22 | 0.47 | <0.001** |
| Consecutive response | -0.14 | -0.20 | -0.08 | <0.001** | -0.14 | -0.20 | -0.08 | <0.001** | -0.12 | -0.21 | -0.02 | 0.021* |
| Consecutive non-contact | 0.05 | 0.00 | 0.09 | 0.059 | 0.05 | 0.00 | 0.10 | 0.052 | 0.66 | 0.50 | 0.81 | <0.001** |
| Consecutive refusal | 0.81 | 0.60 | 1.03 | <0.001** | 0.84 | 0.61 | 1.08 | <0.001** | 1.03 | 0.65 | 1.41 | <0.001** |
| Consecutive non-refusal | -0.01 | -0.22 | 0.19 | 0.905 | -0.01 | -0.22 | 0.21 | 0.956 | 0.49 | 0.19 | 0.79 | 0.001** |
| Consecutive charity donations | 0.06 | 0.00 | 0.12 | 0.044* | 0.06 | 0.00 | 0.12 | 0.046* | 0.20 | 0.10 | 0.30 | <0.001** |
| Change from interview to other | 0.14 | -0.24 | 0.53 | 0.462 | 0.16 | -0.23 | 0.55 | 0.416 | 0.88 | 0.40 | 1.35 | <0.001** |
| Change from other to refusal | 1.07 | 0.67 | 1.46 | <0.001** | 1.03 | 0.61 | 1.44 | <0.001** | 0.88 | 0.36 | 1.40 | 0.001** |
| Constant | -4.64 | -7.06 | -2.22 | <0.001** | -4.65 | -7.09 | -2.21 | <0.001** | | | | |
| Pseudo R-Squared | 0.147 | | | | | | | | | | | |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level; pooled logit regression model includes the following controls: gender, age, education, capital, born in Australia, English only spoken at home, Indigenous status, another health card, carer status, online/offline population and SEIFA

In addition to these predictors, consecutive participation, consecutive non-refusal, and change from "interview" to "other" all had a positive effect on attrition in the fixed-effect model only. In comparison to the nonresponse models, fewer paradata-derived predictors had statistically significant coefficients in the pooled and random-effects models, which should translate into lower prediction power than for nonresponse.

All in all, the strongest online panel paradata predictors of attrition were a change from other to refusal, consecutive refusal, consecutive response, and charity rate. Participation, non-contact rates and consecutive non-contact were less reliable predictors. While the random-effect model should, theoretically, have proved a better model by which to identify potential attritors, the fixed-effect model should have proved more accurate in establishing when the attrition would actually happen and what the strongest indicators would be for that specific (non)respondent group.

### 5.4.5 Accuracy of predictions for nonresponse: addressing RQ2, RQ3 and RQ4

To extend the analysis in Section 5.4.3 and to answer research questions RQ2 (predictor choice), RQ3 (modeling choice) and RQ4 (required length of time series), a set of pooled logit and random-effect logistic regression models for nonresponse were used (see Table 5.4 for regression including data for all 30 waves).

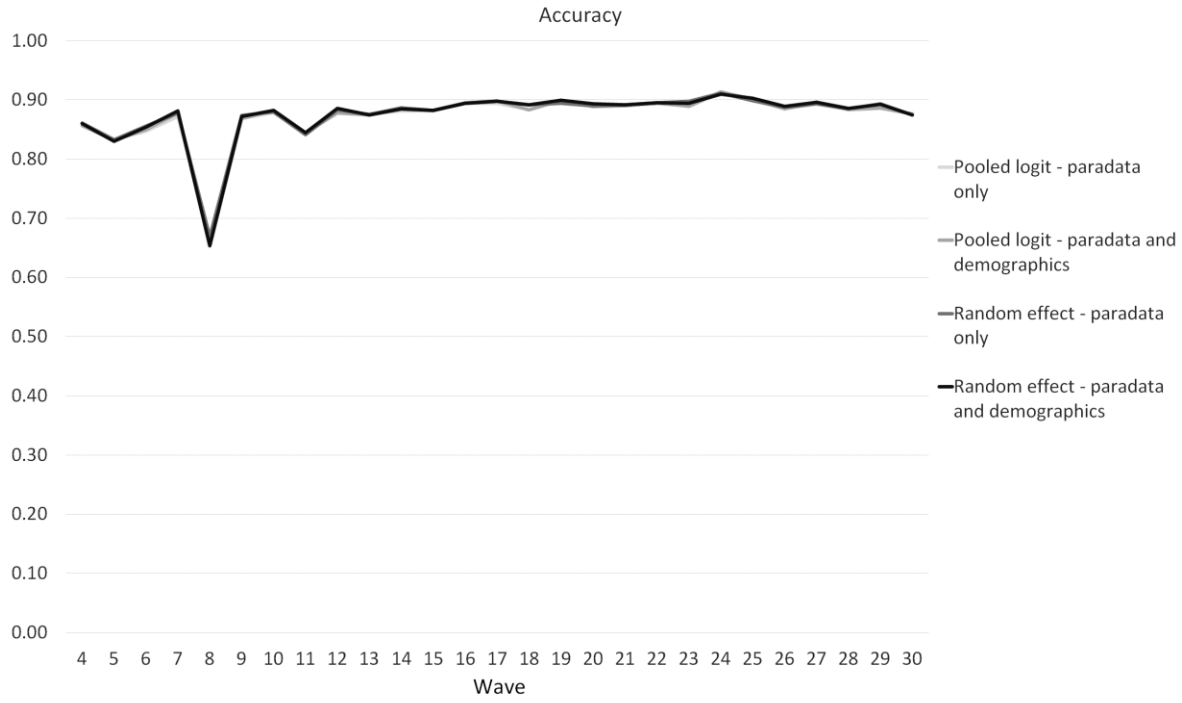
Firstly, we have to emphasize that the accuracy of identifying voluntary attritors was fairly low, i.e., recall was equal to less than 20% in any models we constructed, with or without socio-demographics, pooled or random-effects, and no matter how many future waves were investigated. Instead, predicting nonrespondents (and treating them) should offer better results in dealing with potential voluntary attrition.

Consequently, we thoroughly investigated how accurately nonrespondents could be identified using their previous panel participation behavior as predictors in the static logit and the dynamic logit regression models. Generally speaking, fixed-effect within-person regression could be more accurate in identifying behavioral indicators of nonresponse and their magnitude (if explanatory variables are correlated with the error term), but it would not be possible to use its model coefficients to calculate the predicted probabilities for each respondent.

The aim of this analysis was not only to compare the accuracy of using logit models against dynamic random-effect models, but also to compare the prediction efficiency of combining two groups of predictors, namely: (1) online panel paradata derived variables only, and (2) online panel paradata derived variables and socio-demographics combined. To compare the prediction power, we presented two key statistics: accuracy and recall. The results are presented for waves 4–30 since we needed at least three waves of data to derive certain behavioral predictors and to avoid multicollinearity. We

used online panel paradata for waves 1–3 to predict nonresponse in wave 4, paradata for waves 1–4 to predict nonresponse in wave 5, and paradata for waves 1–29 to predict nonresponse in wave 30. Accuracy was used as a metric for correct identification of both respondents and nonrespondents in the subsequent wave, while recall, calculated as true positives divided by all actual positives, was used as a metric for correct identification of nonrespondents only. Since the propensity for survey completion was about four times as high as nonresponse in Life in Australia™, accuracy of any model (or even random selection) should naturally be higher than recall. As we worked with full online panel paradata including response numbers for all 30 waves of data collection, we did not need to estimate nonresponse in the subsequent waves to determine the target number of nonrespondents identified with our prediction models, something that would need to be done in real life situations. This way, precision as the third metric typically reported in data science to evaluate algorithms, equals to recall and thus does not need to be reported. We identified nonrespondents by selecting the units with the highest probability of nonresponse calculated with our models. Remembering that wave 8 was a partial survey, the results for all 30 waves are presented in Figures 5.4 and 5.5.

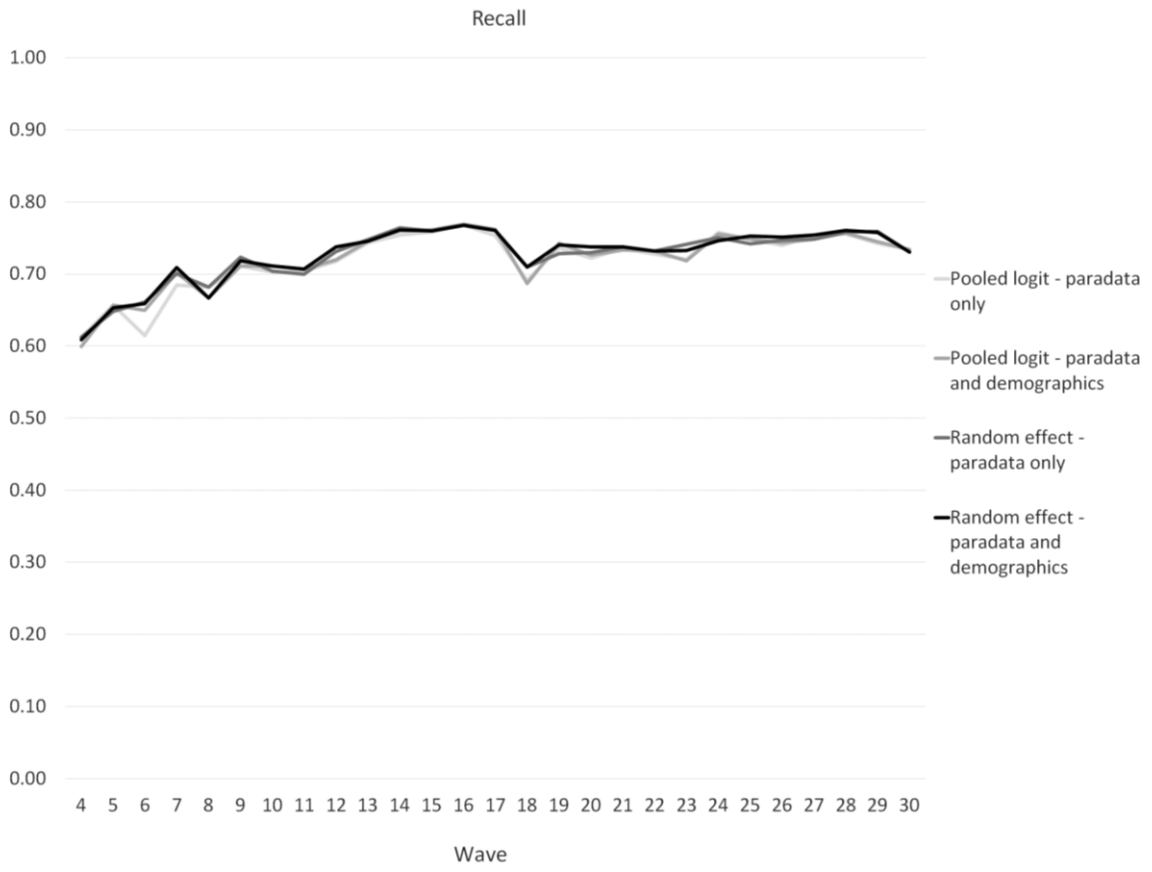
Figure 5.4: Predictive power for response and nonresponse combined, waves 4–30 (Accuracy)




The accuracy curves in Figure 5.4 show the total accuracy of identification of both respondents and nonrespondents in a certain wave. To answer RQ4, we can conclude that we already achieve more than 87% accuracy with six waves of data. It is also evident that the prediction accuracy improved further over time with more data, it peaked in wave 24 (91%), and slightly declined in the remaining six waves. Wave 8 is an exception, since only about 100 panellists were invited to participate. To

answer RQ3 and RQ4: we observed very little to no differences in accuracy between logit and random-effect models, and models with or without socio-demographic predictors. Initially, models with online panel paradata predictors were more accurate, since there were about 4% of panellists with incomplete socio-demographic data, and this missingness was also associated with a lower propensity to respond in a particular wave. We corrected this problem with multiple imputations, resulting in about 4% improved accuracy of models including socio-demographic variables. Now, there is almost no difference (see Figure 5.4).

Figure 5.5: Predictive power for nonresponse, waves 4–30 (Recall)




The recall curves in Figure 5.5 show how accurate is identification of nonrespondents in a certain wave. It is evident that the predictive power improved over time with more data, but it peaked earlier than accuracy (see Figure 5.4) – in wave 16 (77%). These results compare favorably to simpler identification approaches, such as identification nonrespondents in wave *k* by looking at nonrespondents in wave *k-1*. That approach is about 6% less accurate on average (for waves with the majority of panellists invited), suffers from a loss of predictive power in a wave subsequent to a wave with a small proportion of all panellists invited (waves 8 and 9), and is generally less reliable (an increased variability of recall, e.g., up to 11% lower accuracy in wave 11 compared to random effect models).

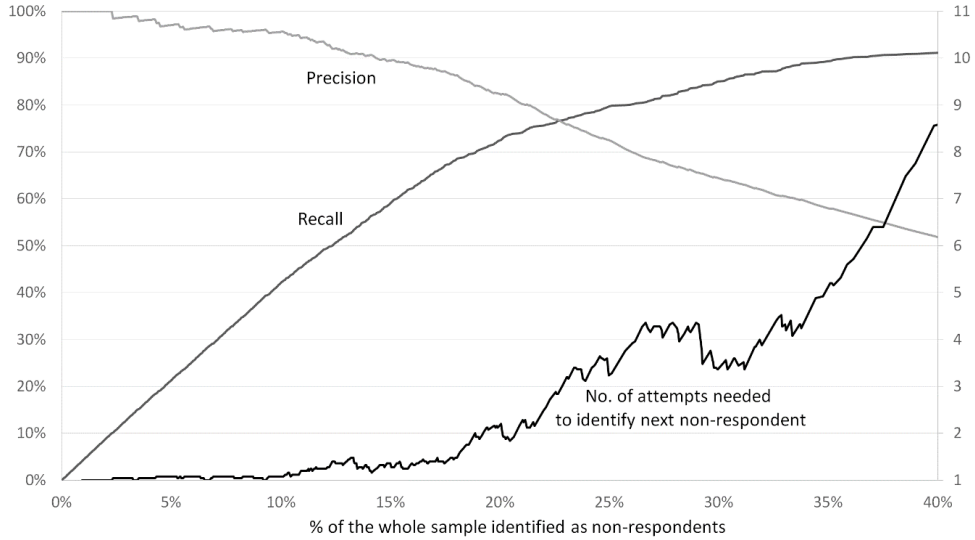
To answer RQ4, we can conclude that we can achieve fairly good accuracy with 15 waves of online panel paradata, identifying more than 3 of 4 nonrespondents in wave 16. After wave 17, about 10% of panellists were retired due to inactivity, which means that a significant portion of the sample, for which nonresponse was easy to predict, was lost. This drop of recall can be seen in wave 18, but it again increased gradually over time and almost reached wave 16 numbers in wave 29 (76%). To answer RQ3: we observed minor differences between different logit models – random-effect models were about 1% more accurate than logit models on average, but in only two waves by more than 2%. To answer RQ2: this time, we again cannot observe any differences between models with different ranges of predictors, and multiple imputations for missing socio-demographic information improved efficiency by about 3% in the models including socio-demographic predictors.

### 5.4.6 Cost-benefit analysis of prediction and post-prediction treatment (RQ5)

To extend the findings, to turn them into practical solutions, and to answer RQ5 (cost-benefit problem), we will show the relationship between recall and precision. It will be presented conditional on the target proportion of panellists with the highest probability of nonresponse, selected to identify nonrespondents. Having in mind that organizations managing online panels could in practice identify potential nonrespondents for different purposes (e.g., see Lugtig 2014), we will show the results of our "cost-benefit" analysis. The "cost" in our case is identifying potential nonrespondents and treating them to prevent them from not participating in future panel surveys; that increases costs of panel management. The "benefit" is identifying those who would not respond in the upcoming survey and successfully convincing them to participate in future panel surveys. However, as identification cannot be 100% accurate, we would also treat respondents who would normally respond without interventions[48]. Our cost benefit analysis is in the form of the number of attempts needed to identify the next nonrespondent by selecting the panellist with the next highest calculated probability of nonresponse (probability calculated with random effect model, range 0-100%). For this particular exercise, we used the data for the wave with the highest recall score (wave 16). The results are shown in Figure 5.6.

[48] Treatment could be any panel management solution proven to increase survey completion of less frequent respondents.

Figure 5.6: The relationship between recall and precision, “cost-benefit” analysis (wave 16, n=2727)




In Figures 5.4 and 5.5 the selected number of panellists with the highest calculated probability of nonresponse equaled to the real number of nonrespondents in the subsequent wave. But in Figure 5.6, we are showing the relationship between precision and recall at different proportions of panellists selected for nonresponse identification. The recall and precision curves cross at about 23%, which was the nonresponse rate in wave 16. Around that proportion of the whole sample identified as nonrespondents, recall curve starts flattening. In practice, the result of this flattening means a higher proportion of false positives. This is confirmed with the line showing the number of attempts needed to identify the next nonrespondent – while almost every panellist with the top 10% (top decile) calculated probability of nonresponse is an actual nonrespondent (value 1 or just above 1), we would need about three attempts, including two false positives, to identify one nonrespondent with a calculated probability around the 75$^{th}$ percentile of probability of nonresponse. We could argue that in that region costs already exceed benefits – for example, to decrease nonresponse, we would offer extra monetary incentives to three potential nonrespondents, but only one of them would actually skip participation in that particular wave without the treatment. With our models, we could correctly identify 90% (or more) of all nonrespondents, but for a high price of about five false positives for one true positive for the last few nonrespondents to reach recall=0.9. This chart shows how different approaches, either more or less conservative or progressive, can be taken based on expected cost-benefit balance.

To answer RQ5, we showed a practical example of identification cost-benefit analysis in a particular wave. We determined that the right balance between “costs” and “benefits”, when identifying

nonrespondents, was around the expected response rate in the upcoming wave[49]. There are a few practical reasons for identification of nonrespondents that could later as well become voluntary attriters. They could be either treated with tailor-made incentives or special panel maintenance approaches (e.g., thank-you or birthday cards) to increase response, which could lead to better representation, higher data quality, more complete time-series, or a delayed recruitment of a refreshed sample. The other aim of identification could as well be inviting panellists conditional on their response propensity to achieve higher response rates while controlling for other representation errors. There might be other uses of accurate identification of less active panellists and all the above should be tested carefully and experimentally. Nonetheless, we would argue – based on the results presented in the paper – that paradata and the types of analyses we have conducted can help with the targeting of interventions.

## 5.5 Discussion

Online panel paradata are able to capture the entire history of panel activity for each member, are considered a new class of paradata (Callegaro 2013), and can be classified as the "prior survey phase" type of paradata (McClain et al. 2019). As such, they offer significant research opportunities from a methodological perspective and can contribute to the development and implementation of various panel management solutions. Baker et al. (2010) argued that at the very least the differences between respondents and nonrespondents should be characterized, although this is in practice seldom carried out. Moreover, the richness of this type of data might also aid understanding of panel members' behavior, predict their future participation, and adjust panel management activities. On the one hand, the longitudinal nature of the data can have negative effects on total survey error (Groves et al. 2009) as nonresponse bias can gradually increase over time due to differential nonresponse and voluntary attrition. On the other hand, in contrast to cross-sectional questionnaire navigation and device paradata, online panel paradata can be restructured into longitudinal panel data for inclusion in different panel data analysis models. The results in this study also partially supported the assumption that controlling for unobserved heterogeneity could improve our understanding of what predicts nonparticipation, and, even more importantly, the accuracy of regression models investigating panel participation (RQ3).

One key outcome of this study was the identification of a number of socio-demographic predictors of nonresponse and attrition. Some of these, such as age and online-offline populations, were predictors of both of these online panel participation outcomes. However, some of these predictors were specific to either nonresponse or attrition, although response/nonresponse rates were in fact good predictors

[49] It could be estimated by reviewing panel survey completion trends over time.

of attrition in a particular wave. After controlling for other socio-demographic characteristics, gender, education and languages other than English spoken at home were identified as significant predictors of nonresponse only. We identified some level of differential nonparticipation (RQ1). The findings on the predictors of nonresponse mostly accord with the findings published by other authors such as Watson and Wooden (2009) who reported lower response rates among the youngest (but also the oldest) participants, the least educated and those not born in Australia, but no differences in response rates between the genders. The findings related to attrition in this study were somewhat similar to those presented in the literature, which still offers contradictory evidence, and it is believed that demographic variables have less explanatory power than socio-psychological variables (Lugtig 2014).

The variables derived from the online panel paradata, such as the survey outcome rates or the consecutive waves with a particular survey income, were shown to be reasonably good predictors of panel participation. In the case of attrition in a particular wave, about one-third (random-effect) and all except for one (fixed-effect) independent variables had an association with the outcome variable. That indicates that identification would be much easier for nonresponse than attrition. In terms of the type of panel data analysis (RQ3), it was concluded that random-effect models were a better approach for identifying nonrespondents and attritors, since the coefficients could be used to calculate the probabilities of the survey outcomes in contrast to fixed-effect models. On the other hand, the fixed-effect within-person regression model provided a better understanding of respondents' behavior prior to nonresponse or attrition. This means that while the probabilities could not be calculated for each panel member, the organization managing the panel could identify those behavior patterns which would generally lead to voluntary attrition (or nonresponse) using fixed-effect models and then implement measures to prevent this from recurring. There are several ways of dealing with potential attritors, such as by tailoring incentives to specific respondents, providing extra information about the panel, or sending thank-you cards (Lugtig 2014). One area of future research for the authors is to use the findings in this article and random assignment of some of these approaches to test for differential treatment effects.

In this study, a promising level of accuracy (and consistency of prediction) was achieved in identifying nonrespondents by using predictors derived from online-panel paradata variables in pooled logit models, and random-effect models controlling for unobserved heterogeneity (in comparison to simpler identification approaches, such as using unit nonresponse in a previous wave). We also presented practical solutions in finding the most suitable cost-benefit balance in prediction and treatment of nonresponse (RQ3). While including socio-demographic predictors in addition to online panel paradata derived predictors did not increase the predictive power (RQ2), controlling for unobserved heterogeneity increased nonrespondent identification by almost 1% on average (RQ3).

The most effective way of improving identification results in practice, in terms of predictor choice, was actually imputing missing values for socio-demographics using multiple imputations. By answering RQ4, we concluded that we can achieve sufficient accuracy with 6 waves (identification of both respondents and nonrespondents) or 15 waves (identification of nonrespondents only), with accuracy slowly increasing over time. Yet, we believe the predictive power could be further enhanced or the models improved in such a way as to achieve the same accuracy with shorter paradata time series. Combining the online panel paradata with other types of paradata, such as questionnaire navigation and/or device paradata, including other good socio-demographic or socio-psychological predictors of nonresponse and voluntary attrition reported in the literature, might increase the accuracy. Combining panel data analysis and machine learning methods, i.e., performing ensemble modeling/stacking for producing better predictions, would be an interesting space for future research in survey methodology as well. An alternative solution for voluntary attrition worth investigating would be the use of the same data in different statistical models, which might be a better fit for survey participation outcomes with a low average rate. Ultimately though, our paper highlights the significant benefit of collecting and making available online paradata for research and panel management purposes.

## 5.6 References


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

Bartolucci, F., & Nigro, V. (2010). A Dynamic Model for Binary Panel Data With Unobserved Heterogeneity Admitting a√ n-Consistent Conditional Estimator. *Econometrica*, 78(2), 719-733.

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

Callegaro, M. (2013). Paradata in web surveys. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information Improving surveys with paradata: Analytic uses of process information* (pp. 261-279). Wiley.

Castiglioni, L., Pforr, K., & Krieger, U. (2008). The effect of incentives on response rates and panel attrition: Results of a controlled experiment. *Survey Research Methods*, *2*(3), 151-158.

Cheng, A., Zamarro, G., & Orriens, B. (2016). Personality as a predictor of unit nonresponse in panel data: an analysis of an internet-based survey. *EDRE Working Paper 2016-12.*

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

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in human behavior*, *26*(2), 132-139.

Frankel, L. L., & Hillygus, D. S. (2014). Looking beyond demographics: panel attrition in the ANES and GSS. *Political Analysis*, *22*(3), 336-353.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167-189.

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

Harris, L., Lee, V. K., Thompson, E. H., & Kranton, R. (2016). Exploring the generalization process from past behavior to predicting future behavior. *Journal of Behavioral Decision Making*, *29*(4), 419-436.

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *Test, 16*(1), 1-22.

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

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based Machine Learning Methods for Survey Research. *Survey Research Methods, 13*(1), 73-93.

Kocar, S. (2020). Attrition. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), *SAGE Research Methods Foundations*. https://www.doi.org/10.4135/9781526421036926973

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information* (Vol. 581). John Wiley & Sons.

Kruse, Y., Callegaro, M., Dennis, J., DiSogra, C., Subias, S., Lawrence, M., & Tompson, T. (2009). Panel conditioning and attrition in the AP-Yahoo! news election panel study. *Proceedings of the 64th Conference of the American Association for Public Opinion Research*, 1-15.

Lugtig, P. (2014). Panel attrition: separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research, 43*(4), 699-723.

Lugtig, P., & Blom, A. (2018, July 25-27). *Using paradata to explain attrition* [Conference presentation]. Methodology of Longitudinal Surveys Conference, Essex, United Kingdom.

Lynn, P. (2017). From standardised to targeted survey procedures for tackling nonresponse and attrition. *Survey Research Methods, 11*(1), 93-103.

McClain, C. A., Couper, M. P., Hupp, A. L., Keusch, F., Peterson, G., Piskorowski, A. D., & West, B. T. (2019). A typology of web survey paradata for assessing total survey error. *Social Science Computer Review, 37*(2), 196-213.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information Improving surveys with paradata: Analytic uses of process information* (pp. 43-72). Wiley.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological bulletin*, *124*(1), 54-74. https://doi.org/10.1037/0033-2909.124.1.54

Roßmann, J., & Gummer, T. (2016). Using paradata to predict and correct for panel attrition. *Social Science Computer Review*, *34*(3), 312-332.

Schifeling, T. A., Cheng, C., Reiter, J. P., & Hillygus, D. S. (2015). Accounting for nonignorable unit nonresponse and attrition in panel studies with refreshment samples. *Journal of Survey Statistics and Methodology*, *3*(3), 265-295.

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

Ward, M. M., & Leigh, J. P. (1993). Pooled time series regression analysis in longitudinal studies. *Journal of clinical epidemiology*, *46*(7), 645-659.

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 157-182). Wiley.

# Chapter 6 Do we have to mix modes in probability-based online panel research to obtain more accurate results?

## 6.1 Introduction

### 6.1.1 Mixing modes in probability-based online panel research

Mixed-mode survey research is becoming increasingly common. The use of email, and in particular, web surveys offers a range of opportunities for mixing modes of data collection (Bryman 2016, p. 232). There are many reasons for employing mixed modes, but the following three are especially common: to reduce costs, to maximize responses, and to save money in longitudinal surveys (Groves et al. 2009, p. 175). In addition to these benefits, probability-based online panels often apply two or more data collection modes to cover both the online and offline populations (Baker et al. 2010). While some probability-based panels collect data online only (e.g., Norwegian Citizen Panel), others combine the online mode with telephone (e.g., Life in Australia™), mail (e.g., GESIS Panel), and face-to-face (e.g., KAMOS) data collection as the offline modes, or even carry out mixed-device data collection (e.g., American Trends Panel, ELIPSS or LISS providing tablets with internet access) (Kaczmirek et al. 2019, pp. 4-5). In addition to recruitment methodology and probabilistic sampling, purposely covering the offline population is one of the characteristics that make probability-based panels different from volunteer/opt-in/access panels. One of the criticisms of these nonprobability-based panels is that they possibly introduce noncoverage bias by systematically excluding the offline population (Pennay et al. 2018).

Generally speaking, mixing modes in probability-based online panel research might be necessary since internet-only samples may not be representative of the general adult population due to significant differences in demographic and other characteristics between the online and offline populations. For example, in the United States in 2015, it was reported that 11% of adults did not self-identify as internet users, and there were differences between the online and the offline populations in terms of age, race, marital status, education, and income (Keeter et al. 2015). In Australia, there are notable differences between online and offline populations in terms of age, location (urban-rural), employment status, qualifications, gender, household income, and country of birth (De Vaus 2013, pp. 76-77). In 2016/2017, it was estimated that about 14% of Australian households did not have home internet access (Australian Bureau of Statistics 2018). In addition, not every person with an internet connection has the skills or inclination to participate online, meaning that an offline survey mode should be included or at least considered in probability-based panel research (Pennay et al. 2016a).

### 6.1.2 Undercoverage bias in online panel research

Including both online and offline populations in online panel research should result in better socio-demographic coverage. For example, the complete LISS panel, which includes both online and offline populations, was found to be closer to the general Dutch population than the internet households only population. Furthermore, non-internet households had lower response rates and higher attrition rates (Leenheer & Scherpenzeel 2013). Socio-demographic bias in data (if observable) can be reduced with different weighting approaches, such as post-stratification weighting which adjusts the sample totals to the population totals using national benchmarks (primary demographics). There might, however, be additional fundamental attitudinal, behavioral, knowledge or other factual differences between the online population (so-called onliners) and the offline population (so-called offliners) which are unobserved in the data, or for which we do not have adequate benchmarks. Kaczmirek et al. (2019) suggest that exclusion of the offline population from online panel research will not only result in socio-demographic representation bias, but in potentially biased estimates for many survey topics.

There has been limited research on the effect of coverage bias in online panels on the accuracy of derived estimates, especially in the case of complete exclusion of the offline population. Furthermore, because internet access and willingness to complete surveys online is changing so rapidly and varies across different country contexts, studies that have been undertaken may need to be updated with more recent data and/or in different geographic/cultural contexts.

Rookey et al. (2008) used a combination of approaches to determine whether the mail mode should be used to obtain responses from the offline population. They observed sizable differences in the estimates for a number of items from different topics. Further, basic post-stratification weighting was inadequate to eliminate these non-socio-demographic differences. Eckman (2016) replicated five published articles which used LISS data but excluded all non-internet households from the samples and then compared the findings. Mean estimates were found to be more sensitive than estimates from multivariate models. When comparing means from five models, between 6.9% and 68.9% of model variables had significant undercoverage. The author also concluded that undercoverage of offliners would not introduce bias into most of the studied multivariate models. Most of the findings and conclusions in the original published articles using the LISS data also would not change. One reason for this might have been the particularly high internet penetration rate in the Netherlands (Eckman 2016, p. 55). In addition to studying socio-demographic coverage bias, Keeter et al. (2015) compared more than 400 survey items to evaluate the size of the bias and found evidence that more than two-thirds of estimates based on full and web-only samples differed by only 1% or less. However,

they identified certain groups with much greater differences between the online and offline populations, such as those 65 years of age and older.

### 6.1.3 Attitudinal, behavioral, and factual differences between the online and offline populations

The literature indicates that there are significant differences between online and offline populations in probability-based online panel research, with or without statistically significant undercoverage bias and its effect on the final estimates. In some cases, adjusting for socio-demographic differences (with weighting or regression models (Rookey et al. 2008)) decreases or eliminates those differences and in other cases, it has little effect (Zhang et al. 2009). The differences between the populations are best captured in topics strongly related to internet access (Eckman 2016) and internet and technology (Keeter et al. 2015). Zhang et al. (2009) reported a series of behaviors and other measures, such as voting actions, political attitudes, civic and political actions, sexual orientation, and self-perception, in which differences between the populations were observed. In addition to the basic socio-demographic differences, such as age, education, or income, Bosnjak et al. (2013) identified personality differences, Blom et al. (2015) identified purchasing power differences, and Keeter et al. (2015) identified political knowledge and financial circumstance differences. Rookey et al. (2008) concluded that online and offline respondents differ in about one-third of attitudinal and behavioral questions, but there were no trends in the direction, questionnaire section, or particular type of question (political opinion, opinion on other countries, ever visited country, healthcare, financial) in which differences were observed.

### 6.1.4 Estimation of survey accuracy with benchmarking

There are at least two ways of estimating the effect of undercoverage bias on the accuracy of estimates. One way is by comparing survey results including the offline population with those excluding this population (see Eckman 2016; Keeter et al. 2015; Rookey et al. 2008). The other approach is to compare the results obtained with and without the offline population with the estimates derived from a representative external data source – usually an expensive and sufficiently large government survey with great attention to data quality and accuracy of survey estimates (Bialik 2018). In certain cases (including in Australia, where this study was undertaken), official statistical agencies are able to compel potential respondents to complete their surveys with the use of financial sanctions for those that do not comply. Generally, the practice of benchmarking is often used to study the accuracy of nonprobability-based online panels in comparison to probability-based ones (e.g., Pennay et al. 2018; Yeager et al. 2011), to perform mode effect analyses (Vannieuwenhuyze & Loosveldt 2013), and to check the accuracy of findings in surveys and determine how to improve

survey quality (Bialik 2018). Benchmarking analysis can also represent added value because the differences in distributions, which could be attributed to measurement mode effects in mixed-mode online panels, can add a net effect on undercoverage bias. Another advantage of high-quality government survey benchmarks is that they are often carried out with single-mode data collection (Vannieuwenhuyze & Loosveldt 2013). On the other hand, the disadvantage of benchmarking analysis is that the required national representative data for non-factual and knowledge items are often not available, and in some cases, there is less trust in the validity of benchmarks (Singh 2011).

### 6.1.5 Outline of the study

Against this background, the primary objective of this study was to establish, from the coverage error perspective, how necessary it is to cover the offline population by mixing modes in online panel survey research (see Total Survey Framework, Groves et al. 2009). To this end, this research addressed the problem of undercoverage bias and its effect on the accuracy/consistency of estimates using a different population context. Additionally, three issues related to undercoverage of the offline population were investigated: (1) undercoverage bias estimation, also conditional on the size of the offline population, (2) identification of attitudinal, behavioral, factual, etc. differences between onliners and offliners, and (3) accuracy estimation relative to the nationally representative estimates with benchmarking analysis. Using the Life in Australia™ survey data and the Online Panels Benchmarking Study (OPBS) 2015 data, the current study aimed to address the following research questions and test the following hypotheses:

*RQ1: How much undercoverage bias would there be if the offline population was partially or completely excluded from probability-based online panel research?*

The aim was to determine the differences in univariate results that could be expected in Australia if the offline population was excluded. To answer this research question, this study explored whether the findings of Eckman (2016) on undercoverage bias could be replicated in Australia. The focus was not on socio-demographic items but rather on other factual, as well as attitudinal, behavioral, and knowledge items for which there are very limited or no benchmarks and which cannot be adjusted with weighting. Simple examples of how much univariate estimates for these items differed when the offline population was excluded are presented, and the results demonstrate how these estimates are affected if the portion of the offline population is manipulated in the combined probability-based online panel sample.

The findings of Eckman (2016, p. 47) on the extent of undercoverage bias in LISS from the Netherlands indicated that 39.8% of survey items exhibited significant bias (five models, 37 out of 93 items). Since it is argued that countries with lower internet penetration rates should experience more

undercoverage bias in probability-based panel surveys (The Netherlands, 2013; 95% [Eurostat 2020], Australia, 2016/2017, 86% [Australian Bureau of Statistics 2018]), the following hypothesis was posed:

H1: Undercoverage bias at the univariate level in the Life in Australia™ surveys will be present for more than 40% of all items.

*RQ2: What question and variable characteristics, such as question topic, represent the biggest differences between onliners and offliners?*

In answering RQ1, an estimate of undercoverage bias is derived and the generalized findings are presented. To explore the bias in more detail, this study also focused on different characteristics of survey questions. The theory suggests that there are significant differences between the online and offline populations in various aspects (Blom et al. 2015; Bosnjak et al. 2013; Eckman 2016; Keeter et al. 2015; Rookey et al. 2008; Zhang et al. 2009), but, to date, there has been no comprehensive comparison of the magnitude of those differences in terms of question topics and other variable characteristics. This study used an innovative approach to determine the consistency of undercoverage bias across topics, controlling for question content, variable type, and number of categories. The aim was to present practical implications for organizations managing probability-based online panels. The following hypothesis, derived from the literature (see Subsection 6.1.3), was posed:

H2: In addition to topics strongly related to internet access and use, there will be observed differences across the majority of topics with no particular trend.

*RQ3: Does post-stratification weighting reduce the differences between onliners and offliners?*

Weighting of data was investigated and the current findings were compared with results in the literature on this topic (e.g., Rookey et al. 2008). In doing so, the aim was to determine if the differences between onliners and offliners are present due to the socio-demographic structure of the sample (e.g., offliners tend to be older) or if there are other fundamental differences between the populations that cannot be (fully) adjusted for with post-stratification weighting.

*RQ4: How much does including the offline population improve the accuracy of estimates relative to the nationally representative benchmarks?*

While we can estimate the extent of bias and determine what survey topics are responsible for more bias than others, this does not provide evidence on the effect of mixing modes and the inclusion of offliners on the accuracy of estimates. Thus, this study used a number of nationally representative benchmarks to answer this final research question and determine if mixing modes improves, deteriorates, or does not significantly alter estimates in probability-based online panel research.

The literature suggests that offline populations are different to those who participate in online surveys, and to produce high-quality estimates in probability-based online panel research in countries with a notable percentage of people with no internet access, offliners should be included. Thus, the following hypothesis was posed:

H3: Online-offline probability-based online panel samples produce more accurate estimates compared to online-only samples.

## 6.2 Methods

The methods for the two disjointed but related perspectives in this study on mixing modes in probability-based online panels are presented separately below: (2.1) Undercoverage bias perspective (providing evidence to answer RQ1-RQ3), (2.2) Accuracy of estimation perspective (providing evidence to answer RQ4).

### 6.2.1 Undercoverage bias perspective

Probability-based online panel survey data were used to answer the first three research questions. The nature of probability-based online panels, especially those with relatively smaller samples like the Life in Australia™ panel, is that the same respondents (so-called panellists) are invited to participate in the majority of surveys. The data used in the undercoverage bias perspective aspect of this study were consequently collected from more or less the same respondents. As there are different types of online panellists with different kinds of response behavior, from frequent respondents to those who voluntarily attrit after participating in a few surveys, the survey subsamples overlapped but did not match (from 78% of the whole panel in Wave 1 to 59% in Wave 14).

#### 6.2.1.1 Data

For this part of the study, data from the Life in Australia™ survey were analyzed. Specifically, six out of the first 16 waves before the first panel refreshment in June 2018 were used in this study. Life in Australia™ is the only probability-based online panel in Australia and was established and is managed by the Social Research Centre. The panel has been used to collect data on important topics for different clients, from academic to government and non-governmental organizations (see the list of studies in Kaczmirek et al. 2019, p. 20). However, as those research projects were funded by different clients, the current study only had access to the data collected for the Australian National University (ANU) as the largest Life in Australia™ client (waves 1, 2, 3, 7, 10, and 14). More information about the surveys is provided in Table 6.4 in the Appendix 6.

#### 6.2.1.2 Population, samples, and data collection modes

In Life in Australia™, the panellists are defined as "Australian residents aged 18 years or older" and were recruited in the second half of the year 2016 (n=3,322). The response rate at the establishment of the panel, calculated as the product of the recruitment rate and the profile rate, was 15.5% (AAPOR RR3 (The American Association for Public Opinion Research 2016)). To undertake recruitment, a dual-frame Random Digit Dialing (RDD) sample design was employed, with a 60:40 (pilot) and 70:30 (the main recruitment effort) split between mobile phone and landline sample frames. The last birthday method was used to select potential panel members in landline frames and the phone answerers were selected for the mobile sample; only one person per household was invited to join the panel. Out of all panellists who were recruited, joined the panel, and were later invited to monthly surveys on different topics, about 87% can be defined as online (onliners) and about 13% as offline panellists (offliners). The online self-completion mode (CAWI) was used to collect data from the online panellists and the telephone mode (CATI) was used to cover the offline population. Data were collected at approximately monthly intervals. An incentives scheme was used for recruitment and monthly data collection - $10 per wave, with panellists either receiving the incentives or donating to charity (Kaczmirek et al. 2019). As can be seen in Table 6.4 (in the Appendix 6), the Life in Australia™ survey sample size decreased with each survey, which is a result of an increasing number of nonrespondents over time, as well as accumulating voluntary panel attrition.

#### 6.2.1.3 Data processing and analysis

To estimate undercoverage bias at the univariate level and present evidence to answer RQ1, the following Equation 6.1 from Eckman (2016) for absolute relative bias was used:

$$absolute\ relative\ bias\ (\overline{Y}_c) = \left|\frac{\bar{y}_c - \bar{y}_{pop}}{\bar{y}_{pop}}\right| \qquad (6.1)$$

where $\bar{y}_c$ is the mean from the online population (excluding offliners) and $\bar{y}_{pop}$ is the mean from the full sample (onliners and offliners). Because the variables were measured in different units, absolute relative bias was estimated and averaged across all items. Since the majority of all items were categorical (nominal and ordinal), dummy variables were created for those variables (e.g., an ordinal variable with five levels generated five dichotomous variables) and their absolute relative bias was compared. In addition to Chi-Square testing (nominal variables), linear (continuous variables), binary logistic (dichotomous variables), and ordinal regression models (ordinal variables) were analyzed. With the response variable being a substantive survey item and the independent variable the population (0=online, 1=telephone), the statistical significance of undercoverage bias was tested, with a significant regression coefficient indicating bias (Eckman 2016).

To extend the bias estimation findings and present evidence to answer RQ2, multiple linear regression models were created (see Equation 6.2):

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon \qquad (6.2)$$

where Y is the effect size, $X_1$ - $X_n$ are the survey item characteristics such as item topic and question content, and $\epsilon$ is the error. Comparison of the distributions of onliners and offliners was carried out by calculating effect sizes as measures of association between pairs of variables, i.e., substantive survey items from Life in Australia™ research (see Table 6.4 in the Appendix 6, six surveys, 368 items) and mode of completion (0=online, 1=telephone). The variable information was coded for a total of 368 variables from six Life in Australia™ waves. Using the European Language Social Science Thesaurus (ELSST) (UK Data Service, n.d.), broad survey item topics were identified and combined into 20 broad distinctive topics – the most common was *values and social capital* (12.8%), followed by *housing* and *finance* (both at 7.3%). To code the question content by type, the classification by Dillman (1978) was used; out of the four types, the combined attitudes and beliefs category was the most common type (65.5%), followed by behaviors (19.0%). The following variable types were used in the models: binary, nominal with 3+ categories, ordinal, and continuous (combining interval and ratio variable types). The most common variable type was ordinal (50.5%), followed by binary (33.7%). The most common/modal categories were used as reference categories in the regression models presented in the Results section. Effect size measures were calculated with both unweighted and weighted data. By weighting survey data, the sample totals were adjusted to the selected population totals for both onliners and offliners separately. It was assumed that weighting would decrease some of the undercoverage bias.

The calculated differences between both populations were based on Cramer's V and Rank-Biserial Correlation measures; a higher coefficient value represented a greater difference between the populations in the concept measured by each one of the 368 substantive survey items. While, in practice, various bivariate measures of association are used for pairs of variables of different types and distributions, such as epsilon squared, eta squared, Spearman's rho, or Pearson's r (see the bivariate effect size review from Kocar 2018), not all of them were suitable for this analysis. For example, Bosnjak et al. (2013), who compared sample composition discrepancies in online panels, used Cohen's d (comparing means) and Hasselblad and Hedges's d (percentages). However, this study had to use an effect size measure for nominal variables which would indicate the same magnitude of association regardless of the number of cells in the contingency table or the degrees of freedom. Since the minimum number of either rows or columns was always two (modes: online and telephone), Cramer's V coefficient could be used, whereby *min(r-1, c-1)=2* always equals Phi and Cohen's w values

(see Cohen 1988 for more information). This enabled comparability of coefficients, which would have been more challenging with larger contingency tables. Secondly, due to the fairly low number of interval and ratio variables in the selected Life in Australia™ data (n=17), and as not all of them were normally distributed, non-parametric tests were used for binary variables (*survey mode:* 0=online, 1=telephone) and for ordinal and continuous substantive survey items. This was considered an acceptable adjustment since the Rank-Biserial Correlation measure is based on the Mann-Whitney U test, and the literature indicates that this test is only 5% less effective than a t-test even when the assumption of normality holds (Lehmann 2004, p. 176).

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

- Selection of all substantive survey items in the Life in Australia™ data (six surveys), excluding: (1) those with less than 20% valid responses (to avoid statistical power issues with small samples of offliners), (2) primary socio-demographics which were not asked in each wave but added to the data from the Life in Australia™ profile dataset, (3) open-ended question items, (4) paradata variables. A total of 368 items were selected;
- Coding of variables, adding information on: broad item topic, type of question content, variable type, and no. of variable categories as predictor variables;
- Calculation of post-stratification raking weights for each of the six Life in Australia™ surveys, for onliners and offliners separately (to balance the samples on key socio-demographics) using the selected demography. Post-stratification weighting was carried out to adjust the samples to match the Australian Census distribution by age, gender, education, state, country of birth (Australia, English-speaking background, non-English-speaking background), and telephone status (mobile, landline, dual user);
- Calculation of Cramer's V and Rank-Biserial Correlation (R-BS) proposed by Glass (1965), for each Life in Australia™ substantive survey item in a pair with *survey mode* (weighted data and unweighted data);
- Creation of a new data matrix with Life in Australia™ survey items as cases (rows), and effect size measures (dependent) and coded survey item information (predictors) as variables (columns);
- Construction of multiple linear regression models with *Cramer's V value* and *Rank-Biserial Correlation coefficient* (weighted and unweighted, a total of four models), with effect sizes calculated for each substantive survey item as the response variable, and *broad item topic*, *question content*, and *variable type* as predictor variables;
- Testing for all assumptions of ordinary least squares (OLS) regression and adjustment of the models according to the assumption test results.

For better statistical power, the Life in Australia™ ordinal variables were included in all models, both the ones for categorical variables (with *Cramer's V value* as the dependent variable) and for models with non-parametric effect sizes as dependent variables (with *Rank-Biserial Correlation coefficient*). Since correlation coefficients range from −1 to 1, and we were only interested in the magnitude of effect sizes and not the direction, an absolute version of *Rank-Biserial Correlation coefficient* with positive values only was used. All data processing and analyses, except for multiple linear regression analyses (Stata), were carried out using R software (additional packages used: *survey* and *sjstats*).

### 6.2.2 Accuracy of estimation perspective

Due to the unavailability of high-quality nationally representative benchmarks for the majority of the Life in Australia™ substantive survey items, only one out of the six data sources analyzed in the first part of this study could be used for this second part of the study[50]. Thus, the Health, Wellbeing, and Technology Survey 2017 (also known as Life in Australia™ Wave 2 or OPBS Replication 2017) (Pennay & Neiger 2020) was analyzed to study the accuracy of estimates relative to nationally representative estimates (answering RQ4). To extend the accuracy findings to different online-offline samples, and to control for several other characteristics of mixed-mode surveys, OPBS 2015 data was also used. The questionnaire for OPBS 2015 was designed based on the availability of high-quality benchmarks for Australia and was later replicated in an online panel sample from Life in Australia™.

#### 6.2.2.1 Data

Data from both studies, OPBS 2015 (Pennay et al. 2016b) and OPBS Replication 2017 (Pennay & Neiger 2020) were collected by or for the Social Research Centre (SRC) using a matching Health, Wellbeing, and Technology questionnaire (for more information, see Table 6.5 in the Appendix 6). The OPBS 2015 data were collected quasi-experimentally by the SRC. The primary aim was to determine the accuracy of survey estimates generated from probability-based surveys and nonprobability based online panels, relative to nationally representative benchmarks (Pennay et al. 2018). The study was later replicated on a probability-based online panel sample to provide another point of comparison (Kaczmirek et al. 2019). Both data files can be used to establish the accuracy of online-only samples in comparison to mixed-mode samples.

#### 6.2.2.2 Samples and data collection modes

The following samples from the two selected surveys listed in Table 6.5 in the Appendix 6 were used in this part of the study: the online and offline telephone subsamples from the OPBS Replication 2017

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

and the online, telephone, and mail subsamples from the OPBS 2015. The OPBS Replication 2017 comprises one mixed-mode probability-based online sample and the OPBS 2015 study data comprises three probability-based samples from the Australian population aged 18 years and over (Address-based sampling survey (A-BS), Standalone RDD survey, RDD end of survey recruitment/'piggybacking' survey) and five nonprobability-based samples of participants in online panels (also aged 18+)[51]. The Standalone RDD survey is the only survey with offline-only data collection, and thus, was not included in this study. Data collection was carried out between October and December 2015 (OPBS 2015) and in January 2017 (OPBS Replication 2017). For more information on the subsamples by data collection modes used in the current analysis, see Table 6.6 in the Appendix 6.

The AAPOR Response Rates 3, which could be calculated for the probability-based surveys, were 12.4% for the RDD "piggybacking" study and 26.5% for A-BS study (Pennay et al. 2016a). The completion rate for OPBS Replication 2017 (Life in Australia™ Wave 2) was 78.6% (Kaczmirek et al. 2019).

#### 6.2.2.3 Benchmarks

Benchmarks from some of the largest government-funded national surveys in Australia were used in this study: the Australian Census 2016 (Australian Bureau of Statistics 2015), National Health Survey 2014-15 (Australian Bureau of Statistics 2015), the National Drug Strategy Household Survey 2013 (Jefferson 2015) and General Social Survey 2014, as well as the Australian Electoral Commission (2015) administrative data (benchmarks from Pennay et al. 2018). These surveys should be considered as the highest quality social research data sources in Australia, and the validity of the benchmarks should be the highest. For more methodological details, see Table 6.7 in the Appendix 6.

#### 6.2.2.4 Data processing, weighting, and analysis

In this part of the research, the results from the various surveys listed in Table 6.6 were compared with the nationally representative benchmarks listed in Table 6.7 (both tables are in the Appendix 6). All substantive measures from the OPBS 2015 study (Pennay et. al 2018) which were later used for assessing the performance of Life in Australia™ (Kaczmirek et al. 2019) were selected for use in this study. Our study partially replicated the approach of these previous studies. To measure bias, the average absolute error (AAE) measure proposed by Yeager et al. (2011) was used (see Equation 6.3), which was computed across three categories (secondary demographics, substantive items [see classification in Pennay et al. 2018], and combined secondary demographics and substantive items):

$$AAE = \sum_{j=1}^{k} \frac{|\widehat{y_j} - y_j|}{k} \qquad (6.3)$$

[51] Nonprobability samples were not a subject of this study.

where $\hat{y}_j$ is the j-th estimate from either the OPBS 2015 or OPBS Replication 2017 survey and $y_j$ is the value for a corresponding benchmark. To estimate the accuracy of the online-only samples, the AAE values (bootstraps with bootstrap weights were used for significance testing) and the values of the two bias measures were compared between the online-only and online-offline samples. The absolute relative bias measure (Eckman 2016, see Section 6.2.1.3) from the undercoverage bias estimation was also used in this part of the article, as well as its absolute version – the absolute mean bias measure presented in Equation 6.4:

$$absolute\ mean\ bias\ (\hat{y}) = |y_{combined} - \hat{y}_{online}| \qquad (6.4)$$

where $y_{combined}$ is the estimate from the online-offline sample and $\hat{y}_{online}$ is the estimate from the online-only sample. Just like for AAE, the average (mean) values across all selected items were calculated for the absolute relative bias and absolute mean bias.

Weighted estimates for the selected items and for all analyzed samples, in addition to the unweighted estimates, were calculated to assess the effect of post-stratification on bias. It was decided to employ a consistent approach across all surveys and samples with no base weights derived. Post-stratification raking weights were calculated for each sample separately, i.e., the online-offline and online-only samples, to balance the samples on key socio-demographics. The same primary demographic benchmarks as Pennay et al. (2018, p. 12) and Kaczmirek et al. (2019) were used, while in contrast, the weighting benchmarks were taken from the Australian Census 2016, which was conducted around the time of the OPBS 2015 and OPBS Replication 2017 studies. Post-stratification weighting was carried out to adjust the samples to the national distributions by gender, age by education, state by capital city in state, country of birth (Australia, English-speaking background, non-English-speaking background), and telephone status (mobile, landline, dual user). All larger weights were trimmed down to a value of 5. The random forest technique was used to impute missing values for the listed weighting variables so as not to exclude any cases with valid values for substantive items.

All data processing and analyses were carried using R software. The following packages were used for functions not directly provided by R's base or stats packages: *Hmisc, missforest, anesrake, sjstats,* and *questionr*.

## 6.3 Results

This section will present the results of all analyses and data simulations. This section is divided into the following subsections: undercoverage bias – extent of univariate bias[52], undercoverage bias – survey item characteristics, and accuracy of estimation – benchmarking.

### 6.3.1 Undercoverage bias – extent of univariate bias

This section addresses the first research question, RQ1, and describes the test of hypothesis H1. To do so, the analysis from Eckman (2016) was partially replicated. To showcase the magnitude of differences between the populations, data were not weighted in the following analyses studying bias.[53]

Table 6.1: Undercoverage bias in six Life in Australia™ waves

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

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

The results in Table 6.1 reveal a fairly significant bias at the univariate level. With between 12.9% and 14.1% of offliners participating in the Life in Australia™ surveys, the results indicated that between 52.2% (Wave 3) and 80.0% (Wave 7) of items exhibited significant undercoverage bias, as determined by significance testing with regression modeling and Chi-Square testing. This indicates that offliners are significantly different than onliners and H1 cannot be rejected; more than 40% of items across all six surveys exhibited significant undercoverage bias. Further, dummy variables were generated from all categorical variables to estimate the average absolute bias; as different statistical tests must be used to test for significant differences in categorical variables, relative distance had to be calculated alternatively, like with sets of dummies. In practice, such results are often reported for one variable

[52] 'Undercoverage bias' investigated in this paper is a hypothetical undercoverage bias which would be the result of completely excluding the offline population. Undercoverage bias is, in practice, measured as attitudinal, behavioral, and factual differences between the populations, as well as the effect of those differences on the estimates in case of exclusion of the offline population. As of 2021, Life in Australia™ is a mixed-mode online panel collecting data from offliners as well.

[53] Since Eckman (2016, p. 46) did not use weights and hypothesis H1 was predominantly based on the findings of that study, post-stratification (raking) weights were not used here in the univariate undercoverage bias part of the analysis for comparability purposes. The effect of weighting on undercoverage bias reduction is addressed in the 'survey item characteristics' and 'benchmarking' subsections of the Results.

category only, e.g., the percentage of people strongly agreeing with a particular statement, which justifies the undercoverage bias calculation with dummies. In this study, the average absolute relative bias was between 4.7% (Wave 10) and 6.4% (Wave 1), which is much more than in the study by Eckman (2016). Absolute relative bias seemed to be associated with significant undercoverage bias as examined with dummy variables (and a limited number of interval/ratio variables), and was less severe than the bias observed with the original variables. As categorical variables were split into dichotomous variables with lower proportions, and onliners and offliners might not differ in every single dimension measured by the variable, undercoverage was significant for a smaller portion (between 34.5% (Wave 3) and 63.8% (Wave 2)) of variables/variable categories.

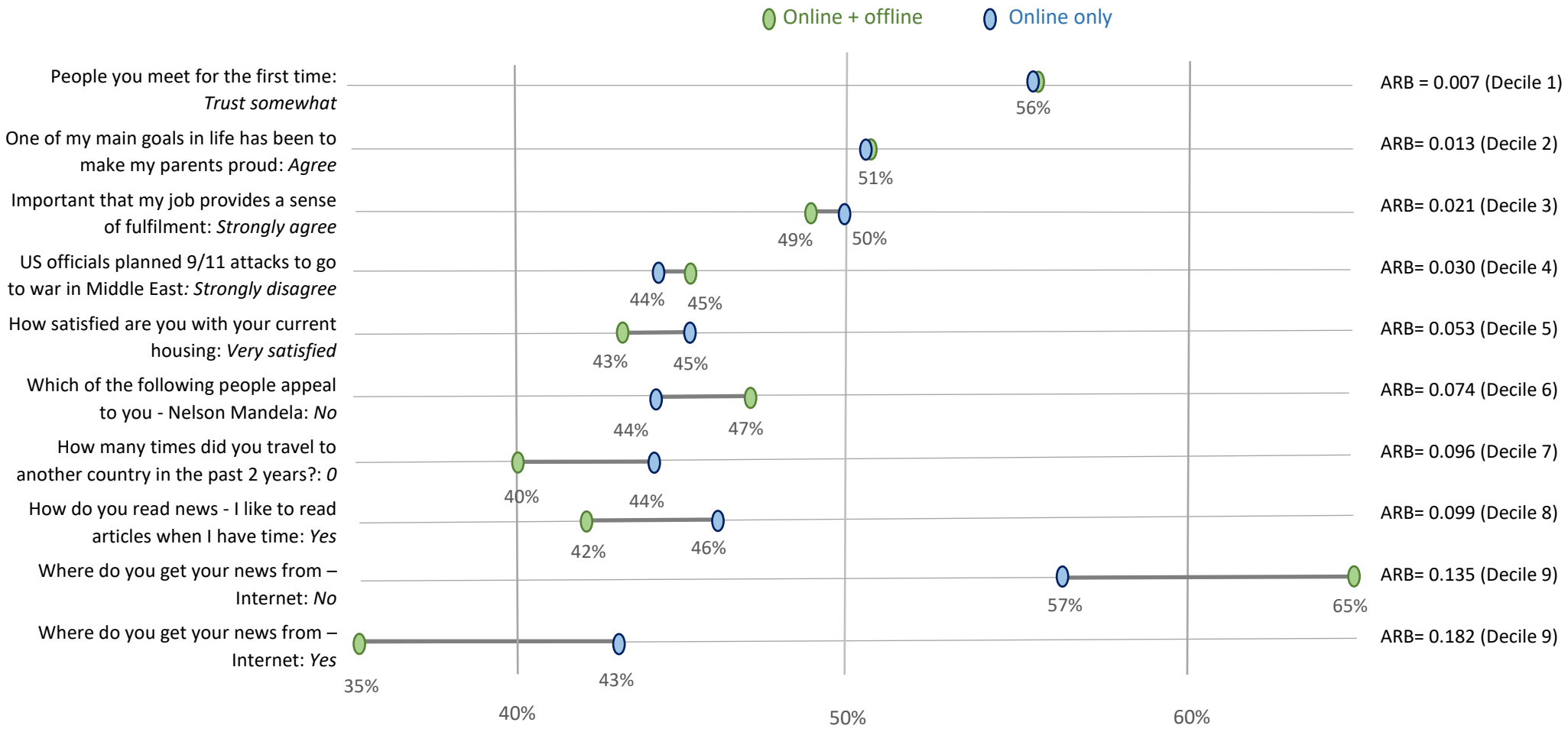
To further illustrate the undercoverage bias and give examples of differences in reported estimates, all 1653 continuous and dummy variables created from categorical variables were ordered by their absolute relative bias (ARB). The items were split into deciles to present all levels of differences in estimates between onliners and offliners. Due to space constraints, only one dummy variable with proportion between 0.4 and 0.6 (online only estimate) from each decile was selected at random for comparability reasons. The differences in estimates for 10 selected items are presented for: (1) onliners and offliners combined (as in the Life in Australia™ surveys), and (2) onliners only.

The results presented in Figure 6.1 show differences in reported estimates in practice, i.e., between real Life in Australia™ estimates and hypothetical estimates if Life in Australia™ was an online-only probability-based panel. For small absolute relative values (Deciles 1-5; half of all items), there were no differences for some items and almost negligible differences for other items. For absolute relative values from Deciles 6-8 (absolute relative bias about 7-10%), there were small differences in estimates. In Decile 9, there were only two variables with proportion between 0.4 and 0.6, but they were derived from the same variable *getting news from the internet*. There were no variables matching our distributional criteria (proportion) in Decile 10. Generally speaking, the differences in reported percentages shown in Figure 6.1 are mostly non-significant. As a rule of thumb, the 95% confidence interval (CI) for a sample sized 3000 is ±1.1% for proportion=10% and ±1.8% for proportion=50%. We previously reported undercoverage bias for more than half of the variables and between one-third and two-thirds of variable categories in the Life in Australia™ surveys (see Table 6.1). Moreover, excluding offliners would mostly result in different proportions being reported (see Figure 6.1). However, we have to note that most confidence intervals of the estimates for onliners and offliners combined and onliners-only overlap, i.e., are “within the margin of error” for estimates presented in Figure 6.1.

The reason for the overlapping CIs is that the effect of undercoverage bias on an estimate is a function of two statistics; the below example demonstrates this. If exactly 50% (p=0.5) of all onliners and 20%

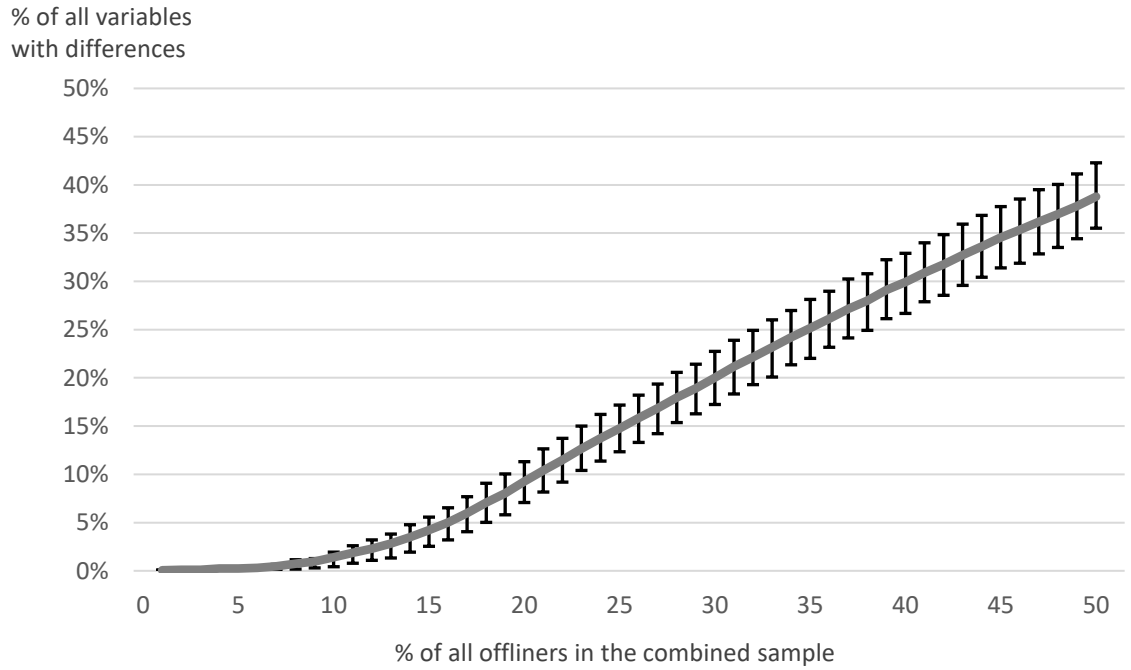
(p=0.2) of all offliners supported a particular political party, and 1 in 10 respondents were offliners, then the reported proportions with and without offliners would have overlapping 95% CIs (0.50±0.018 onliners only, 0.47±0.018 onliners and offliners). Ceteris paribus, if 10% (p=0.1) of offliners supported the party instead of 20% (p=0.2), the CIs would not overlap anymore (0.50±0.018 onliners only, 0.46±0.018 onliners and offliners). A statistically significant difference would also be observed if 20% (p=0.2) of all offliners still supported the party, but there would be twice the proportion of them in the sample, 2 in 10 (0.50±0.018 onliners only, 0.44±0.018 onliners and offliners). Hence, undercoverage bias is more severe if there are larger differences between onliners and offliners and, in the case of notable differences between the populations, if the proportion of offliners in the sample is larger.

Figure 6.1: Differences in reported unweighted proportions if onliners were included or excluded



To extend the findings above and to address the issue of the size of the offline population and its effects on the changed estimates (RQ1), a simulation was performed with real Life in Australia™ survey data (with the dummies generated previously). Online-offline samples were created with between 1% and 50% offliners, and the results were compared to onliners-only estimates. To add data for offliners to the original samples of onliners, sampling with replacement (bootstrapping) was performed and 1000 samples were created for each proportion of offliners (1-50%). All continuous and dummy variables (out of n=1653) with non-overlapping 95% CIs were counted, conditional on the proportion of offliners in the samples.

Figure 6.2: Median percentage of items with changed estimates (based on % of offliners), 95% CI




The results presented in Figure 6.2 show that almost all continuous and dummy variables would have overlapping CIs if the proportion of offliners in the probability-based online panel research was lower than 10%. At 15%, i.e., at about the proportion of households without internet connection in Australia, only between 2-5% of items would have non-overlapping CIs. At 20% of offliners, the portion increases to between 7-12%. After that, the percentage of variables with significant differences increases almost linearly, at about a 1-percentage point increase in items with differences for a 1-percentage point increase in offliners in the sample. In samples consisting of about 40% offliners and 60% onliners, about 30% (±3%) of items would have non-overlapping CIs in comparison to the samples with 100% onliners. The GESIS Panel includes about that proportion of offliners who respond by mail, which means that excluding offliners might not be the best idea in their context, assuming that the magnitude of differences between the populations is comparable to the Australian characteristics.

### 6.3.2 Undercoverage bias – survey item characteristics

To identify the differences between onliners and offliners, which may be more generalizable than only comparing the distributions of individual items (univariate bias) or their dummies, four multiple linear regression models were constructed, as explained in Section 6.2.1.3. Modeling was then performed to address the second and third research questions RQ2 and RQ3 and to test hypothesis H2.

After running the ordinary least squares regressions, the assumption of linear regression was tested in all models. Since there were a number of outliers affecting the normality of the residuals, a few units (i.e., items) were removed based on the following criteria for outlier detection: standardized residuals (as discrepancy measures), leverage (as a distance measure), Cook's distance and DFBETA (as influence measures). In the end, nine outliers out of 351 nominal or ordinal variables were removed from the Cramer's V models and nine outliers out of 202 ordinal or continuous were removed from R-BS coefficient models. It was observed that a number of outliers in the Cramer's V models were *internet* broad topic survey items, and removing them decreased the clearly inflated Adjusted R-Squared coefficients from 0.445 to 0.349 (weighted) and 0.375 to 0.286 (unweighted), respectively. At the same time, the Root Mean Square Errors, as an absolute measure of fit, decreased significantly after removing outliers, which indicates a better absolute fit for both models. While a number of *internet* topic survey items were identified as outliers and removed from the model, the remaining ones were intentionally left in the model to compare the magnitude of differences between *internet* and other topics. In the models with R-BS coefficient values as dependent variables, Adjusted R-Squared increased and Root Mean Square Errors decreased after removing outliers, which meant a better absolute and relative fit in those regression models. Lastly, as the effect sizes were derived from the data collected from the same respondents in the same wave and partially matching respondents in different waves (due to unit nonresponse and voluntary attrition), we had to identify a way of dealing with dependencies in the data so as not to violate any assumptions of ordinary least squares regression. The literature suggests approaches such as panel data analysis, bootstrapping regression models, and regression with clustering. Here, it was decided to carry out a combination of bootstrapping and clustering. Bootstrapping was carried out to mitigate the problem of dependencies and calculate standard errors more accurately (Fox 2015). Clustering was carried out to deal with regression model errors potentially being independent across clusters but correlated within clusters, i.e., waves with a unique sample composition (Cameron & Miller 2015). This was performed using Stata 13.

The results in Table 6.2 reveal some non-negligible differences between onliners and offliners which can be observed for the vast majority of topics. Given that the reference category for *values and social*

*capital* was fairly average in terms of the mean effect size, the non-significant coefficient should be interpreted as no difference between that topic and *values and social capital.* The most significant topical differences measured with Cramer's V were observed for *international relations,* followed by *internet*, although several *internet* items with the highest effect size values were removed after outlier detection, as explained above. Out of the other topics, *public figures and health, media* and *finance* (the last one only after weighting) had average effect sizes and *household and family, science and technology,* and *government and policy* had below-average effect sizes. *Household and family* stood out as a topic with very few average differences between the online and offline population.

Table 6.2: Ordinary least squares regression models with predictors of differences between onliners and offliners (carried out with bootstrapping and clustering – clusters as Life in Australia™ waves)

| | Cramer's V, weighted data | | Cramer's V, unweighted data | | R-BS coefficient, weighted data | | R-BS coefficient, unweighted data | |
|---|---|---|---|---|---|---|---|---|
| Predictors | Beta coef. | p value | Beta coef. | p value | Beta coef. | p value | Beta coef. | p value |
| **Broad topics** | | | | | | | | |
| Values and social capital | 0 | | 0 | | 0 | | 0 | |
| Environment | 0.032 | 0.244 | 0.032 | 0.258 | 0.100 | 0.062 | 0.084 | 0.000** |
| Finance | 0.024 | 0.000** | -0.010 | 0.680 | -0.049 | 0.000** | -0.024 | 0.000** |
| Gender equality | 0.003 | 0.714 | 0.002 | 0.627 | 0.042 | 0.219 | 0.049 | 0.000** |
| Government and policy | -0.015 | 0.000** | -0.026 | 0.000** | -0.030 | 0.000** | -0.037 | 0.000** |
| Health | 0.032 | 0.000** | 0.016 | 0.000** | 0.007 | 0.010* | 0.002 | 0.508 |
| Household and family | -0.063 | 0.000** | -0.069 | 0.000** | 0.063 | 0.148 | -0.155 | 0.007** |
| Housing | 0.004 | 0.886 | -0.003 | 0.844 | 0.023 | 0.632 | 0.031 | 0.348 |
| Internet | 0.114 | 0.000** | 0.166 | 0.000** | 0.328 | 0.000** | 0.466 | 0.000** |
| Labor, employment, work | -0.004 | 0.610 | -0.045 | 0.000** | 0.019 | 0.026* | 0.252 | 0.003** |
| Lifestyle | 0.006 | 0.522 | -0.008 | 0.340 | 0.025 | 0.428 | 0.023 | 0.000** |
| Multiculturalism | 0.009 | 0.611 | -0.018 | 0.555 | 0.034 | 0.291 | 0.083 | 0.003** |
| Politics and elections | -0.017 | 0.038* | -0.015 | 0.023* | -0.038 | 0.000** | -0.024 | 0.231 |
| Science and technology | -0.021 | 0.001** | -0.062 | 0.000** | 0.020 | 0.435 | -0.020 | 0.001** |
| Wellbeing | 0.005 | 0.450 | -0.032 | 0.005** | -0.024 | 0.164 | -0.063 | 0.000** |
| Discrimination | -0.023 | 0.013* | -0.012 | 0.067 | | | | |
| International relations | 0.160 | 0.000** | 0.180 | 0.000** | | | | |
| Media | 0.029 | 0.001** | 0.039 | 0.000** | | | | |
| Public figures | 0.069 | 0.000** | 0.093 | 0.000** | | | | |
| Other | -0.001 | 0.898 | 0.012 | 0.139 | 0.029 | 0.013* | 0.063 | 0.000** |
| **Type of question content** | | | | | | | | |
| Attitudes and beliefs | 0 | | 0 | | 0 | | 0 | |
| Behaviors | -0.006 | 0.415 | -0.006 | 0.608 | -0.031 | 0.261 | 0.003 | 0.430 |
| Attributes | 0.008 | 0.608 | 0.048 | 0.001** | 0.078 | 0.000** | 0.277 | 0.000** |
| Knowledge | -0.024 | 0.064 | -0.070 | 0.000** | | | | |
| Variable type | | | | | | | | |
| **Ordinal** | 0 | | 0 | | 0 | | 0 | |
| Nominal | -0.002 | 0.515 | -0.002 | 0.718 | | | | |
| Binary | -0.039 | 0.003** | -0.059 | 0.000** | | | | |
| Interval/ratio | | | | | 0.083 | 0.046* | 0.088 | 0.025* |
| **No. of variable values** | 0.003 | 0.000** | 0.002 | 0.000** | -0.003 | 0.000** | -0.005 | 0.000** |
| **Constant** | 0.108 | 0.000** | 0.135 | 0.000** | 0.106 | 0.000** | 0.129 | 0.000** |
| **N** | 342 | | 342 | | 194 | | 194 | |
| **Adjusted R-Squared** | 0.349 | | 0.286 | | 0.416 | | 0.563 | |
| **Root Mean Square Error** | 0.053 | | 0.066 | | 0.066 | | 0.083 | |

*p<0.05, **p<0.01

The R-BS models showed that the differences between onliners and offliners were captured the most prominently in *internet*, but also in *household and family, environment,* and *multiculturalism* (the last one only in the unweighted data). The topics with below-average differences were *finance* (in contrast to the Cramer's V model), *politics and elections,* and *government and policy.* Except for the *internet* topic (and to some extent *international relations*), there were no observable trends – in some cases,

weighting decreased bias in others it had no effect; effect sizes differed substantially between Cramer's V and R-BS models for the same topics; topics with above and below-average effect sizes could not be grouped further into broader homogenous topics with more or less undercoverage bias. Thus, based on the evidence presented in Table 6.2, hypothesis H1 cannot be rejected.

To address RQ3, weighted and unweighted estimates of the differences between onliners and offliners are presented. The results show that post-stratification weighting reduced some of the differences between onliners and offliners, which is consistent with some literature on this topic (see Rookey et al. 2008). After post-stratification weighting, both the Cramer's V coefficients for topics and mean Rank Biserial coefficients for topics were decreased (see constants and coefficients), but most of the magnitude of the effect size remained. Nevertheless, on average, the differences between onliners and offliners were small (see the interpretation of effect sizes in Cohen 1988, pp. 79-81). Moreover, the effect of weighting on the decreased magnitude of differences can be observed for *attributes* as a type of question content. This should come as no surprise since *attributes* are, generally speaking, other "non-weighting" socio-demographic or factual information about respondents and are associated with primary socio-demographics used in post-stratification weighting. As no other type of question content category stood out as a predictor of differences in the weighted models, it can be concluded that the differences between onliners and offliners, when controlling for primary demographics, are fairly stable across question content.

On the other hand, the differences measured with *binary variables* were smaller than those measured with *ordinal variables* (the reference category) in the Cramer's V models, and the differences measured with *continuous variables* were greater than those measured with *ordinal variables* in the R-BS Coefficient models. Moreover, the *number of variable values* had a statistically significant effect in all four models. There might be methodological explanations for this finding, such as measurement mode effects or potentially inconsistent effect size measures affected by particular variable characteristics like the number of values/categories. Mode effects as differences in responding are a result of different factors, such as interviewer administration. In the case of binary variables, the difference might be smaller due to acquiescence, i.e., tendency to agree with the interviewer. At the same time, the measures of the magnitude of effect size (Cramer's V, R-BS Coefficient) might be more dependent on the number of categories/ranges of continuous variables (see the coefficient for *No. of variable categories)* than theory suggests (see Cohen 1988; Glass 1965). We make this argument assuming that the magnitude of differences between onliners and offliners is consistent across different types of questions and variables. These results indicate that regression modeling and controlling for variable characteristics, in contrast to techniques such as carrying out ANOVA, provide more robust results.

### 6.3.3 Accuracy of estimation – benchmarking

Finally, benchmarking was performed to establish how the observed differences between onliners and offliners affected the accuracy of estimates relative to the nationally representative benchmarks (see Table 6.3). Our focus was on the comparison of the Life in Australia™ online-offline and online-only samples. Additionally, different probability-based samples from the OPBS 2015 study were compared in an attempt to generalize the current finding to the population that cannot, refuses to, or prefers not to respond online (so-called offliners). With this benchmarking analysis, the aim was to address RQ4 and test hypothesis H3.

While the majority of benchmarks, including weighting benchmarks, were updated to the Australian Census 2016 figures, our findings on the accuracy of different probability-based samples did not differ substantially from the findings of Pennay et al. (2018) and Kaczmirek et al. (2019) who used the 2011 Australian Census benchmarks. However, the primary focus here was on the comparison of the accuracy of estimates if the offline population was completely excluded. Firstly, the results indicated that the Life in Australia™ estimates for 18 items from OPBS would differ very little if no offliners were included. The average absolute relative bias was 0.028 (2.8%), which is about half that of the average for the items from Life in Australia™ that were analyzed in the first part of this paper (see Table 6.1, far right column); this should be attributed to the introduction of post-stratification weighting in this benchmarking analysis. Secondly, the average absolute mean bias differed by less than 1 percentage point (0.89), and since the difference in the average absolute error from the mean between the combined and online-only samples was equal to less than the average absolute mean bias (0.33=5.74-5.41), it can be concluded that excluding offliners can, in some cases, even improve derived estimates relative to benchmarks. This can be confirmed by reviewing the estimates for the following items, *Australian citizen, couple with dependent children, enrolled to vote, not Indigenous, living at last address 5 years ago, general health status (very good), life satisfaction (8 out of 10), and psychological distress, Kessler 6 (low),* whereby the online-only estimates are closer to the benchmarks, albeit with only a very small improvement if excluding offliners. Similar results were obtained when comparing A-BS and RDD Piggybacking combined and online-only estimates, whereby the absolute mean bias and absolute relative mean bias were much larger, possibly because of larger proportions of offline respondents in those surveys (RDD piggybacking 268/560=47.9%, A-BS 330/538=61.3%).

Table 6.3: Benchmarking results, comparing accuracy relative to the benchmark with and without the offline population, weighted estimates

| | | **Life in Australia™** | | **A-BS** | | **RDD piggybacking** | |
|---|---|---|---|---|---|---|---|
| Survey item | Benchmark | Online+offline (n=2,580) | Online only (n=2,166) | Online+offline (n=538) | Online only (n=208) | Online+offline (n=560) | Online only (n=292) |
| Australian citizen | **87.12** | 0.53 | 0.41 | 2.97 | 4.19 | -0.67 | 2.60 |
| Couple with dependent children | **38.35** | -11.16 | -10.70 | -11.11 | -11.50 | -12.53 | -15.81 |
| Currently employed | **61.61** | 5.38 | 6.69 | 5.38 | 3.06 | 6.78 | 6.26 |
| Enrolled to vote | **78.47** | 7.21 | 7.10 | 7.80 | 7.54 | 4.53 | 10.03 |
| Home ownership with a mortgage | **28.82** | 2.22 | 2.68 | 9.28 | 11.62 | 7.69 | 10.00 |
| Not Indigenous | **97.73** | -0.23 | 0.09 | 0.62 | 1.09 | 0.51 | 2.21 |
| Language other than English (speak only English) | **76.50** | 8.62 | 8.69 | 4.56 | 0.53 | 7.73 | 5.13 |
| Living at last address 5 years ago | **56.85** | 1.50 | 0.50 | -0.96 | -9.25 | -1.40 | -1.97 |
| Most disadvantaged quintile for area-based SES | **20.00** | -6.71 | -7.77 | -5.51 | -9.01 | -9.59 | -6.05 |
| Resident of a major city | **66.80** | 4.15 | 5.08 | 8.17 | 12.34 | 3.95 | 4.92 |
| Voluntary work (none) | **79.39** | -17.07 | -17.08 | -18.35 | -17.04 | -17.74 | -12.40 |
| Wage and salary income $1000–1249 per week | **13.80** | -1.64 | -2.22 | 0.06 | 0.54 | -0.07 | 5.64 |
| Consumed alcohol in last 12 months | **81.87** | 3.62 | 5.03 | 4.24 | 4.59 | 1.96 | 3.47 |
| Daily smoker | **13.52** | -1.97 | -3.47 | -4.03 | -7.11 | 2.88 | 2.63 |
| General health status (very good) | **36.20** | -2.96 | -1.49 | 2.07 | 10.71 | -2.08 | 11.09 |
| Life satisfaction (8 out of 10) | **32.60** | -1.24 | -0.73 | -3.43 | -2.92 | -1.68 | -1.60 |
| Has private health insurance | **57.10** | 3.43 | 7.03 | 4.08 | 5.95 | 2.17 | -0.30 |
| Psychological distress, Kessler 6 (low) | **82.20** | -17.76 | -16.65 | -13.92 | -19.14 | -11.75 | -9.61 |
| Average absolute mean bias (online+offline and online only) | | 0.89 | | 2.69 | | 3.05 | |
| Average absolute relative mean bias (online+offline and online only) | | 0.028 | | 0.076 | | 0.095 | |
| Average absolute error from the mean (combined) | | 5.41 | 5.74 | 5.92 | 7.67 | 5.32 | 6.21 |
| Average absolute error from the mean (secondary demographics) | | 5.53 | 5.75 | 6.23 | 7.31 | 6.10 | 6.92 |
| Average absolute error from the mean (substantive items) | | 5.16 | 5.73 | 5.30 | 8.40 | 3.75 | 4.78 |

*p<0.05, **p<0.01

Despite observing differences in the average absolute errors from the mean between samples with or without offliners, none of those differences tested with bootstrapping were statistically significant at $p<0.05$. Hence, H3 is rejected and it can be concluded that including the offline population does not improve the quality of estimates in the analyzed surveys, even if the portion of offliners is much higher than in the Life in Australia™ data. It should be noted that OPBS 2015 respondents were not encouraged/'web-pushed' to participate online in the same way as Life in Australia™ respondents, and the offline sample in those surveys consisted of "traditional" offliners as well as respondents who could respond online but only preferred to respond via telephone or mail at that particular time. When comparing all average absolute errors (combined, secondary demography, substantive items) and CIs estimated using bootstrap for all samples, there were only statistically significant differences between the Life in Australia™ online-offline respondents and A-BS online-only respondents (at $p<0.05$, A-BS less accurate).

## 6.4 Discussion and recommendations

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

Making a decision on (temporarily) excluding the offline population is a multi-dimensional problem. One could argue that the offline population should be included no matter the costs due to the offline population being fundamentally different to the online population; this has been supported by evidence from multiple studies (Eckman 2016; Keeter et al. 2015; Rookey et al. 2008). Similarly, the undercoverage bias analysis described here revealed statistically significant bias for more than half of all studied variables from all surveys (not rejecting hypothesis H1). Yet, the magnitude of differences between the populations, as well as the size of the offline population, should be a factor in the decision making, as the effect of undercoverage is a function of these two dimensions. With statistically significant but relatively small differences, and with a small proportion of offline respondents in the general population (in countries with high internet penetration rates and high-level internet literacy), there might be a much less significant effect of undercoverage than one would expect. Last but not least, the differences between onliners and offliners do not always work against the accuracy of probability-based online panels (or mixed-mode research in general). Based on the evidence

presented in this study, exclusion of the offline population generally does not affect the derived estimates much and can even push them closer to the nationally representative benchmark as the best approximation of the truth.

The findings of this research are based on data from one country only (Australia) and country-specific effects cannot be ruled out. The results indicate that inclusion of the offline population in probability-based online panel research seems to be, to some extent, unnecessary from the coverage error and accuracy perspectives. This could potentially be generalized to other developed countries with high internet penetration rates, narrower socio-economic and demographic distributions, and consequently, relatively minor non-factual differences between those with and without internet connection. At the very least, offliners could be temporarily excluded for certain topics which the current study identified as lesser predictors of differences between the populations, such as *household and family, government and policy,* or partially, *finance*. On the other hand, it might be more prudent to think reversely - what items should never be included in probability-based online panel surveys if data are collected from an online sample only, e.g., *internet* or *international relations* items in Life in Australia™. However, overall, the current study observed differences across the majority of topics with no particular trends, and thus, could not reject hypothesis H2. This is in line with the findings of Rookey et al. (2008) and other authors who have reported differences for different topics (Blom et al. 2015; Bosnjak et al. 2013; Eckman 2016; Keeter et al. 2015; Zhang et al. 2009). The observed bias might well be a result of a combination of fundamental differences between the populations (potential undercoverage bias), differential nonresponse in panel studies over time, and measurement mode effects.

Moreover, the evidence from this study suggests that while differences between onliners and offliners are present in probability-based mixed-mode research in Australia, any negative impacts on data accuracy should be minimal for the majority of topics, question contents, and variable types, even relative to the nationally representative benchmarks. The findings of this study led to rejection of hypothesis H3, which predicted that online-offline probability-based online panel samples would produce more accurate estimates compared to online-only samples. In the future, it would be worth exploring if undercoverage bias and its effect on survey estimates (benchmarking) decrease at the bivariate or multivariate level, as previously reported by Eckman (2016) for probability-based online panels and by Biddle et al. (2018) for opt-in panels.

The current analyses were limited, to some extent, by the number of studied items and their characteristics. With a larger sample of items and variables with available benchmarks, possibly from questions related to different broad topics and with more continuous variables, future studies would have greater statistical power and better evidence for data-informed decision making. The current

findings would have to be adjusted in that case. This study presents a combined approach to studying undercoverage bias and its effects on data accuracy, and as this was examined in the Australian context only, future research should focus on online-offline population differences in other countries. This is particularly pertinent in regions with both lower internet penetration rates and wider socio-economic and demographic distributions. Such studies could help establish how necessary mixed modes and inclusion of the offline population is in a particular country's context.

## 6.5 References


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

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

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

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

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

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

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

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

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

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

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

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

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

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

Eurostat. (2020, September). *Digital economy and society statistics - households and individuals.* https://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_households_and_individuals

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

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

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

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

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

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

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

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

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

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* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.4225/87/FSOYQI

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

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

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

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

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

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

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

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

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

## Appendix 6

Table 6.4: Life in Australia™ survey data collected for the ANU (used in the undercoverage bias part of the study)

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

Table 6.5: Data files used in the accuracy estimation part of the study

| Title of the survey | Month and year | Subsample sizes | Data collection | Data DOI |
|---|---|---|---|---|
| Online Panels Benchmarking Study 2015 | June 2015 | n=4,757 (1,699 prob, 3,058 non-prob) | Online, telephone, mail | 10.4225/87/FSOYQI |
| Health, Wellbeing and Technology Survey (OPBS Replication) 2017 *(Life in Australia™ Wave 2) Wave 2)* | January-February 2017 | n=2,580 | Online, telephone | 10.26193/YF8AF1 |

Table 6.6: Subsamples in the benchmarking component of this study

| Study | Sampling | Survey | Sample by data collection mode | n |
|---|---|---|---|---|
| OPBS Replication 2017 | Probability-based | Life in Australia™ Wave 2 | online | 2,166 |
| | | Life in Australia™ Wave 2 | offline (telephone) | 414 |
| OPBS 2015 | Probability-based | Address-based sampling | online | 208 |
| | | Address-based sampling | offline (telephone and mail) | 330 |
| OPBS 2015 | Probability-based | RDD “piggybacking” | online | 292 |
| | | RDD “piggybacking” | offline (telephone and mail) | 268 |

Table 6.7: Benchmarking data sources and nationally representative benchmarks

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

# Chapter 7 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

## 7.1 Introduction

### 7.1.1 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, p. 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, p. 175). In probability-based online panel research, panel organizations 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 (Jans 2008). That is one of the reasons why certain panel organizations 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, pp. 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 (Jans 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[54]), 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, n.d.). 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 2013, pp. 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

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

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 (Dillman et al. 2014; Tourangeau et al. 2000). 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).

Mode-related measurement errors are present in different ways, such as acquiescence response bias, social desirability and satisficing (Groves et al. 2009; Jans 2008). 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, p. 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 (Baker et al. 2010; Krosnick et al. 1996). The direction of biases from these mode effects are more difficult to predict *a priori*.

### 7.1.2 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 nonprobability 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.

### 7.1.3 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 (Beulens 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 & van der 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 modeling (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 (Beulens 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 Loosveldt (2013) 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 & Loosveldt 2013, pp. 99-101).

### 7.1.4 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 (Funk et al. 2011; King & Nielsen 2019; Rosenbaum 2020). More recently, more classic matching methods for casual inference 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 favor of alternative methods, such as Mahalanobis distance matching or coarsened exact matching (King & Nielsen 2019). Besides matching methods based on modeling, 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 2019). 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 (Iacus et al. 2012).

### 7.1.5 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 organization 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. (2010). 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 matching methods, 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 matching methods, 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.

## 7.2 Methods

### 7.2.1 Data

We will analyze the Life in Australia™ Wave 2 (2017) (Pennay & Neiger 2020) 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 analyze 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 are 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.

### 7.2.2 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 (A-BS) 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 2016) (online, telephone, mail modes).

2. Standalone dual-frame random digit dialing (DFRDD) CATI Survey.
3. Recruitment at the end of an established DFRDD survey/piggyback recruitment (online, telephone, mail modes).

The A-BS survey and the survey using piggyback recruitment allowed mixed modes of completion. For the A-BS 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 (The American Association for Public Opinion Research 2016)) were 12.4% for the DFRDD piggyback sample, 17.9% for the standalone DFRDD sample and 26.5% for the A-BS 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 7.1.

Table 7.1: 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 dialing

While those samples differ based on the sampling approach applied (Sample origin in Table 7.1), we combined samples based on the mode used (Mode in Table 7.1), since all of the surveys were probability-based. For example, to identify mode effects, CATI mode respondents from A-BS,

standalone DFRDD CATI, DFRDD piggybacking, and DFRDD-recruited panel (Wave 2) samples will be compared to the PAPI respondents from A-BS and DFRDD piggybacking. For more detail, see Table 7.2.

Table 7.2: Subsamples by mode combined for this matching and mode effects analysis

| Sample by mode | Source | Sample origin |
|---|---|---|
| CAWI (n=2,666) | Wave 2 (n=2,166) | 1 |
| | OPBS A-BS (n=208) | 2 |
| | OPBS DFRDD/piggybacked (n=292) | 4 |
| CATI (n=1,371) | Wave 2 (n=414) | 1 |
| | OPBS A-BS (n=128) | 2 |
| | OPBS DFRDD/standalone (n=601) | 3 |
| | OPBS DFRDD/piggybacked (n=228) | 4 |
| PAPI (n=242) | OPBS A-BS (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 dialing; OPBS = Online Panels Benchmarking Study 2015; PAPI = paper self-administered mode

By combining two studies and four samples into three targeted subsamples[55], we increased the statistical power as well as enable mode effects analysis for three distinctive modes.

### 7.2.3 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 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 (2019) 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. Further technical details about post-survey adjustments, including propensity score calculation, Mahalanobis distance matching and coarsened exact matching, are provided in Chapter 9, section 9.2.2.

*1. Socio-demographic controls in regression models without matching*

This approach is similar to poststratification weighting and MM Calibration (see Vannieuwenhuyze & Loosveldt 2013), and it was conducted to identify differences in distributions of response variables as

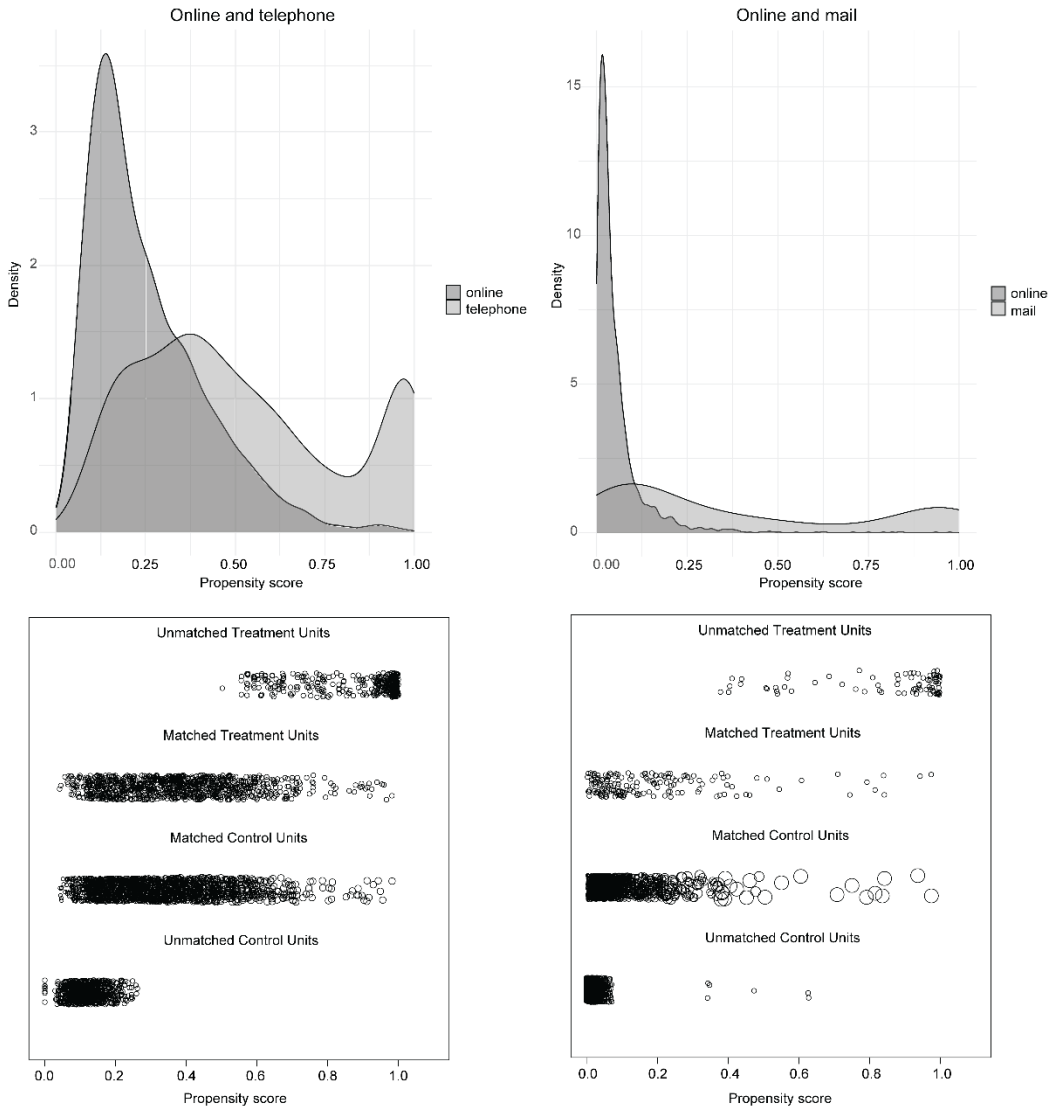
[55] Potential temporal effects due to the time gap in data collection, as well as sample composition effect (in OPBS studies), were controlled by including *sample source* variable (see samples in Table 7.2) as a predictor/control in all regression models (for more information, see Subsection 7.2.4 Data Analysis and Table 7.6 in the Appendix 7).

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 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 matching (King & Nielsen 2019).

*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 7.2). The matching results in Figure 7.1 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 7.1: 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 (colored blue in Table 7.6), 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 7.3 summarises the matching approaches and results. We purposely tried to optimise the 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 7.3: Matching parameters and sample sizes by matching method

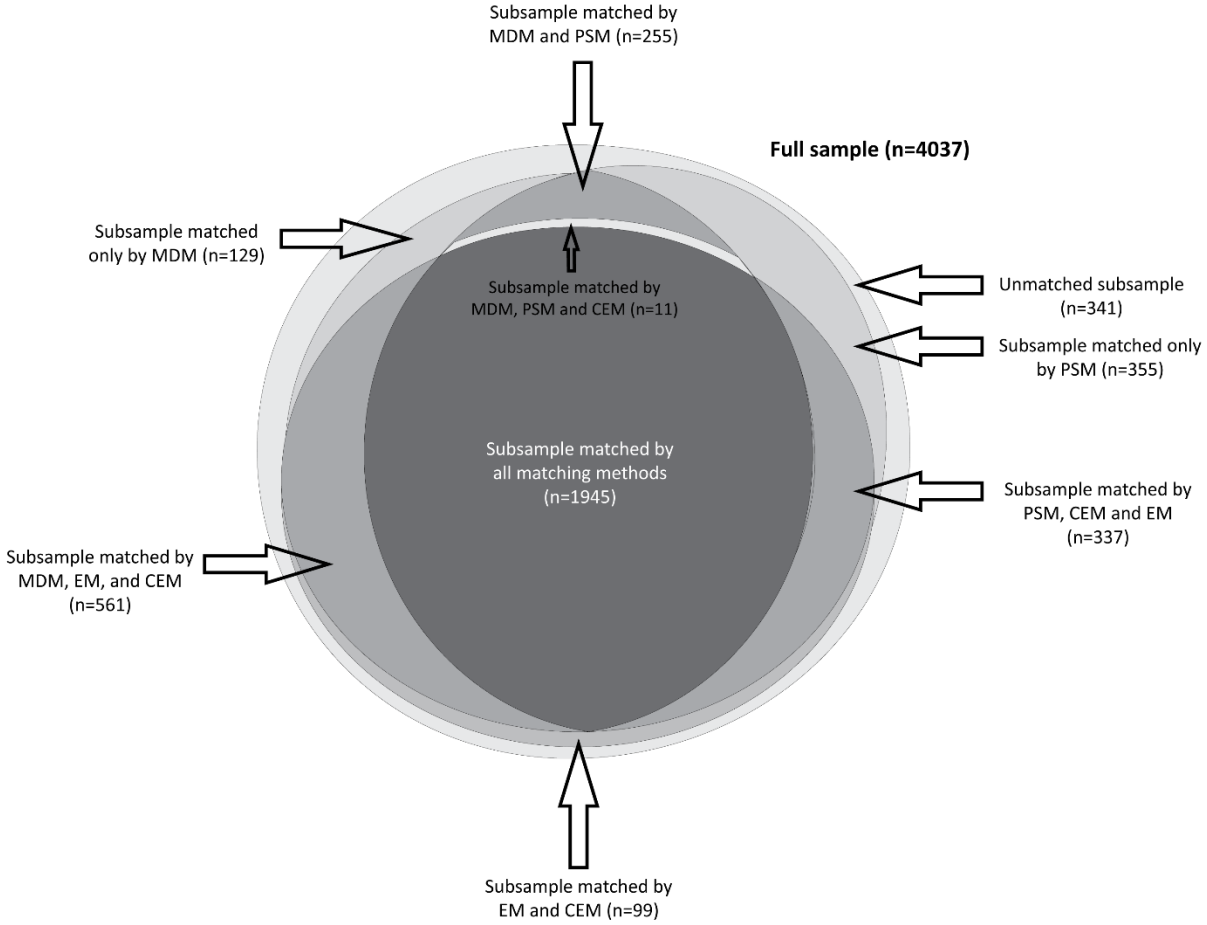
| Matching method *(R package used)* | Online-telephone samples matching | | | | Online-mail samples matching | | | |
|---|---|---|---|---|---|---|---|---|
| | **Original sample size** | **Matching parameters** | **Matched sample size** | **Average SMD[56]** | **Original sample size** | **Matching parameters** | **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

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 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 favor of CATI, which is apparent from Figure 7.1 as well.

[56] 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 7.2.6).

Figure 7.2: Matching results for the online-telephone sample




In Figure 7.2 we present the results of 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.

### 7.2.4 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 7.3, 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 modeling 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 matching and thus sample size reduction can lead to a loss of statistical power.

### 7.2.5 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, pp. 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, pp. 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 7.6 in the Appendix 7). 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).

### 7.2.6 Selection of items as controls in 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 A-BS 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[57]), 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 (2019).

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 behavioral, 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

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

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 matching. See Table 7.6 for more information – PSM and MDM variables are colored 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 7.6 in the Appendix 7 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.

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.

## 7.3 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.

### 7.3.1 Item 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 7.5 in the Appendix 7). 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 7.4).

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 7.4). 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.

### 7.3.2 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 7.6). 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 was a total of 27 items (out of 35) with distributional differences identified by regression modeling 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 7.6). 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 modeling, controlling for socio-demographics (see Table 7.6). 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 behaviors than telephone respondents. Responding to an interviewer might be related to underreporting of particular harmful behaviors 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 (also see Chapter 8), 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 (also see Chapter 6), 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 using matching methods for casual inference.

### 7.3.3 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.

### 7.3.4 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 7.2).

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.

### 7.3.5 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 7.4 and 7.5, 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 change significantly, *ceteris paribus*.

### 7.3.6 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 (colored green in Table 7.6). 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 7.6 for all coefficients).

## 7.4 Discussion and recommendations

Mixed-mode surveys are increasingly common and seem to be the standard in probability-based online panel research. Panel organizations providing measurement equivalence like ELIPSS panel (e.g., tablets for all panellists) are more of an exception than not (see Kaczmirek et al. 2019, pp. 4–5). The evidence from this research, as well as from the study on longitudinal panel mode effects (see Chapter 8), 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 matching methods 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 Chapter 8), 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 matching methods, 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. Using matching methods for casual inference 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 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, pp. 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 utilize 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 matching methods. 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.

## 7.5 References


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

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

Australian Communications and Media Authority. (n.d.). *Mobile-only Australia: living without a fixed line at home*. Retrieved January 10, 2020, from https://www.acma.gov.au/publications/2020-12/report/mobile-only-australia-living-without-fixed-line-home

Australian Government. (2016, February 22). *PSMA Geocoded National Address File (G-NAF)*. https://data.gov.au/data/dataset/19432f89-dc3a-4ef3-b943-5326ef1dbecc

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

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.

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

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

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

Cernat, A., Couper, M. P., & Ofstedal, M. B. (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, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research*, *36*(1), 131-148.

De Leeuw, E., & van der Zouwen, J. (1988). Data quality in telephone and face-to-face surveys: A comparative analysis. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283–299). John Wiley & Sons.

De Leeuw, E. 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. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 5959–5967.

De Leeuw, E. D., 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 P. J. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. De Leeuw, & E. West (Eds.), *Experimental methods in survey research: techniques that combine random sampling with random assignment* (pp. 151–165). John Wiley & Sons.

Dennis, J. M., 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. *60th Annual Conference of the American Association for Public Opinion Research*, 1-26.

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

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons.

DiSogra, C., Cobb, C., Chan, E., & Dennis, J. M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings, 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), 615-639.

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

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

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

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

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

Jans, M. (2008). Mode effects. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 475-480). Sage.

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

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

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

Kennedy, C., Ackermann, A., Turakhia, C., Emerson, M., & James, A. (2012, May 17-20). *Mode effects measurement and correction: a case study* [Conference presentation]. 67th Annual Conference of the American Association for Public Opinion Research, Orlando, United States of America.

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.

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, W. R. (1996). Satisficing in surveys: Initial evidence. *New directions for evaluation, 1996*(70), 29-44.

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* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.4225/87/FSOYQI

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

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

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R. (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, August 18). *Matching Methods for Causal Inference: A Machine Learning Update.* https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/

Suzer-Gurtekin, Z. T., Valliant, R., Heeringa, S. G., & De Leeuw, E. D. (2018). Mixed-mode surveys: design, estimation, and adjustment methods. In T. P. Johnson, B. E. Pennell, I. A. L. Stoop, & B. Dorer

(Eds.), *Advances in comparative survey methods: multinational, multiregional, and multicultural contexts (3MC)* (pp. 409-430). John Wiley & Sons.

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

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

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

## Appendix 7

Table 7.4: 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 7.5: 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 7.6: 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 |
| | No | RC | | | | | | | | | | |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a2_2 - internet connection (dial-up, ISDN) | Yes | | 0.04 | 0.44 | -0.12 | -0.74 | -0.05 | 0.95 | 0.47 | 1.67 | 0.48 | 1.64 |
| a2_3 - internet connection (mobile device) | No | RC | | | | | | | | | | |
| | 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-dependent child(ren) | | 0.45† | 0.11 | 0.17 | 0.57 | 0.90* | -0.44 | 0.32 | 0.33 | 0.32 | 0.36 |
| | 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

# Chapter 8 Panel mixed-mode effects: does switching modes in probability-based online panels influence measurement error?

## 8.1 Introduction

### 8.1.1 Mixing modes in online panel research

Organizations managing nonprobability-based panels tend to use only the online mode to collect data from their panellists, whereas most probability-based online panels try to allow for the fact that internet penetration is still not close to 100% (Baker et al. 2010). The aim of probability-based panels is to represent the general population, and people without computer or internet access should be included in the panel.

This comes with a cost, and the balance between measurement equivalence and coverage is important (Blom et al. 2016). One way to reduce noncoverage bias is to provide respondents with computer hardware and internet access. An alternative is to collect data using a mix of modes, including interviewer-administered modes (Baker et al. 2010). Blom et al. (2016) listed four European probability-based panels which included the offline population. Different organizations managing probability-based online panels use different approaches to find the right balance between measurement errors, representation errors and costs. The French ELIPSS Panel focuses more on measurement equivalence by subjecting panellists to the same stimulus; consequently, all of them receive a tablet computer with internet access to fill out online questionnaires. On the other hand, the German GESIS panel is a mixed-mode panel, with the offline population participating via mailed paper questionnaires. The trade-off is that this does not guarantee measurement equivalence. More equivalence is offered by the Dutch LISS Panel and the German Internet Panel (GIP), but still less than in the case of the ELIPSS panel, because of different devices and browsers being used – only the offline population receives tablets (Blom et al. 2016).

Life in Australia™ is, similar to the GESIS panel, a mixed-mode panel. However, it uses a different offline mode – interviewer-administered telephone mode. To reduce representation and response bias, some respondents in Life in Australia™ end up being part of both online and offline populations during the panel lifecycle. Some offliners do not provide an email address at recruitment, but provide one in later waves, which means that they start by participating offline and later switch to responding online. Also, onliners are first contacted and reminded via email, then text message, but might be contacted over the phone later if they do not respond online after a certain time; a small percentage of them even respond over the phone (SRC 2018). Similar changes between online and offline populations happened in the LISS, GIP and ELIPSS panels. Those panels include between 7% and 10%

of online panellists who were previously offline (Blom et al. 2016), which should not come as a surprise because the stability of mode preferences of longitudinal respondents is fairly low (Baghal & Kelley 2016).

### 8.1.2 Measurement mode effects in mixed-mode panel research

In mixed-mode probability-based online panels, improvements in representativeness may come at the cost of measurement error. Generally speaking, there are two specific hypotheses about the possible impact of shifting from one mode to another: social desirability response bias and satisficing. Social desirability is the tendency of certain respondents to report more socially desirable, acceptable answers or those in sync with the popular opinion, rather than choosing answers reflecting their true feelings or thoughts (Grimm 2010). It is a consequence of two separate factors: self-deception and other-deception (Nederhof 1985). Satisficing occurs when a respondent generates valid but not necessarily accurate or thoughtful answers to survey questions by decreasing their cognitive effort. Although social desirability should be a more significant issue in the case of interviewer surveys, self-administration might have the potential for a higher incidence of satisficing because of the ease of responding (Baker et al. 2010).

Mixing interviewer-administered and self-administered modes should, therefore, result in an increased measurement mode effect bias. In a longitudinal design, this may cause measurement inequivalence both between waves and between different modes (Cernat 2015), because change cannot be measured accurately if the respondent is presented with the same question stimulus in each wave (Dillman 2009). It has been argued (De Leeuw 2005) that mode effects should be an important survey design consideration and should be reduced as much as possible, and, in a longitudinal design, mode experiments should be carried out to help adjust for measurement mode effects.

The concept of social desirability is made up of four nested characteristics, from large scale to small scale: cultural characteristics, personality characteristics, data collection mode and item characteristics. It is, therefore, important to control for question wording and mode of data collection, particularly in mixed-mode or cross-cultural research (Callegaro 2008). Social desirability is more prevalent in survey modes allowing respondent identification, and in data collection in the presence of other people, and is related to questions on widely accepted attitudes, and behavioral and social norms (Grimm 2010).

Social desirability response bias can also be observed in interviewed-administered surveys measuring satisfaction, with respondents reporting higher satisfaction with their jobs (Kim & Kim 2016), products in market research (Albert & Tullis 2013), family, social life, health and financial troubles (Keeter et al.

2015), and democracy. Those interviewed on the phone also reported more trust in the ruling party (Zimbalist 2017) and more positive ratings of political figures (Keeter et al. 2015).

The main reason for satisficing is that some respondents tend to make the task of responding as easy as they can. That leads to using ranges or rounding values, making ratings following a few simple principles or not considering questions seriously (Tourangeau et al. 2000, p. 254). Satisficing, which tends to have a higher incidence in self-administered surveys, can be observed in different ways: non-differentiation, non-substantive responses, rapid completion (speeding), and response-order effects, such as primacy and recency (Baker et al. 2010). Primacy is a tendency to choose the first-offered answers, and recency is a tendency to choose among last categories regardless of the content (Dillman et al. 2014, pp. 104–105). It is expected that computer administration will yield the opposite response-order effect (primacy) to oral administration (recency), and so will give different distributions of responses (Baker et al. 2010).

### 8.1.3 Panel conditioning, panel fatigue and (in)stability of responses over time

Compared with cross-sectional research, panel research has more sources of measurement bias, such as panel conditioning and panel fatigue, which are not only interesting in themselves but should be considered when analysing measurement mode effects specific to panels. Panel conditioning occurs when a sample unit's response is influenced by prior survey participation or contacts, and introduces so-called 'time-in-sample' bias. Because it affects responses in future waves, the estimates for certain supposedly stable concepts vary significantly over time (Cantwell 2008). For example, in the study conducted by Halpern-Manners and Warren (2012), conditioned respondents reported lower unemployment rates and higher incidence of leaving the labor force than respondents who participated in the survey for the first time, even after controlling for attrition and mode effects. Panel fatigue, on the other hand, occurs when the quality of data from a particular respondent diminishes because they stay in the panel, providing data, for too long. It results in unit nonresponse, item nonresponse, satisficing, and other forms of lower-quality data (Lavrakas 2008).

These sources of measurement bias could be the reason for lower or higher stability of answers, or lower or higher quality of data in panel research over time. Although some authors report panel conditioning for a limited number of knowledge items only in online panel research (see Kruse et al. 2009), other authors report significant effects of panel conditioning in longitudinal research. For example, Cernat (2015) reported an effect of panel conditioning on stability, with the results showing that stability increases in time even if no mode differences are apparent. In Australia, Wooden and Li (2014) presented similar findings – repeated participation resulted in a clear and gradual reduction in

the dispersion of the target variables. Moreover, Sturgis et al. (2009) reported a reduction in the fraction of non-substantive answers over time as a form of panel conditioning.

### 8.1.4 Research questions

In this study, we answer the following research questions about the presence of respondent-level panel measurement mode effects after controlling for socio-demographic characteristics of panellists and panel conditioning:

1. *Is switching from interviewer-administered mode (telephone, offline) to self-administered mode (online) (or vice versa) in probability-based online panels associated with changes in answers over time?*
2. *Does switching from interviewer-administered mode (telephone, offline) to self-administered mode (online) (or vice versa) in probability-based online panels influence satisficing?*
3. *Does switching from self-administered mode (online) to interviewer-administered mode (telephone, offline) influence social desirability?*
4. *Does switching from self-administered mode (online) to interviewer-administered mode (telephone, offline) (or vice versa) influence item nonresponse?*

To answer these questions, the remainder of the paper is structured as follows: Section 8.2 of this paper presents the data, methods and survey items selected to study panel mode effects; Section 8.3 presents results of the measurement mode effect analysis; and Section 8.4 discusses the results and practical implications of respondents switching modes in online panel research.

## 8.2 Methods

### 8.2.1 Data

The data used in this study were collected for the Australian National University by the Social Research Centre using its probability-based panel known as Life in Australia™. Five of six datasets used in this study are from ANUPolls, quarterly surveys of Australian public opinion (Centre for Social Research and Methods, n.d.).

Socio-demographic variables were from Life in Australia™ paradata. Certain predictors, such as mode switching, were derived from panel participation variables from Life in Australia™ online panel paradata files.

### 8.2.2 Population, sample and data collection modes

The population in this research can be defined as 'Australian residents aged 18 years or more', and the results from the surveys are generalisable to the Australian population. The response rate for the establishment of Life in Australia™, calculated as the product of the recruitment rate and the profile rate, was 15.5% in 2016 (*n* = 3,322) and 12.2% in August 2018 (refreshment, n = 267). To undertake recruitment, a dual-frame random-digit dialing (RDD) sample design was employed, with a 30:70 (40:60 in pilot) split between landline and mobile phone sample frames in 2016; a single-fame RDD mobile sample design was employed in 2018. 'Last birthday' method was used to select potential panel members in landline frames and the phone answerers in the mobile sample, although only one person per household was invited to join the panel. To cover online panellists, the online web self-completion mode was used. To collect data from offline panellists, the telephone mode was used (Social Research Centre 2018).

Because we studied changes in responses over time, only those respondents participating in at least two waves (see Table 8.1) were included (n = 2,542). About 1% (wave 21 $\rightarrow$ wave 22) and about 3% (wave 1 $\rightarrow$ wave 3) of panellists changed the survey administration mode between the analyzed waves (see Table 8.2). Although we worked with a relatively large sample of Life in Australia™ respondents whose answers were potentially sensitive to panel conditioning, we ended up with a relatively small sample size of panellists who changed the mode of data collection (127 respondents, who switched modes a total of 172 times). That might negatively affect the reliability of results related to panel mode effects, because small samples often leave the null hypothesis unchallenged.

Table 8.1: Survey data used in this study

| Title of survey | Month and year | Wave |
|---|---|---|
| Australian Personas Survey, 2016 | December 2016 | 1 |
| ANU Poll 2017 Housing | March 2017 | 3 |
| ANU Poll 2017 Job Security | October 2017 | 10 |
| ANU Poll 2018 Populism | August 2018 | 19 |
| ANU Poll 2018 Data Governance | October 2018 | 21 |
| ANU Poll 2018 Population | November 2018 | 22 |

### 8.2.3 Data analysis

Measurement mode effects in cross-sectional mixed-mode studies are usually tested using binomial, ordinal and multinomial logistic regressions, partial proportional and proportional odds models, ordinary least squares models (OLS), or structural equation modeling (SEM) (Jäckle et al. 2010). In this study, because we worked with panel data, we used:

- binary logit regression (pooled)

- multinomial logistic regression (pooled)
- multiple linear regression (OLS, pooled)
- fixed- and random-effect panel logit regression
- fixed- and random-effect panel OLS regression.

To establish what panel data model should be used in controlling for unobserved heterogeneity, either fixed- or random-effect, we performed the Hausman test for endogeneity[58] (Hausman 1978) each time.

### 8.2.4 Selection of data items to investigate panel mode effects

Because Life in Australia™ is not primarily used to measure longitudinal changes over time, few items are repeated. In ANUPolls, four variables are included in each wave to measure change over time. The items that appeared in all six waves of Life in Australia™ online panel data used in this study were:

- satisfaction with the way Australia is heading ('satisfaction')
- the most important problem facing Australia ('1st problem')
- the second most important problem facing Australia ('2nd problem')
- party support in federal election for the House of Representatives ('party support').

### 8.2.5 Statistical models

Panel mode effects were investigated by studying changes in responses to the same questions over time, conditional to changes in survey administration modes for panellists over time, and controlling for the extent of panel conditioning and socio-demographic characteristics of panellists.

In our models, the derived dependent variables measuring changes in responses from the same respondents over time attempted to capture certain concepts described in the literature on measurement mode effects in cross-sectional studies – that is, those that can be observed with panel data:

- Changes in answers in some or all of the four substantive items (stability): logit regression with a binary dependent variable coded as 0 – no change, 1 – any change; multiple regression

[58] The Hausman test is a test for model misspecifications and deals with endogeneity, i.e., regressors correlating with the error term. It compares two different estimators of the model parameters. Under the null hypothesis, the preferred panel model is random effect (assuming there is no correlation between regressors and the error term), and the alternative hypothesis is the preferred model is fixed effect (assuming a correlation between regressors and the error term) (Hausman 1978).

analysis with a continuous dependent variable *number of answer changes in a wave* (range 0–4).

- Change from substantive to non-substantive answers and vice versa, all four substantive items (sensitivity): multiple regression analysis with a continuous dependent variable *number of changes between substantive and non-substantive answers in a particular wave* (range −4 to 4, where negative values represent increase in the total number of non-substantive answers).
- Change from any substantive answer to the first listed answer and vice versa, satisfaction and party support items (primacy effect): multinomial regression with a nominal dependent variable coded as 0 – no change, 1 – other answer to the first one, 2 – the first answer to other.
- Change from any substantive answer to the last listed answer and vice versa, satisfaction (recency effect): multinomial regression with a nominal dependent variable coded as 0 – no change, 1 – other answer to the last one, 2 – last answer to other.
- Increase of satisfaction, change from less popular answers to more popular answers (1st problem, party support), change from other answers to 'environment'[59] (1st problem), vice versa (social desirability): multinomial regression with a nominal dependent variable coded as 0 – no change, 1 – change to socially desirable answer, 2 – change to socially less desirable answer.

We also derived a number of regressors for our regression models. The following independent variables are included, with information about which model they are included in:

<u>Mode effects</u>

- Change of mode, binary regressor coded as 0 – no change, 1 – any change (panel mode effects), model with *any change of answers* as the dependent variable only.
- Change of mode, nominal regressor coded as 0 – no change, 1 – online to telephone, 2 – telephone to online (panel mode effects), all other models.

<u>Panel conditioning</u>

- Number of times a respondent was asked the same question before responding in a particular wave (the extent of panel conditioning) – we assume that panel conditioning will have a greater effect if certain questions are asked more times.

[59] Reporting 'environment' as the most (or the 2nd most) important problem facing the country could be a potentially socially desirable answer, at least in the Australian context.

- Time in months since previously asked the question (the extent of panel conditioning) – we assume that panel conditioning is less severe if the gap in time between asking the respondents the same question is greater; this item could also have an effect on the propensity of changing answers that cannot be consistent over time.

Panel fatigue

- The total number of waves a respondent participated in over the first 22 waves of Life in Australia™ data collection. It should be kept in mind that we also include non-ANU-based surveys in this calculation, which made up 16 of the first 22 waves of data collection. This control variable had to be added, because the literature reports that sensitivity of answers is related to item nonresponse, but so is panel fatigue, which is further associated with non-substantive responses, panel nonresponse and voluntary attrition (for more details, see Lavrakas 2008).

Demographic controls

- Age group
- Gender
- Education
- State
- Country-of-birth group.

## 8.3 Results

### 8.3.1 Descriptive statistics

First, we present the results of the descriptive analysis of the composition of the sample in this study (Table 8.2). We have calculated and added propensities (predictive margins from logistic regression models) for the panel phenomena empirically investigated in the next sections of this paper – propensity of particular socio-demographic subgroup to participate in a panel wave, propensity to change answers to the same questions over time, and propensity to change mode (online to telephone or vice versa) over time (holding other factors constant).

In the full Life in Australia™ sample, females, people between 55 and 75 years of age, the more educated (bachelor degree or higher and certificate/diploma/trade), people living in the Australian Capital Territory, Tasmania or South Australia, and Australian-born people are overrepresented compared with Australian Census 2016 results. The youngest (18–24 years of age), people living in New South Wales, and the least educated are the most underrepresented (Australian Bureau of Statistics 2016). In the seven waves of data that we use in this study, females, respondents 55 years

of age and older, people living in South Australia, the most educated, and Australian-born people have a higher propensity to participate in surveys. The propensity to answer questions consistently over time is higher for females and people with certificate/diploma/trade, and year 11 or less education levels. The propensities to change modes are very low and are fairly consistent across all socio-demographic groups.

Table 8.2: Distribution of socio-demographic variables and calculated propensities for participation, changing answers and changing modes between waves

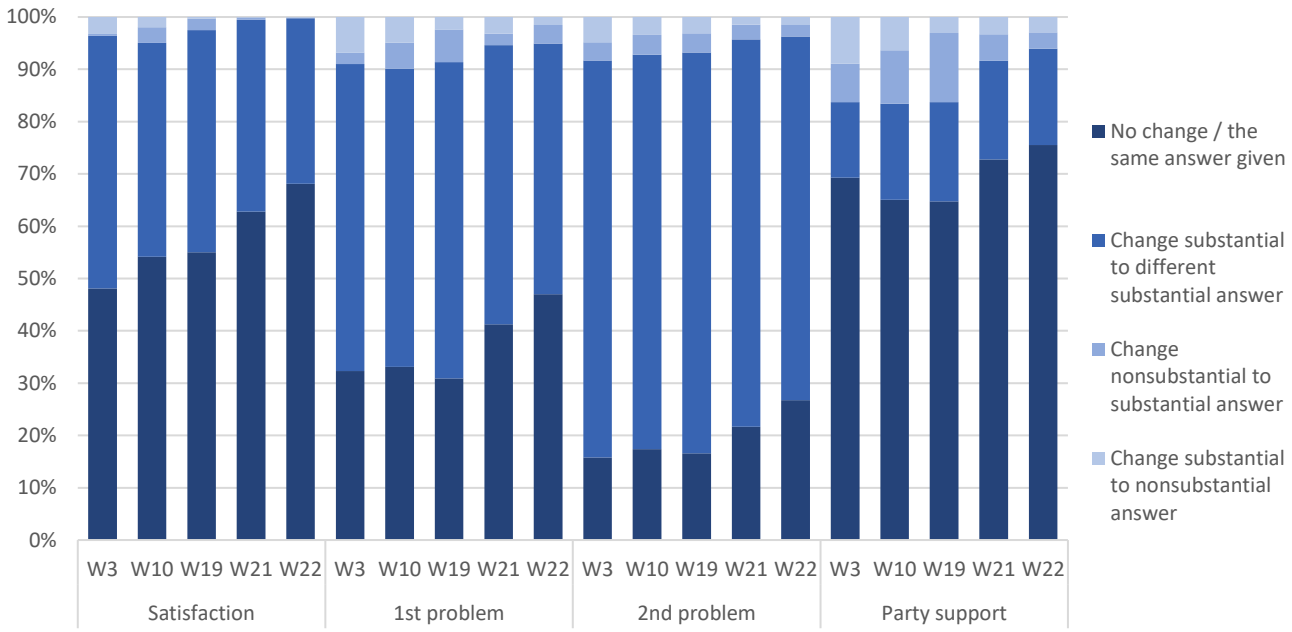
| Control variable | % | Propensity to participate in wave | Propensity to change answer* | Propensity to change mode |
|---|---|---|---|---|
| | | Margin (95% CI) | Margin (95% CI) | Margin (95% CI) |
| **Gender** | | | | |
| Male | 48.0% | 0.666 (0.657, 0.675) | 0.827 (0.816, 0.837) | 0.014 (0.010, 0.017) |
| Female | 52.0% | 0.693 (0.684, 0.701) | 0.854 (0.845, 0.863) | 0.019 (0.015, 0.022) |
| **Age group** | | | | |
| 18-24 years | 9.2% | 0.512 (0.488, 0.536) | 0.864 (0.837, 0.891) | 0.009 (0.002, 0.016) |
| 25-34 years | 14.9% | 0.609 (0.591, 0.626) | 0.814 (0.793, 0.835) | 0.015 (0.008, 0.022) |
| 35-44 years | 14.9% | 0.640 (0.622, 0.657) | 0.826 (0.806, 0.845) | 0.012 (0.006, 0.018) |
| 45-54 years | 17.9% | 0.684 (0.669, 0.699) | 0.840 (0.823, 0.856) | 0.018 (0.012, 0.023) |
| 55-64 years | 19.6% | 0.729 (0.715, 0.742) | 0.845 (0.830, 0.859) | 0.015 (0.010, 0.020) |
| 65-74 years | 16.0% | 0.787 (0.773, 0.801) | 0.856 (0.840, 0.871) | 0.022 (0.015, 0.028) |
| 75 or more years | 7.4% | 0.707 (0.684, 0.729) | 0.851 (0.827, 0.875) | 0.021 (0.011, 0.030) |
| **Education** | | | | |
| Bachelor or higher | 37.2% | 0.742 (0.732, 0.751) | 0.822 (0.810, 0.833) | 0.011 (0.008, 0.014) |
| Certificate/diploma/trade | 35.8% | 0.643 (0.632, 0.654) | 0.858 (0.846, 0.869) | 0.024 (0.019, 0.029) |
| Year 12 or equivalent | 12.3% | 0.682 (0.663, 0.700) | 0.841 (0.819, 0.862) | 0.017 (0.009, 0.025) |
| Year 11 or less | 14.8% | 0.605 (0.587, 0.623) | 0.858 (0.840, 0.876) | 0.012 (0.007, 0.017) |
| **State** | | | | |
| New South Wales | 30.0% | 0.663 (0.651, 0.675) | 0.838 (0.825, 0.851) | 0.015 (0.010, 0.019) |
| Victoria | 25.2% | 0.680 (0.667, 0.692) | 0.843 (0.829, 0.857) | 0.015 (0.010, 0.020) |
| Queensland | 19.4% | 0.688 (0.674, 0.702) | 0.852 (0.837, 0.867) | 0.015 (0.010, 0.020) |
| South Australia | 8.3% | 0.748 (0.727, 0.769) | 0.839 (0.816, 0.862) | 0.018 (0.010, 0.026) |
| Western Australia | 11.4% | 0.676 (0.657, 0.694) | 0.829 (0.807, 0.850) | 0.019 (0.011, 0.027) |
| Tasmania | 2.6% | 0.65 (0.609, 0.690) | 0.846 (0.802, 0.889) | 0.029 (0.009, 0.048) |
| Northern Territory | 1.0% | 0.585 (0.517, 0.653) | 0.888 (0.822, 0.953) | 0.061 (0.009, 0.112) |
| Australian Capital Territory | 2.3% | 0.673 (0.629, 0.716) | 0.816 (0.768, 0.863) | 0.017 (0.000, 0.034) |
| **Country of birth** | | | | |
| Australian born | 71.8% | 0.701 (0.693, 0.708) | 0.838 (0.830, 0.846) | 0.017 (0.014, 0.020) |
| Mainly NESB background | 16.1% | 0.604 (0.587, 0.621) | 0.856 (0.838, 0.873) | 0.015 (0.008, 0.022) |
| Mainly ESB background | 12.2% | 0.657 (0.638, 0.675) | 0.842 (0.822, 0.861) | 0.013 (0.007, 0.019) |

ESB = English-speaking background; NESB = non-English-speaking background; [a] the propensity to change answer to either of: satisfaction with the country heading, problem no. 1 in Australia, or party preference (problem no. 2 excluded) questions between consecutive waves completed

### 8.3.2 Stability of answers over time

Further to the propensity to change any answer in consecutive waves participated in, Figure 8.1 presents propensities for changing answers to any of the four specific questions repeated in ANU-commissioned Life in Australia™ surveys, over time. Different types of changes in answers to individual questions are presented as well.

Figure 8.1: Types of changes in answers from the same respondents over time




*W = wave*

The results show that party support is the item with the greatest consistency of answers over time, and 2nd problem is the item with the least consistency, which is why it was not included in the analysis presented in Table 8.2. Overall, there seem to be two factors associated with the propensity to change answers: the number of waves participated in and the time between waves participated in. The propensity for consistency generally increases over time and when the time between waves decreases. The same conclusion can be drawn for changes between substantive and non-substantive answers. The relationship between the factors is to be further explained using statistical modeling.

### 8.3.3 Panel mode effects controlled for panel conditioning

Because the focus of this research is on predictors of changes in answers over time, respondents who participated in at least two waves were included in the studied sample. In practice, this means that there was a certain level of panel conditioning present for all included respondents. The results are presented by the different types of measurement mode effects related to mixed-mode research and described in the literature:

- any change in answers (instability of answers)
- change between substantive and non-substantive answers (sensitivity)
- change from a substantive answer to the first answer (primacy)
- change from a substantive answer to the last answer (recency)
- change to potentially socially desirable answers (social desirability).

While satisficing and social desirability were controlled for by demographics and panel conditioning, sensitivity-associated item nonresponse was also controlled for by panel fatigue.

#### 8.3.3.1 Instability of answers

The analysis of the types of changes in answers from the same respondents over time, results of which are shown in Figure 8.1, are here extended with multivariate analysis of instability of answers over time as an indicator of general measurement mode effects and panel conditioning effects. The results of logit regression analysis (a pooled regression model) and dynamic logit regression analysis are presented in Tables 8.3 and 8.4. We carried out the Hausman test to look for a correlation between errors and regressors in the models, so we could choose between using fixed effects and random effects models. The results showed that fixed-effect models were the appropriate solution for controlling for unobserved time-invariant heterogeneity in all models except for the one with party support changes in answers as the dependent variable.

Table 8.3: Logit regression, random-effect and fixed-effect within-person logistic regression results; dependent variable: *any change of answers*

| **Substantive repeated item** | **Predictor of any change in answers over time** | **Logit regression model (pooled)** | | | | **Fixed-effect logit regression model** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| Satisfaction | Mode change (any) | 0.28 | -0.03 | 0.58 | 0.073† | 0.19 | -0.28 | 0.67 | 0.430 |
| | No. times question asked | -0.15 | -0.18 | -0.12 | <0.001** | -0.23 | -0.26 | -0.19 | <0.001** |
| | Months since question asked | 0.03 | 0.02 | 0.04 | <0.001** | 0.03 | 0.01 | 0.04 | <0.001** |
| 1st problem | Mode change (any) | 0.14 | -0.19 | 0.46 | 0.416 | -0.19 | -0.67 | 0.29 | 0.430 |
| | Times question asked | -0.09 | -0.12 | -0.06 | <0.001** | -0.16 | -0.19 | -0.12 | <0.001** |
| | Months since question asked | 0.05 | 0.03 | 0.06 | <0.001** | 0.05 | 0.04 | 0.07 | <0.001** |
| 2nd problem | Mode change (any) | -0.08 | -0.48 | 0.32 | 0.696 | -0.23 | -0.81 | 0.35 | 0.440 |
| | Times question asked | -0.11 | -0.15 | -0.08 | <0.001** | -0.17 | -0.21 | -0.12 | <0.001** |
| | Months since question asked | 0.04 | 0.03 | 0.05 | <0.001** | 0.04 | 0.02 | 0.05 | <0.001** |

| | | | | | | Random-effect logit regression model | | | |
|---|---|---|---|---|---|---|---|---|---|
| Party support | Mode change (any) | 0.26 | -0.06 | 0.57 | 0.108 | 0.21 | -0.21 | 0.63 | 0.325 |
| | Times question asked | -0.06 | -0.09 | -0.03 | <0.001** | -0.09 | -0.12 | -0.05 | <0.001** |
| | Months since question asked | 0.05 | 0.04 | 0.06 | <0.001** | 0.06 | 0.05 | 0.08 | <0.001** |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level, †significant at the 0.1 level; socio-demographic controls in the models: gender, age groups, education, state, country of birth

Regression analysis shows that mode changes, either online → telephone or telephone → online, are not predictors of changes in answers in any of the eight models for four response variables. For items with the lowest propensity for changing answers – satisfaction and party support – mode change predictors are almost statistically significant in the pooled models (satisfaction at $p<0.1$, party support at $p<0.15$, see Table 8.3). Further, the results show that both indicators of panel conditioning in all models were statistically significant predictors of changes in answers, although there was very little difference between the coefficients of pooled and dynamic logit models. Panel conditioning generally affects the changes in two different ways: the more times a question is asked, the lower the probability of change; and the longer the gap (measured in months) between a question being asked and then repeated, the higher the probability of change. The changes in answers to the satisfaction question seem to be more affected by how many times the question was asked than the changes in answers to the party support question.

Table 8.4: Multiple linear regression and fixed-effect within-person regression results; dependent variable: *number of changes of answers in a particular wave*

| **Derived variable** | **Predictor of any changes in answers over time** | **Multiple linear regression model (pooled)** | | | | **Fixed-effect regression model** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| No. of any changes in answers | Mode change (any) | 0.15 | 0.01 | 0.30 | 0.039* | -0.01 | -0.19 | 0.16 | 0.881 |
| | No. times questions asked | -0.08 | -0.09 | -0.06 | <0.001** | -0.10 | -0.12 | -0.09 | <0.001** |
| | Months since questions asked | 0.03 | 0.03 | 0.04 | <0.001** | 0.03 | 0.03 | 0.04 | <0.001** |
| | Constant | 2.06 | 1.96 | 2.16 | <0.001** | 2.22 | 2.17 | 2.27 | <0.001** |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level, †significant at the 0.1 level; socio-demographic controls in the models: gender, age groups, education, state, country of birth

The results from Table 8.4 explain the relationship between changes in answers, mode changes and panel conditioning from a slightly different perspective. This time, the dependent variable in the models is the number of changes in answers between two consecutive waves (range 0–4). The results fully support the findings of modeling with individual survey items. However, this time, the mode

change is a statistically significant predictor of the number of changes in answers in the pooled OLS model – mode changes increase the propensity to change answers.

#### 8.3.3.2 Sensitivity

Sensitivity as a result of interviewer administration – the measurement mode effect concept that can result in item nonresponse or non-substantive answer selection – was studied with static and dynamic regression models. Multiple linear regression analysis results (a pooled regression model) and dynamic regression analysis results with a continuous dependent variable are presented in Table 8.5. Based on the results of the Hausman test, we decided to carry out fixed-effect modeling. In these two regression models, the predictor variable *any mode change* from the previous models is split into *online to telephone* and *telephone to online* mode changes. Also, the *number of changes between substantive and non-substantive answers in a particular wave* (range 0–4) is this time controlled for panel fatigue as well (see Section 8.2.5 for more details).

Table 8.5: Multiple linear regression and fixed-effect within-person regression results; dependent variable: *number of changes between substantive and non-substantive answers in a particular wave*

| **Derived variable** | **Predictor of changes between substantive and non-substantive answers** | **Multiple linear regression model (pooled)** | | | | **Fixed-effect regression model** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| No. of changes between substantive and non-substantive answers in particular wave | Mode change: online to telephone | 0.06 | -0.06 | 0.17 | 0.336 | 0.10 | -0.09 | 0.29 | 0.309 |
| | Mode change: telephone to online | 0.01 | -0.10 | 0.12 | 0.872 | 0.06 | -0.09 | 0.21 | 0.448 |
| | No. times questions asked | -0.01 | -0.02 | 0.00 | 0.088† | 0.01 | -0.02 | 0.04 | 0.414 |
| | Months since questions asked | 0.02 | 0.01 | 0.02 | <0.001** | 0.02 | 0.02 | 0.02 | <0.001** |
| | Panel fatigue indicator | 0.010 | 0.007 | 0.013 | <0.001** | 0.01 | 0.00 | 0.01 | 0.026* |
| | Constant | -0.11 | -0.16 | -0.05 | <0.001** | -0.17 | -0.20 | -0.13 | <0.001** |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level, †significant at the 0.1 level; socio-demographic controls in the models: gender, age groups, education, state, country of birth

The results show that the mode changes and the number of times questions were asked were not statistically significant predictors of the number of changes between substantive and non-substantive answers in a particular wave. On the other hand, both the number of months since the questions were asked and the panel fatigue indicator had an effect on the number of changes between substantive and non-substantive answers. The longer the gap in months between a question being asked and then repeated, and the more times respondents participated in Life in Australia™ research, the higher the

probability of changes in the substantive answer direction. Panel fatigue, which was highly correlated with the number of times questions were asked, had a fairly small effect on answer changes.

#### 8.3.3.3 Primacy

Primacy as a response-order effect was studied with static and dynamic regression models. The results of logit regression and random-effect regression models (after performing the Hausman test) with *change in satisfaction* and *change in party support* answers as the dependent variables are presented in Table 8.6. This time, the type of answer changes – both substantive answer to first-offered answer and first-offered answer to other substantive answer – were considered as well.

Table 8.6: Logit regression and random-effect within-person regression results; dependent variable: *change of answers from any to first-offered answer (and vice versa)*

| **Substantive repeated survey item** | **Type of change** | **Predictor of primacy change over time** | **Logit regression model (pooled)** | | | | **Random-effect logit regression model** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| Satisfaction | Substantive answer to first offered answer | Mode change: online to telephone | -0.43 | -1.85 | 0.98 | 0.549 | -0.31 | -1.83 | 1.21 | 0.691 |
| | | Mode change: telephone to online | 0.59 | -0.25 | 1.43 | 0.170 | 0.66 | -0.30 | 1.61 | 0.177 |
| | | No. times question asked | 0.18 | 0.11 | 0.26 | 0.000** | 0.23 | 0.15 | 0.31 | 0.000** |
| | | Months since question asked | 0.06 | 0.04 | 0.09 | 0.000** | 0.07 | 0.04 | 0.09 | 0.000** |
| | First offered answer to other substantive answer | Mode change: online to telephone | 0.33 | -0.84 | 1.50 | 0.583 | 0.42 | -0.84 | 1.68 | 0.512 |
| | | Mode change: telephone to online | 0.83 | -0.02 | 1.68 | 0.055† | 0.84 | -0.09 | 1.78 | 0.078† |
| | | No. times question asked | 0.09 | 0.02 | 0.16 | 0.016* | 0.10 | 0.02 | 0.17 | 0.010* |
| | | Months since question asked | -0.09 | -0.12 | -0.05 | 0.000** | -0.09 | -0.13 | -0.05 | 0.000** |
| Party support | Substantive answer to first offered answer | Mode change: online to telephone | 0.11 | -0.92 | 1.13 | 0.836 | 0.16 | -0.99 | 1.32 | 0.783 |
| | | Mode change: telephone to online | 0.39 | -0.45 | 1.24 | 0.362 | 0.44 | -0.51 | 1.40 | 0.362 |
| | | No. times question asked | -0.05 | -0.11 | 0.02 | 0.135 | -0.05 | -0.12 | 0.02 | 0.181 |
| | | Months since question asked | 0.02 | -0.01 | 0.04 | 0.149 | 0.02 | 0.00 | 0.05 | 0.107 |
| | First offered answer to other substantive answer | Mode change: online to telephone | -0.17 | -1.35 | 1.00 | 0.772 | -0.12 | -1.42 | 1.17 | 0.851 |
| | | Mode change: telephone to online | -0.72 | -2.14 | 0.69 | 0.315 | -0.90 | -2.42 | 0.62 | 0.245 |
| | | No. times question asked | 0.03 | -0.03 | 0.09 | 0.335 | 0.03 | -0.03 | 0.10 | 0.330 |
| | | Months since question asked | 0.00 | -0.03 | 0.02 | 0.709 | 0.00 | -0.03 | 0.02 | 0.793 |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level, †significant at the 0.1 level; socio-demographic controls in the models: gender, age groups, education, state, country of birth

The results show that the mode changes *online to telephone* and *telephone to online* are not statistically significant predictors of primacy-related changes in answers. However, the change in the other substantive answer to the satisfaction question after switching to the online mode had a

significant effect at the p<0.1 level in both static and dynamic models. Moreover, both panel conditioning indicators are statistically significant predictors of primacy-related satisfaction answer changes. Although *the number of times question asked* had a positive effect on both types of answer changes (answer change effect), *months since question asked* showed only a primacy effect in this particular case for the satisfaction item. On the other hand, we did not observe any effects of predictor variables on the party support primacy-related answer changes.

#### 8.3.3.4 Recency

Recency as a response-order effect was studied with static and dynamic regression models. The results of logit regression and random-effect regression models (after performing the Hausman test) with *change in satisfaction* answers as the dependent variable are presented in Table 8.7. The type of answer changes – both substantive answer to last-offered answer and last-offered answer to other substantive answer – were considered as well.

Table 8.7: Logit regression and random-effect within-person regression results; dependent variable: *change of answers from any to last-offered answer* (and vice versa)

| **Substantive repeated survey item** | **Type of change** | **Predictor of recency change over time** | **Logit regression model (pooled)** | | | | **Random-effect logit regression model** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| Satisfaction | Substantive answer to last offered answer | Mode change: online to telephone | 1.06 | 0.35 | 1.78 | 0.004** | 1.05 | 0.17 | 1.92 | 0.019* |
| | | Mode change: telephone to online | -0.60 | -2.01 | 0.82 | 0.408 | -0.78 | -2.30 | 0.75 | 0.319 |
| | | No. times question asked | -0.13 | -0.20 | -0.06 | 0.000** | -0.13 | -0.21 | -0.06 | 0.001** |
| | | Months since question asked | 0.02 | 0.00 | 0.05 | 0.086† | 0.03 | 0.00 | 0.06 | 0.033* |
| | Last offered answer to other substantive answer | Mode change: online to telephone | 0.33 | -0.70 | 1.35 | 0.533 | 0.26 | -0.88 | 1.40 | 0.659 |
| | | Mode change: telephone to online | 0.72 | -0.06 | 1.51 | 0.072† | 0.75 | -0.13 | 1.64 | 0.096† |
| | | No. times question asked | 0.04 | -0.03 | 0.11 | 0.278 | 0.04 | -0.03 | 0.12 | 0.288 |
| | | Months since question asked | 0.06 | 0.04 | 0.09 | 0.000** | 0.07 | 0.04 | 0.10 | 0.000** |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level, †significant at the 0.1 level; socio-demographic controls in the models: gender, age groups, education, state, country of birth

The results show that the *online to telephone mode change* predictor had a statistically significant effect on recency-related *substantive answer to last-offered answer* change of answers in both logit and random-effect logit models. The other mode change, *telephone to online,* had a significant positive effect on 'change away from recency' at p<0.1 in both models. The predictor *number of times question asked* had a statistically significant negative effect on recency (answer change effect), and *months since question asked* had a positive effect on any recency-related answer changes in dynamic models and in one of the two logit regression models.

### 8.3.3.5 Social desirability

Social desirability as a type of response bias, related to reporting more socially desirable, acceptable answers or those in sync with the popular opinion, was studied with static and dynamic regression models. The results of logit regression and fixed-effect regression models (after performing the Hausman test), with changes to socially desirable answers as dependent variables, are presented in Tables 8.8 and 8.9.

Table 8.8: Logit regression and fixed-effect within-person regression results; dependent variable: *increased satisfaction and changed support of a 'popular' party* (and vice versa)

| **Substantive repeated survey item** | **Type of change** | **Predictor of change associated with social desirability over time** | **Logit regression model (pooled)** | | | | **Fixed-effect logit regression model** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| Satisfaction | Increased satisfaction | Mode change: online to telephone | -0.10 | -0.67 | 0.48 | 0.742 | -0.70 | -1.83 | 0.44 | 0.232 |
| | | Mode change: telephone to online | 0.37 | -0.14 | 0.88 | 0.155 | 0.24 | -0.47 | 0.94 | 0.512 |
| | | No. times question asked | -0.08 | -0.12 | -0.05 | 0.000** | -0.10 | -0.15 | -0.06 | 0.000** |
| | | Months since question asked | 0.06 | 0.05 | 0.07 | 0.000** | 0.07 | 0.05 | 0.08 | 0.000** |
| | Decreased satisfaction | Mode change: online to telephone | 0.25 | -0.26 | 0.77 | 0.336 | 0.25 | -0.77 | 1.28 | 0.629 |
| | | Mode change: telephone to online | 0.51 | 0.01 | 1.00 | 0.044* | 0.26 | -0.45 | 0.96 | 0.479 |
| | | No. times question asked | -0.22 | -0.25 | -0.18 | 0.000** | -0.31 | -0.36 | -0.27 | 0.000** |
| | | Months since question asked | -0.01 | -0.02 | 0.01 | 0.249 | -0.02 | -0.04 | 0.00 | 0.028* |
| Party support | Any other answer to "popular" opinion answer | Mode change: online to telephone | 0.66 | -0.06 | 1.38 | 0.074† | 0.00 | -1.19 | 1.20 | 0.997 |
| | | Mode change: telephone to online | 0.19 | -0.60 | 0.98 | 0.634 | 0.34 | -0.75 | 1.43 | 0.545 |
| | | No. times question asked | -0.05 | -0.11 | 0.00 | 0.057† | -0.15 | -0.22 | -0.07 | 0.000** |
| | | Months since question asked | 0.03 | 0.01 | 0.05 | 0.004** | 0.03 | 0.01 | 0.06 | 0.007** |
| | "Popular" opinion answer to any other answer | Mode change: online to telephone | 0.56 | -0.25 | 1.36 | 0.174 | 0.29 | -1.17 | 1.75 | 0.694 |
| | | Mode change: telephone to online | 0.13 | -0.71 | 0.98 | 0.758 | -0.30 | -1.52 | 0.91 | 0.628 |
| | | No. times question asked | -0.03 | -0.08 | 0.03 | 0.317 | -0.07 | -0.15 | 0.00 | 0.064† |
| | | Months since question asked | -0.02 | -0.04 | 0.01 | 0.186 | 0.00 | -0.03 | 0.03 | 0.856 |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level, †significant at the 0.1 level; socio-demographic controls in the models: gender, age groups, education, state, country of birth

The results presented in Table 8 show that the *mode change online to telephone* is not a statistically significant predictor of the social desirability–related changes of satisfaction or 1st problem answers. On the other hand, *mode change telephone to online* positively affects decreased satisfaction (pooled model only). Those respondents reported lower satisfaction in the self-administered mode. On the other hand, *mode change online to telephone* positively affects changing party support answers to the

'popular opinion' answers, but at p<0.1 and in the pooled model only. Those respondents supported the two biggest Australian parties with a slightly higher propensity in the interviewer-administered mode.

On the other hand, *the number of times questions asked* negatively affected any satisfaction and other changes to 'popular opinion' answers (party support, 1st problem), and positively affected changes between 'environment' and other answers and vice versa (answer change effect). Variable *months since question asked* positively affected changes to socially desirable answers: increased satisfaction and selecting 'popular opinion' answers to party support and 1st problem 'environment' answer (social desirability effect, Table 8.9).

Table 8.9: Logit regression and fixed-effect within-person regression results; dependent variable: *changing answers to popular opinion about most important problems in the country* (and vice versa)

| **Substantive repeated survey item** | **Type of change** | **Predictor of change associated with social desirability over time** | **Logit regression model (pooled)** | | | | **Fixed-effect logit regression model** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Coef | L 95% CI | U 95% CI | p value | Coef | L 95% CI | U 95% CI | p value |
| 1st problem | Any other answer to "popular" opinion answer | Mode change: online to telephone | 0.44 | -0.09 | 0.97 | 0.106 | 0.41 | -0.41 | 1.23 | 0.330 |
| | | Mode change: telephone to online | 0.09 | -0.47 | 0.65 | 0.747 | -0.13 | -0.85 | 0.59 | 0.732 |
| | | No. times question asked | -0.06 | -0.09 | -0.02 | 0.002** | -0.10 | -0.14 | -0.05 | 0.000** |
| | | Months since question asked | 0.00 | -0.02 | 0.01 | 0.955 | 0.01 | -0.01 | 0.02 | 0.550 |
| | "Popular" opinion answer to any other answer | Mode change: online to telephone | -0.24 | -0.93 | 0.45 | 0.497 | -1.05 | -2.23 | 0.13 | 0.080† |
| | | Mode change: telephone to online | 0.05 | -0.52 | 0.63 | 0.856 | -0.01 | -0.82 | 0.81 | 0.990 |
| | | No. times question asked | -0.03 | -0.07 | 0.00 | 0.083† | -0.03 | -0.08 | 0.02 | 0.209 |
| | | Months since question asked | 0.02 | 0.00 | 0.03 | 0.027* | 0.04 | 0.02 | 0.05 | 0.000** |
| 1st problem | Any other answer to *Environment* answer | Mode change: online to telephone | 0.36 | -0.50 | 1.22 | 0.409 | 0.18 | -1.05 | 1.41 | 0.778 |
| | | Mode change: telephone to online | 0.19 | -0.66 | 1.04 | 0.661 | -0.28 | -1.50 | 0.94 | 0.654 |
| | | No. times question asked | 0.06 | 0.00 | 0.12 | 0.039* | 0.16 | 0.08 | 0.24 | 0.000** |
| | | Months since question asked | 0.06 | 0.04 | 0.08 | 0.000** | 0.09 | 0.06 | 0.12 | 0.000** |
| | *Environment* answer to any other answer | Mode change: online to telephone | 0.48 | -0.45 | 1.41 | 0.312 | -0.78 | -2.47 | 0.91 | 0.365 |
| | | Mode change: telephone to online | -0.26 | -1.42 | 0.90 | 0.665 | -0.41 | -1.98 | 1.17 | 0.612 |
| | | No. times question asked | 0.09 | 0.03 | 0.15 | 0.004** | 0.18 | 0.10 | 0.27 | 0.000** |
| | | Months since question asked | 0.00 | -0.03 | 0.02 | 0.716 | 0.03 | 0.00 | 0.06 | 0.037 |

Notes: Coef = model regression coefficient, L 95% CI = lower limit of the 95% confidence interval, U 95% CI = upper limit of 95% confidence interval, *significant at the 0.05 level, **significant at the 0.01 level, †significant at the 0.1 level; socio-demographic controls in the models: gender, age groups, education, state, country of birth

## 8.4 Discussion and recommendations

The existing literature on measurement mode effects in probability-based mixed-mode research is limited. One of the limitations to understanding measurement mode effects better is the inability to fully disentangle measurement mode effects from subsample composition effects. To achieve this, an optimal randomised design would have to be applied, which means that all onliners and offlliners in treatment and control groups would have to have an equal non-zero probability of being assigned to either the online or the offline mode. However, this kind of randomisation is almost impossible, because most offline respondents cannot or refuse to respond online, and the cost of administering telephone compared with online delivery mode means that survey companies are unlikely to significantly (and randomly) increase the number of offline respondents. In this study, we instead used the fact that certain respondents, although the percentage is small and non-random, appear in both modes over time and respond to a limited number of repeated questions. Consequently, we could not only study mode effects related to questionnaire administration, we could also assess how much of a measurement error may be introduced by allowing respondents to respond in different modes, especially if we would like to measure changes over time in a quasi-longitudinal design.

An important finding of this study was that answers from the same respondents vary greatly over time, even for items for which a slightly higher consistency would be expected, such as party preference, and for short time gaps between survey interviews. Stability of answers differed significantly between different political attitude items, but very little instability could be explained by socio-demographic characteristics. At the same time, respondents switching modes affected stability to a smaller extent than we expected based on the relevant literature. This might be a result of there being only a small subsample of mode switchers. We observed several coefficients that indicated an impact of switching modes on changing answers, consistent with the measurement mode effect literature, but the effects were often significant at the $p< 0.1$ and not the $p<0.05$ pr $p<0.01$ levels. Bigger samples – for example, with additional panel survey data with mode switchers or people with a higher propensity to change modes – would increase the statistical power and may show the effects to be more statistically significant.

Nevertheless, we found a few notable measurement biases after switching modes; the mode change was a statistically significant predictor of the number of changes in answers – switching decreases the stability of answers, has a positive effect on recency when switching to interviewer-administered telephone mode, and has a negative effect on social desirability in the self-administered mode for a limited number of items. These findings are in line with the theory on measurement mode effects in online panels (Baker et al. 2010). On the other hand, many different changes in answers could be

better explained by panel conditioning. Generally, the more times the same questions are asked over time, the lower the probability of changes, and the longer the gap (measured in months) between asking questions, the lower the stability of answers. This is consistent with findings of Cernat (2015) and Wooden and Li (2014) on the effect of panel conditioning on reliability and stability of answers. In our study, the analysis of individual types of changes, normally attributed to measurement mode effects, offered mixed evidence for both indicators of the extent of panel conditioning. Both regressors were associated with something we called 'answer change effect', but in some cases in a positive and in other cases in a negative way. The number of months since the question was asked was slightly more strongly associated with phenomena normally attributed to measurement mode effects than the other indicator of panel conditioning. The contribution of this study is to present evidence on the severity of panel conditioning effects when respondents are conditioned repeatedly with short time intervals, sometimes being asked the same question in consecutive months. The existing literature on panel conditioning (e.g., Cernat 2015; Sturgis et al. 2009; Wooden & Li 2014) mostly studied this source of measurement error in longitudinal studies, where the time between data collection waves is much longer. We can conclude that panel conditioning seems to play an important role in the stability of answers; researchers should pay extra attention if the same question is asked several times in a short period of time, which might prevent respondents from reporting naturally changed attitudes over time.

In this study, we faced a number of limitations. Because we investigated measurement mode effects as changes in responses to the same questions from the same respondents over time, we had to control the effect of switching modes with the other sources of measurement errors specific to panel research. The reason for this is that, in this study design, all respondents were conditioned in at least one wave before providing the same or different answers in the next wave. The subsample used in this study therefore consisted of respondents who participated in at least two waves out of six for which we could find repeated items measuring political attitudes. Infrequent respondents were not included in the sample, which might have introduced some representation bias. Moreover, panel conditioning had to be controlled in the models in a slightly different way. We did not compare distributions of the selected response variables between those who answered the question for the first time and those who had been conditioned by being asked the same question in the past. With this study design, we instead controlled measurement mode effects with the effect of the extent/severity of panel conditioning. Also, we note that some respondents, who are panellists in Life in Australia™, regularly participate in other cross-sectional and/or nonprobability-based panel research. Unfortunately, we could not control for potential panel conditioning as a result of survey participation outside of Life in Australia™ research. One out of five different concepts related to

measurement bias – sensitivity – was investigated through item nonresponse and non-substantive answers. Because these concepts are associated with another source of measurement errors specific to panel research – panel fatigue – we included an indicator of panel fatigue in the models investigating factors affecting sensitivity. Last but not least, our findings might be less generalisable in fields outside political attitudes research, because all of the survey items used in this study were from ongoing political poll research.

The contribution of this study is, first and foremost, in identifying certain measurement mode effects, such as recency or social desirability, as a result of panellists switching modes of data collection in online panel research. We also noted that switching modes might induce more measurement error due to satisficing and social desirability if the proportion of mode switchers was higher. Although measurement mode effects themselves do not seem to affect the accuracy of estimates in a very negative way, combining them with panel conditioning, as well as voluntary attrition and nonresponse, may lead to less accurate estimations. The future research on the accuracy of estimation of attitudinal changes over time in probability-based panels should, therefore, focus on studying concurrent sources of survey errors specific to online panels and their effects on accuracy.

## 8.5 References


Albert, W., & Tullis, T. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics* (2nd ed.). Morgan Kaufmann.

Australian Bureau of Statistics. (2016). *2016 Census – cultural diversity*. Findings based on use of ABS TableBuilder data.

Baghal, T. A., & Kelley, J. (2016). The stability of mode preferences: implications for tailoring in longitudinal surveys. *Methods, Data, Analyses, 10*(2), 143–166.

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

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

Callegaro, M. (2008). Social desirability. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 825-826). Sage.

Cantwell, P. J. (2008). Panel conditioning. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 566-567). Sage.

Centre for Social Research and Methods. (n.d.). *ANUPoll*. Retrieved March 1, 2019, from https://csrm.cass.anu.edu.au/research/surveys/anupoll

Cernat, A. (2015). The impact of mixing modes on reliability in longitudinal studies. *Sociological Methods & Research, 44*, 427–457.

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics, 21*(5), 233–255.

Dillman, D. A. (2009). Some consequences of survey mode changes in longitudinal surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 127-140). Wiley.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons.

Grimm, P. (2010). Social desirability bias. In W. Kamakura (Ed.), *Wiley international encyclopedia of marketing – vol 2: Marketing research* (pp. 258-259). Wiley.

Halpern-Manners, A., & Warren, J. R. (2012). Panel conditioning in longitudinal studies: evidence from labor force items in the current population survey. *Demography, 49*(4), 1499–1519.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, *46*(6), 1251–1271.

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

Keeter, S., McGeeney, K., Igielnik, R., Mercer, A., & Mathiowetz, N. (2015). *From telephone to the web: the challenge of mode of interview effects in public opinion polls*. Pew Research Center.

Kim, S. H., & Kim, S. (2016). Social desirability bias in measuring public service motivation. *International Public Management Journal, 19*(3), 293–319.

Kruse, Y., Callegaro, M., Dennis, J., DiSogra, C., Subias, S., Lawrence, M., & Tompson, T. (2009). Panel conditioning and attrition in the AP-Yahoo! news election panel study. *Proceedings of the 64th Conference of the American Association for Public Opinion Research*, 1-15.

Lavrakas, P. J. (2008). Panel fatigue. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 569-570). Sage.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: a review. *European Journal of Social Psychology, 15*(3), 263–280.

Social Research Centre. (2018). *ANUPoll 27 – Data governance in Australia Technical Report*. The Social Research Centre.

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: the psychology of panel conditioning. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 113–126). Wiley.

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

Wooden, M., & Li, N. (2014). Panel conditioning and subjective well-being. *Social Indicators Research 117*(1), 235–255.

Zimbalist, Z. (2017). 'Fear-of-the-state bias' in survey data. *International Journal of Public Opinion Research, 30*(4), 631–651.

# Chapter 9 Comparing and improving the accuracy of nonprobability sample surveys

## 9.1 Introduction

It has become increasingly evident that traditional surveys are becoming less responsive to measuring and understanding emerging and complex social issues. Traditional surveys often fail to accurately measure individual behavior, attitudes and perceptions on issues such as migration and fertility (Tourangeau et al. 2014). Recent notable failures of polls to predict the outcomes of referenda and elections have shown that the way in which data are collected from the population must be responsive to people's dynamic lifestyles, choices and attitudes. The widespread availability of and access to the internet and social media leads to a quick diffusion of ideas that may rapidly shift social attitudes and behaviors.

While we are currently witnessing a period of great social unrest, traditional survey methods are proving to be inadequate for capturing new-to-emerge, quick-to-change social events. This means sociologists and demographers are often struggling to catch up with trends identified in more nimble fields (e.g., psychology or journalism). Fortunately, the internet has ushered in new web-based methods that have been widely utilized in other fields (e.g., political science or market research). Web-based surveys are advantageous given their convenience, quick turn-around times, and relatively low respondent costs. Additionally, they allow tests for consistency and reliability to be performed in a timelier manner than telephone and interviewer administered surveys.

The main difference between traditional surveys and online surveys is that traditional surveys use probability sampling, which means that there is a random selection of survey participants based on established statistical and survey methods, whereas online surveys are nonprobability-based and rely on individuals self-selecting to participate in surveys. This respondent self-selection has been the main cause for concern, rendering the usual survey sampling theoretical approaches inapplicable and leading to issues about biases, lack of generalizability and causal inference (Baker et al. 2013; Elliott & Valliant 2017; Mercer et al. 2018). Furthermore, the absence of an underlying statistical theory limits the use and acceptability of nonprobability surveys for a number of reasons.

Firstly, respondents are not selected based on probability sampling. Even though they may be 'randomly' selected, it is often not possible to work out their chance of being selected into the survey. This means that certain people who regularly fill out such surveys skew the data, thus skewing the results. If these respondents tend to be "distinct" in particular ways, then the results from nonprobability online surveys are not generalizable to the general population. This leads to the second

limitation whereby online survey respondents have been found to have different characteristics and behaviors to respondents from more traditional surveys (i.e., internet access tends to be positively associated with income and education and negatively associated with age). In many instances empirical testing confirms that survey administration by computer leads to higher reports of socially undesirable behavior than interviewer administration (Dillman et al. 2009; Kaplowitz et al. 2004). While this is associated with measurement mode effect in mixed-mode surveys, it can lead to more accurate reporting in online-only surveys. Additionally, computer-administered surveys generally have higher levels of (item) nonresponse due to their questionnaire design (Couper 2000), while on average yielding 12% lower response rates than other modes (Daikeler et al. 2020), which has potential to introduce more nonresponse bias. Thirdly, there is no sampling frame for the internet. While virtually all internet users have an email address there is no list comprising all of these email addresses that could be used to draw a random sample. Individuals also tend to have several email addresses, or may share a single address with family members, and addresses may fall into disuse without being deactivated. Fourthly, the internet does not have universal coverage: in Australia, between 12% (Datareportal 2020) and 14% of the population (Australian Bureau of Statistics 2018a) has no access to the internet and this lack of access is concentrated amongst those of older age, rural location, Indigenous ethnicity, and lower education levels: groups which are increasingly important to policymakers, hence limiting the utility of internet-based surveys. In the United States (US) context, certain socio-demographic groups are less likely to have access to the internet (noncoverage bias), which can be of greater concern than nonresponse bias among those groups (Couper et al. 2007; Couper et al. 2018). Collectively, these limitations mean that the data collected online are less reliable than those gathered by traditional survey methods, and may not accurately represent trends in the general population. However, online nonprobability surveys are significantly less expensive, more flexible, quickly implemented and easier to complete than traditional surveys. As governments are increasingly being asked to solve complex and urgent social problems with limited budgets, robust statistical tools for analyzing online data are vital. These perceived advantages have to be weighed against the fact that there is self-selection of respondents, who often receive incentives, which renders design-based methods of survey inference inapplicable and raises concerns about the potential for biased results (Baker et al. 2013; Mercer et al. 2017).

Our research aims are three-fold. Firstly, we evaluate and quantify the differences in survey estimates obtained from the same survey administered through a probabilistic sampling framework in contrast with those from a non-probabilistic framework. Secondly, we examine the representativeness of the various survey estimates against three categories of benchmark variables: primary demographics (such as age and gender); secondary demographics (such as citizenship and employment status); and

non-demographics (such as alcohol consumption and life satisfaction). Finally, we compare and contrast the performance of different post-survey adjustment methods on reducing bias in nonprobability-based surveys (such as raking or matching methods).

## 9.2 Background and literature review

The textbook definition of a probability sample is one in which every unit in the population of interest has a known, non-zero, chance of being selected for the sample through a random process. Differently stated, these selection probabilities ensure that every unit in the population has a unique chance of being selected into the sample. This randomisation is a key design attribute of probability sampling, and enables the calculation of standard errors, confidence intervals, and making generalized inferences regarding the target population of interest from the sample. However, while most (probability) surveys have known selection probabilities, whether people respond cannot be controlled for, in spite of all the best efforts of survey practitioners, and ultimately it is sample inclusion not selection that matters (Rivers 2013). Trends of high nonresponse rates with a large proportion of probability-based surveys reporting response rates of under 10% (Kennedy & Hartig 2019), and the associated nonresponse biases may lead to flawed results and problems in statistical inference (Baker et al. 2010; Baker et al. 2013). However, the fact that the selection probabilities for a sample are unknown does not imply that they cannot be estimated or adjusted for in a nonprobability sample, just as adjustments are used in probability-based surveys to compensate for issues around coverage and response (Rivers 2013).

### 9.2.1 Accuracy of nonprobability samples

Just as there is a whole spectrum of surveys that purport to be probability-based, there is a whole gamut of online nonprobability-based surveys, from the opt-in click-through unsolicited surveys which are advertised on websites, to more structured recruitment of a panel of respondents who receive incentives for participation in surveys. For the former nonprobability surveys, as a result of the idiosyncratic designs which make it difficult to work out the rates of contact, response, and (non)coverage it is almost impossible to make reasonable statistical inferences from data obtained in this manner (Rivers 2013). However, in the latter the characteristics of the sample of those recruited may closely resemble the population being studied and identifying the conditions under which valid statistical inferences can be made using the realized sample is important (Mercer et al. 2017). This selection bias – which is the systematic differences between a statistical estimate and the true population parameter – can be controlled for using a number of different approaches, underpinned by an existing framework based on causal inference used in numerous fields such as epidemiology, political science and economics (Heckman 1979; Hug 2003; Rothman et al. 2008).

Using notation from Elliott and Valliant (2017) and Valliant (2020), we define *s* as the sample selected to participate in a survey using a nonprobability recruitment strategy (for instance using an opt-in panel). Our problem is to make inferences about *s* to the full population *U*. We assume, (i) every unit in the population has some probability of being included in the sample, and (ii) there is a structural model based on the observed sample which can be used to describe the variables we are interested in measuring. Under (i) and (ii) we can correct for any noncoverage (or selection bias) through reweighting schemes (such as poststratification or raking or other regression-based methods) for accurate statistical inference. More importantly, Valliant (2020) and Elliott and Valliant (2017) showed that conditions (i) and (ii) are not necessary and sufficient, meaning that you do not need both conditions to hold. This implies these reweighting or matching schemes can be used to (a) estimate the probability of response and (b) calibrate to known benchmark population totals, to correct for any selection biases in the estimates derived from nonprobability sample surveys (Matei 2018). We distinguish between matching approaches which are non-parametric methods of controlling for any confounding effects in the observed data, and reweighting approaches which adjust the observed sample according to a specified set of control variables through statistical procedures. However, the key goal of both approaches is to ensure that there is no (or little) bias in the observed data, meaning that the empirical distribution of the observed data is similar to the population (Baker et al. 2013; Elliott & Valliant 2017; Mercer et al. 2017; Mercer et al. 2018; Valliant 2020).

### 9.2.2 Post-survey adjustments in nonprobability samples

The problem facing nonprobability samples is related to the distorted representation of observed data as a consequence of the way in which the decisions by the survey practitioner or the respondents influence how they are selected into the sample. Note that this problem of selection bias arises also in probability samples that do not have simple random sampling of the underlying population since no matter how big the sample size, a sample not selected using simple random sampling produces a realization of the population distribution that does not accurately describe the true population distribution (Cornesse et al. 2020; Elliott & Valliant 2017). In fact, post-survey adjustments which correct for the unequal probabilities of selection are common in most probability surveys: virtually no probability sample uses simple random sampling. As such, in both probability and nonprobability samples, the objective for inference is to ensure that the composition of the sampled units with respect to the observed characteristics either matches or can be adjusted to match the population of interest. For that matter, in nonprobability samples, analytical steps are often carried out, post-survey, in order to account for any differences (perceived or otherwise) between the observed sample, and the true population distribution. These post-survey adjustments have the dual purpose of reducing

the bias and producing more accurate population estimates (Elliott & Valliant 2017; Mercer et al. 2017).

There are a number approaches which have been proposed to improve accuracy and inference for data collected under a nonprobability sample. These approaches basically are predicated from the issues facing probability samples caused by differences in response and coverage of surveys. To cope with these issues, statistical adjustments typically correct for any systematic biases. Many of these approaches have been adapted to cope with systematic biases in nonprobability samples (Cornesse et al. 2020; Elliott 2009; Rivers 2007).

This study compares six primary methods of reweighting and matching survey data: raking, generalized regression estimation (GREG), propensity score weighting (PSW), multilevel regression and poststratification (MRP), Mahalanobis distance matching (MDM) and coarsened exact matching (CEM).

Reweighting methods directly adjust the sample distribution to the target population distribution, to achieve the desired sample composition in the presence of nonresponse and/or other factors. On the other hand, matching attempts to create a balanced nonprobability sample which closely resembles the characteristics of a probability sample from the 'true' population (when compared with a selected array of auxiliary (often non-demographic) characteristics) (Bethlehem 2016; Cornesse et al. 2020). In order to ascertain the performance of the different approaches, we evaluate how closely the final adjusted model estimates compare to the population characteristics, and thereby provide an assessment of the accuracy of the nonprobability samples. Through contrasting the results before and after adjustment, we are also able to quantify the improvement of the various adjustment processes. The majority of studies that have empirically evaluated how nonprobability surveys compare with probability samples have reiterated the importance of the availability of population benchmark information and how this is related with the accuracy the nonprobability samples (MacInnis et al. 2018; Pennay et al. 2018). Our objective in this study is to determine which post-adjustment approach fares best, and we achieve this through trialling out different scenarios that vary in regards to the availability of external data for good quality benchmarking (described in Section 9.2.4).

#### 9.2.2.1 Raking

For most surveys, the most common method for weighting is raking, also known as iterative proportional fitting. Here, the usual way is to start off with a set of variables where the population distribution is known, the sample is then divided into mutually exclusive cells and then repeatedly adjust the weights for each individual within each cell until the sample distribution is perfectly aligned with the population distribution (for the selected set of variables). Raking is simple to implement since

it relies only on knowing the marginal distributions of the selected variables, and there is evidence that the utility of a large set of variables diminishes, implying that in most cases using key demographic and socio-economic variables is often sufficient to reduce the selection bias (Kalton & Flores-Cervantes 2003).

#### 9.2.2.2 Generalized regression estimation (GREG)

Generalized regression estimation (GREG) is a calibration approach where the sampling weights are adjusted to make certain the survey estimators match to the set of known population totals (benchmarks). In contrast to raking which repeatedly reweights the sample to the marginal distributions of the known population totals, the GREG estimator is based on the minimizing the distance measure between the sample and the benchmark information and it is supposedly more efficient and provides more accurate population estimates (Deville & Särndal 1992). While this is more efficient, the GREG estimator becomes less precise the larger the number of benchmarks since there is more volatility introduced into the cross-classified information (and in these instances, simple raking to the marginal estimates might be better (Deville et al. 1993)). To compensate for this, the GREG estimation is performed for fewer benchmarks (for our situation we use the primary and secondary demographic information).

#### 9.2.2.3 Propensity score weighting (PSW)

In the simplest version of probability-based sampling, survey respondents are assumed to have a non-zero chance of being included in the sample (referred to as the sample selection probability), and weighting each sample individual by the inverse of its sample selection probability removes any selection bias (Cochran 1977). This weighting corrects for having different types of people over- or under-represented in the sample when compared to the population. When data are collected through a nonprobability-based sample, we can use the same ideas. But here the difference is that for a nonprobability-based sample, the selection probabilities cannot be easily computed, and in most cases are not known. However, the fact that selection probabilities from a nonprobability sample are unknown does not mean that they cannot be estimated (Rivers 2013). In PSW, the technique here is to first create a synthetic population, which is assumed to “represent” the full target population, or to use external high-quality data representative of the population. Second, pseudo-inclusion probabilities are estimated which leads to a probability-based (or synthetic) reference sample which is combined with the nonprobability sample. The pseudo-inclusion probabilities for the nonprobability cases can subsequently be estimated using binary (i.e., probit or logistic) regression modeling (Schonlau & Couper 2017; Valliant 2020). Like in calibration (raking), PSW is efficient in bias reduction

if the weighting variables and the propensity of response in the nonprobability sample are (strongly) associated with outcome variables (Rosenbaum & Rubin 1983; Valliant & Dever 2011).

#### 9.2.2.4 Multilevel regression and poststratification (MRP)

The MRP approach (Gelman 2007; Gelman & Little 1997) is based on assuming the existence of a super-population model which can be fitted to the analytic survey variables and can be used to project the observed sample to the full population. The key assumption here is that sampled and nonsampled data are driven by an underlying model and this model can be revealed by analyzing the sample responses. In the presence of nonresponse, this model also specifies the relationship between the observed units and the unobserved data (Brick 2013). As this (model-based) approach relies on the existence of a model that truly represents the population based on the observed sample, different realizations of the observed sample can lead to different models, and as such it can appear to be less flexible than the pseudo-random approach (in PSW) which produces a unique set of pseudo-inclusion probabilities (Valliant 2020). For MRP we begin by creating a set of post-strata through cross-tabulating the set of survey covariate predictors, based on the model parameters. The estimated (nonprobability) proportion in each post-strata, is given by relative size in each post-strata multiplied by the estimated mean value. This mean value can be estimated by fitting mixed effects (multilevel) model which smooth the noisy estimates in the post-strata with fewer data through borrowing strength from the overall or nearby information. This has the effect of correcting the model estimates for any differences between the sample population (here the opt-in panel) and the target population. In political science, this approach is useful in obtaining state-level predictions based on relatively small national samples (for example, Bon et al. 2019; Park et al. 2004; Park et al. 2006; Wang et al. 2015). Poststratification requires knowledge of the joint distribution of the poststratification variables in the target population, while other reweighting methods (such as GREG and raking) require only the knowledge of the marginal distribution of the adjustment variables (Deville & Särndal 1992).

#### 9.2.2.5 Mahalanobis distance matching (MDM)

The idea of matching is similar to PSW or poststratification, and the objective is to create groups containing one or more observations from both the reference sample and the nonprobability sample that are similar on a set of auxiliary variables believed to be associated with the probability of selection. Weights can be added to adjust the distribution of the nonprobability sample to the reference sample. There is one critical difference between PSW or poststratification with matching, and it is that observations that are unmatched are often discarded from the matched dataset. This can lead to invalid inferences due to the loss of information from those unmatched data (Mercer et al. 2017). To mitigate against this, most matching packages now automatically identify those

observations in the reference sample for which there are no counterparts in the nonprobability sample (Ho et al. 2011; Stuart 2010).

In MDM, we measure the distance between a pair of observations, $y_i$ and $y_j$, with the Mahalonobis distance calculated as presented in Equation 9.1:

$$M(y_i, y_j) = \sqrt{(y_i - y_j)^T S^{-1} (y_i - y_j)} \qquad (9.1)$$

where $S$ is the sample covariance matrix of y. Two observations are matched if they have the minimum distance out of a set of pairs, and the simplest way of doing this is through nearest neighbour matching, where the set of nonprobability sample units are sequentially matched to the nearest reference sample unit based on the minimum (Mahalonobis) distance. Since the population of possible match-pairs exponentially increases as the nonprobability sample size increases, usually some procedure is used to remove pairs that are unreasonably distant through defining calipers which are chosen cut-offs for which the maximum distance is allowed (Stuart & Rubin 2008).

#### 9.2.2.6 Coarsened exact matching (MDM)

Coarsened exact matching (CEM) is similar to the MDM, but the key difference is that calipers (cut-offs) are not required to remove unreasonably bad matches (Iacus et al. 2011). First, each variable is coarsened (recategorized into fewer groups, for instance 10-year age groups instead of 5-year groups are used), and second, units with the same values of the coarsened variables are placed in a single stratum, and finally within each stratum, the units in the nonprobability sample are weighted to be equal to the number of units in the reference sample. Strata without at least a single nonprobability sample or reference sample unit, are given a zero weight which effectively prunes them from the dataset. After matching, the coarsening is reversed which results in a final analytic (matched sample) comprised of both the uncoarsened values of the stratification variables and the unpruned units and as such the inference is generally improved because it achieves a better balance between the empirical distributions of reference sample and the nonprobability sample (Iacus et al. 2009; Stuart 2010).

### 9.2.3 Measures of survey quality – Total Survey Error framework

Assessing quality in surveys requires an objective standard to which the survey estimates can be compared. However, measures of quality have predominantly developed for application to probability-based surveys, and as such cannot be used for nonprobability surveys due to the violation of three key assumptions. First, the sampling frame does not cover the whole population in its entirety. Second, not every sample unit has a unique non-zero probability of selection. Third, and relatedly, it

is not straightforward to calculate this probability of selection. The fundamental problem of nonprobability surveys is that the internet does not have a comprehensive sampling frame which means that there is no way to randomly draw a sample for which everyone has a positive chance of being selected. Therefore, to model and make inferences from the data obtained through nonprobability when compared with probability surveys, we will use the Total Survey Error framework (Groves et al. 2009; also see Biemer 2010; Groves & Lyberg 2010). The Total Survey Error framework has been developed to provide a comprehensive overview of all possible sources of sampling and non-sampling errors and give a systematic measure of survey quality that encompasses not just accuracy but also bias. The Total Survey Error paradigm attempts to account for, and assess, many sources of error that arise through the survey process. We primarily use information from the Australian quinquennial Census as benchmarks since censuses offer universal coverage of the population by definition. However, for some instances we will use administrative record data and information drawn from large government surveys. Benchmarks are often not available for questions on attitudes and behaviors not studied by the government in a complete enumeration of the population, and as a result we use these measures as characteristics of interest for inference. Under the Total Survey Error perspective, a survey error is defined as the deviation of the survey response from its true underlying value (i.e., the population benchmark). This error can occur through bias or variance, where the bias term captures the systematic (selection) errors that are shared by nonprobability samples. The variance term captures the sampling variation and accounts for the variation due to the differences in survey protocols, statistical modeling or weighting adjustments. Through adopting this framework, we can ensure that the analysis does not conflate selection bias and non-sampling errors (Shirani-Mehr et al. 2018).

### 9.2.4 Scope of this study

Following from Mercer et al. (2017), we use the general framework which emphasizes the characteristics of the realized sample (regardless of how it was generated), and therefore correct for any self-selection bias in survey inference (Groves 2006; Keiding & Louis 2016; Little & Rubin 2002). The authors identify three components that determine whether or not the presence of self-selection ultimately leads to biased survey estimates. These are

(1) Exchangeability – for all sampled units, are all confounding variables known and measured?
(2) Positivity – does the sample contain the full spectrum of differences in the target population, or does it systematically miss particular segments of the population?
(3) Composition – with regard to the confounding variables, do the sample and population distributions match, and if not, can they be adjusted to match?

These components of self-selection bias are not fundamentally different for nonprobability samples, but what differs between probability and nonprobability samples are the underlying assumptions which lead to individuals becoming members of nonprobability samples (Kennedy et al. 2016; MacInnis et al. 2018; Pfeffermann et al. 2015).

Notwithstanding, this framework can be useful in investigating if there is (a) improved inference of sample data from a nonprobability survey, and (b) through comparing different post-survey adjustment methods we can ascertain their suitability/performance under various conditions. There have been a number of authors – for instance, DiSogra et al. (2011), Baker et al. (2013), Mercer et al. (2017), Mercer et al. (2018), and Valliant (2020) – who have undertaken similar research into the performance of different methods, and also discussed the requirements with respect to the external data sources for the various approaches.

Therefore, we will primarily examine a range of survey estimates against three categories of population benchmarks: primary demographics (such as age and gender); secondary demographics (such as citizenship and employment status); and non-demographics (such as alcohol consumption and life satisfaction). We will investigate the performance of different post-survey estimates under four realistic scenarios (which differ in the availability of external/auxiliary data, see Table 9.1 for more information):

- Scenario 1 – availability of census aggregated statistics and nonprobability-based data including primary demographics only
- Scenario 2 – availability of census aggregated statistics and nonprobability-based data including both primary and secondary demographics
- Scenario 3 – availability of census aggregated statistics, one other representative source of non-demographic benchmarks (i.e., a large national survey) and nonprobability-based data with matching non-demographic survey items
- Scenario 4 – availability of census aggregated statistics, a smaller scale probability-based survey data[60], and nonprobability-based data with matching non-demographic survey items.

[60] The difference between Scenarios 3 and 4 is the type and the source of auxiliary survey data available for post-survey adjustment. Under Scenario 3, we have access to a large-scale nationally representative survey (large sample, e.g., 20,000+, with higher accuracy), e.g., National Drug Strategy Household Survey. Under Scenario 4, we can use a smaller probability-based sample (e.g., n=600), but with an ability to collect tailor-made data including key covariates which could help mitigate bias after matching or propensity scoring weighting (e.g., 'webographic' variables). Under this scenario, data collectors attempting to improve the accuracy of their nonprobability samples could conduct a smaller-scale probability-based survey (e.g., a probability-based sample from Online Panels Benchmarking Study to improve inference in opt-in panel samples).

Table 9.1: Post-survey adjustment scenarios (based on auxiliary data availability)

| **Scenario** | **Covariates available in nonprobability data (and are matching in probability data)** | **Probability data source** | **Type of probability data used in post-survey adjustment** |
|---|---|---|---|
| Scenario 1 | Primary demographics | Population census | Aggregated/ tabular data |
| Scenario 2 | Primary and secondary demographics | Population census | Aggregated/ tabular data |
| Scenario 3 | Primary demographics, secondary demographics and non-demographics (e.g., health-related covariates) | Population census, large national survey (e.g., on health) | Microdata/ unit record files |
| Scenario 4 | Primary demographics, non-demographics (e.g., 'webographics') | Population census, smaller-scale non-government survey | Microdata/ unit record files |

The compulsory nature of official statistics in Australia means that general social surveys achieve relatively high response rates – for instance, the Australian Health Survey, which was used as a source of most of our non-demographic benchmarks, had a response rate of around 80% (Australian Bureau of Statistics 2018b). In contrast, similar health surveys in the US and United Kingdom (UK) achieve less than half that response (Harrison et al. 2020; Leeper 2019). This might be partly attributable to the fact that, under a parliamentary act, participation in official data collections is compulsory in Australia (not just for the census, as is the case in the UK or the US). As such the information available for primary and secondary demographic population-level benchmarks are measured to a high degree of accuracy. This is one advantage our study has over similar studies conducted in other countries.

This study will address the following research question: *How accurate are nonprobability online samples in comparison to probability samples and to what extent can inference be improved by using post-survey adjustment methods under different scenarios?*

## 9.3 Methods

### 9.3.1 Data

#### 9.3.1.1 Original Online Panel Benchmarking Study (2015 OPBS)

In 2016-17, around the time the 2015 OPBS was carried out, 86% of Australian households could access the internet at home – this figure was up from 56% in 2004-05 (Australian Bureau of Statistics 2018a). The offshoot of this is that the volume of survey research conducted online has exponentially increased, leading to a rapid proliferation of nonprobability online surveys. Unfortunately, this has not been accompanied by a clearer understanding of the issues pertaining to inference from such surveys.

To inform the debate in Australia, the 2015 Online Panels Benchmarking Study (OPBS) was designed. The 2015 OPBS administered the same questionnaire to eight samples, made up of three probability samples and five nonprobability samples collected from volunteer/access/opt-in online panels. For two of the probability-based samples, a dual-frame telephone sampling methodology was employed, while the third used an address-based sampling frame. Each sample aimed to achieve approximately six hundred completed interviews (Pennay et al. 2018). The design was similar to the US study by Yeager et al. (2011) which compared the accuracy of seven online samples and two probability samples.

#### 9.3.1.2 Life in Australia™ – probability-based online panel: OPBS Replication (2017 OPBS)

Life in Australia™ is a probability-based internet panel for the Australian general adult population. Panellists are recruited via their landline or mobile phones to take part in incentivized monthly surveys. Participants receive a small reward to join the panel, and receive payments of $10-$15 for each survey they complete (with an option to donate to charity). Since the recruitment of panellists was through probability-based dual-frame sampling, the results from the surveys are generalizable to the Australian population (Kaczmirek et al. 2019), which means that sampling errors and confidences intervals can be derived. The recruitment was completed in December 2016, and the final sample obtained was 3,322 panellists. The overall recruitment rate as a product of recruitment and profile rates was 15.5% (Kaczmirek et al. 2019).

In January-February 2017, all active Life in Australia™ panellists were asked to participate in the replication of the OPBS; the Social Research Centre administered the same questionnaire used for the original 2015 OPBS. This was the second wave of Life in Australia™. To take into account the population with no access to the internet, the study also contacted panel members who happened to be offline (roughly 450 members). The completion rate was 78%, and 2,580 total interviews were achieved from panellists. We refer to this as the Online Panel Benchmarking Study Replication (2017 OPBS). Table 9.2 provides a description of the two datasets used in the study.

Table 9.2: Data files used

| Title of the study | Sampling | Data collection period | Mode of data collection | Total sample size | Data DOI |
|---|---|---|---|---|---|
| Online Panels Benchmarking Study (2015 OPBS) (Pennay et al. 2016) | Probability and nonprobability | June 2015 | Online, telephone, postal | n=4,757 | 10.4225/87/FSOYQI |
| Online Panels Benchmarking Study Replication (2017 OPBS) *(Life in Australia™ Wave 2)* (Pennay & Neiger 2020) | Probability | January-February 2017 | Online, telephone | n=2,580 | 10.26193/YF8AF1 |

### 9.3.2 Population, sampling and samples

Both the 2015 and 2017 OPBS surveys collected information from an in-scope population of all Australians aged 18 years and over. The studies were carefully designed to assess accuracy of nonprobability online panel samples relative to probability-based surveys using different probabilistic sampling methodology through using the same data collection instrument to provide data on the demographic, social characteristics and wellbeing of people in Australia (Kaczmirek et al. 2019; Pennay et al. 2018).

The OPBS 2015 study data comprised three probability-based samples: (i) an address-based sampling (A-BS) survey with Geocoded National Address File (G-NAF) as a sampling frame, (ii) a standalone dual-frame Random Digit Dialing (RDD) survey sample, and (iii) a RDD end-of-survey recruitment sample, also known as 'piggybacking' survey sample (Tourangeau & Smith 1985).

For the purpose of the 2015 OPBS study, five nonprobability online panels collected data from about 600 of their panellists each. We will analyze accuracy of the whole nonprobability sample combined[61] (n=3,058) and for two purposely selected nonprobability samples, the most and the least accurate. The OPBS Replication 2017 survey comprised of one probability-based mixed-mode (online, telephone) sample with a cumulative response rate as a product of overall recruitment and survey completion rates of 12.2%. Generally speaking, there were notable differences in response between the subsamples listed in Table 9.3, which might result in different levels of nonresponse error. The hope is that we can mitigate against this in our analysis through effective post-survey adjustment procedures.

[61] Combining data from several volunteer panels can increase their overall accuracy (Cornesse et al. 2020) and can be thus considered a solution to mitigate representation bias in nonprobability surveys, and is as such a subject of this study. We were particularly interested in the effectiveness of post-survey adjustment on combined data from different nonprobability sources, in comparison to individual opt-in panel samples.

Table 9.3: Studies and subsamples analyzed

| Study | Subsample | Response rate | n |
|---|---|---|---|
| Online Panels Benchmarking Study (2015 OPBS) | Address-based sampling | 26.2% | 538 |
| | Standalone RDD (dual-frame) | 14.7% | 600 |
| | RDD “piggybacking” (dual-frame) | 9.8% | 560 |
| | 5 volunteer panel samples* | 2.6%-15.4%** | 3,058 |
| Online Panels Benchmarking Study Replication (2017 OPBS) | Life in Australia™ Wave 2 | recruitment: 15.5%, survey completion 78.6% | 2,580 |

*Besides the combined nonprobability sample, we will analyze data separately for the most accurate panel (Panel 3, n=601) and the least accurate panel (Panel 1, n=601) (based on the results from Kaczmirek et al. 2019, p. 25). We will not analyze data for all 5 nonprobability panels separately due to space constraints. However, through comparing the best and worst performing nonprobability panel, we can get an indication of the variation in the bias and accuracy of different panel providers. **For nonprobability samples, response rates cannot be calculated and sample yield is reported instead (Pennay et al. 2018).

### 9.3.3 Benchmarks

To replicate benchmarking analysis from Pennay et al. (2018) and Kaczmirek et al. (2019), we will use the same benchmarks but from updated data sources collected closer in time to 2015 OPBS and 2017 OPBS studies. The sources of benchmarks are listed in Table 9.4. As mentioned previously, census data from the Australian Census 2016, electoral registration information from the Australian Electoral Commission, and social and health characteristics from the government funded surveys are considered as the best quality sources of nationally representative benchmarks in Australia with the highest validity. Benchmarks will be divided into primary, secondary demographics and substantive items (see Pennay et al. 2018). Table 9.4 provides a description of the benchmarks used in the study.

Table 9.4: Benchmarking data sources and nationally representative benchmarks

| **Study** | **Data collection mode** | **Sample size** | **Benchmarks** ***[a]primary demographics, [b]secondary demographics, [c]substantive items (non-demographics)*** |
|---|---|---|---|
| Australian Census 2016 | self-administered online, F2F | N=23,401,892 persons | Age (in categories) [a]<br>Gender[a]<br>State[a]<br>Residence in state capital city[a]<br>Country of birth[a]<br>Australian citizenship[b]<br>Employment status[b]<br>Home ownership[b]<br>Indigenous status[b]<br>Language other than English[b] |
| National Drug Strategy Household Survey (NDSHS) 2016 | self-administered paper-based or online, CATI | n=23,749 persons | Household status[b]<br>Daily smoker[c]<br>Alcoholic drink of any kind in the past 12 months[c] |
| National Health Survey 2014-15 | F2F | n=19,259 persons | Psychological distress (Kessler 6)[c]<br>General health[c]<br>Private health insurance[c]<br>Wage and salary income[b] |
| General Social Survey 2014 | F2F | n= 12,932 persons | Life satisfaction[c] |
| Australian Electoral Commission (2015) | administrative data | N=16,405,465 persons | Enrolled to vote[b] |

F2F– face-to-face; CATI – Computer-assisted telephone interviewing

## 9.3.4 Data analysis

### 9.3.4.1 Benchmarking analysis

To carry out our benchmark analysis, we need to balance against variance and bias in the final estimates. There are a wide variety of measures estimating the bias, such as the number of statistically significant differences from the benchmarks, the average absolute error (AAE) (including measures of uncertainty of the AAE, such as the standard deviation of the AAE or the range and ranking) (see Dutwin & Buskirk 2017; MacInnis et al. 2018; Yeager et al. 2011). To provide a measure of the variance, we compute the mean squared error. The mean square error is a function of both the bias and the variance, and as such it is a good measure of the overall accuracy of the different approaches. It is usual practice to take the square root of the mean square error (RMSE) to mitigate against the undue influence of extreme values. The aim of the study is to find the approach which is robust under the different scenarios. As such we present results using the AAE and RMSE to give an absolute measure of the error and the variability measure of the error, respectively.

The AAE was used by Yeager et al. (2011) to compare impact of different weighting approaches for probability and nonprobability surveys in the US. The same measure was used by Pennay et al. (2018) and Kaczmirek et al. (2019), who replicated the study design in Yeager et al. (2011) for Australia.

Our study follows all three of these previous studies, and the AAE is calculated as presented in Equation 9.2:

$$AAE = \sum_{j=1}^{k} \frac{|\widehat{y_j} - y_j|}{k} \qquad (9.2)$$

where $\widehat{y_j}$ is the j-th estimate (of a survey item) and $y_j$ is the value for a corresponding (population) benchmark.

And similarly, the RMSE is computed as presented in Equation 9.3:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{k} (\widehat{y_j} - y_j)^2}{k}} \qquad (9.3)$$

where *k* is the number of benchmarks, $\widehat{y_j}$ is again the j-th estimate from either OPBS surveys, and $y_j$ is the value for a corresponding benchmark.

To explore the generalizability of these findings, we calculate AAE and RMSE for 12 secondary demographics, 6 substantive items, and all 18 survey items with corresponding benchmarks combined. Most probability and nonprobability surveys apply adjustment for primary benchmarks as a standard approach, and for the majority of surveys the differences between the sample and population for primary benchmarks is expected to be minimal (Cornesse et al. 2020; Mercer et al. 2017). Therefore, we focus our analysis for the secondary demographics and substantive items, and explore the bias and variability relative to the different nationally representative (secondary and substantive) benchmarks.

The analysis was facilitated by the statistical coding environment and language R (R Core Team 2020) to carry out all data processing, post-survey adjustments, imputation of missing values[62] and benchmarking analyses. Besides R base or stats packages, the following packages were used: *Hmisc* (Harrell et al. 2020), *missForest* (Stekhoven 2013), *fastDummies* (Kaplan 2020), *anesrake* (Pasek 2018), *sjstats* (Lüdecke, D. 2020), *questionr* (Barnier et al. 2020), *MatchingFrontier* (King et al. 2015), and *cem* (Iacus et al. 2020).

#### 9.3.4.2 Post-survey adjustment approaches and parameters

To improve inference in nonprobability samples, we will test a number of post-survey adjustment methods and techniques:

[62] For matching and calibration only, and not for estimation.

- raking[63]
- generalized regression estimation (GREG)
- multilevel regression and poststratification (MRP)
- propensity score weighting (PSW)
- Mahalanobis distance matching (MDM)
- coarsened exact matching (CEM).

For more information about each of these methods, see Subsection 9.2.2. Post-survey adjustment details are presented in Table 9.5.

[63] In probability samples, a two-stage process can be used for weighting, first calculating a design weight (for the unequal probability of sample members being selected) and second raking (to reduce possible nonresponse). As the same process cannot be used for weighting nonprobability samples, and as the findings on the accuracy of nonprobability samples would not change (see Kaczmirek et al. 2019), we used a consistent one-stage raking approach across all samples.

Table 9.5: Post-survey methods, items, and parameters

| Method | Scenario | Covariates | Source of covariates | Other post-survey adjustment characteristics |
|---|---|---|---|---|
| **Raking** | Scenario 1 | Primary demographics(1) | Australian Census 2016 *(Australian Bureau of Statistics 2016)* | We applied weight trimming to ensure that the maximum weight after post-survey adjustment was 5. |
| | Scenario 2 | Primary demographics(1), secondary demographics (additional covariates with the largest absolute error) | Australian Census 2016 | |
| | Scenario 3 | Primary demographics1), all matching additional covariates from a large-scale survey | Australian Census 2016, National Drug Strategy Household Survey 2016 *(Hewitt 2017)* | |
| **GREG** | Scenario 1 | Primary demographics(1) | Australian Census 2016 | |
| | Scenario 2 | Primary demographics(1), secondary demographics (additional covariates with the largest absolute error) | Australian Census 2016 | |
| **MRP** | Scenario 1 | Primary demographics(1) | Australian Census 2016 | |
| **CEM** | Scenario 3 | All matching covariates from a large-scale survey | National Drug Strategy Household Survey 2016 | Pruning(3) of maximum 50% of all nonprobability sample units, later adjusted to match primary demographic(1) benchmarks |
| | Scenario 4 | Selected with dominance analysis(2) out of all available matching covariates (based on logit regression) | OPBS 2017 Replication sample *(Pennay & Neiger 2020)* | |
| **PSW** | Scenario 4 | All available matching covariates (excluding those with corresponding benchmarks) | OPBS 2017 Replication sample | Adjusted to match primary demographic(1) benchmarks |
| **MDM** | Scenario 4 | All available matching covariates (excluding those with corresponding benchmarks) | OPBS 2017 Replication sample | The same matched sample sizes as for CEM (Scenario 4), later adjusted to match primary demographic(1) benchmarks |

(1) Primary demographics were gender, age, state, capital city in state, education, country of birth and interaction effects between age and education, and state and capital city in state.

(2) Dominance analysis is used to compare the relative importance of predictors in regression models by comparing $R^2$ or Pseudo $R^2$ coefficient with different ranges of selected predictors (Budescu 1993). In practice, with dominance analysis we can select the covariates which distinguish probability and nonprobability samples the most in a multivariate setting.

(3) Removing those units from nonprobability data which cannot be matched with any unit from a probability sample.

The literature explains that the selection of covariates should be based on the relationship with nonresponse and noncoverage while preserving validity of the sample by including core demographics like gender; in the case of calibration, we also have to have in mind that selecting too many covariates

can lead to significant variance inflation and inability for raking algorithm to converge (Battaglia et al. 2009). In determining the optimal number of covariates (and their interactions), we used the Akaike Information Criteria (AIC) to select the minimum set of primary benchmarks, while maintaining the quality of poststratification information.

In the case of CEM with Australian Census benchmarks and good quality probability and nonprobability data for benchmarking (i.e., under Scenario 4), selecting many categorical covariates with a number of categories would lead to a low number of positive matches and substantial severe pruning. Carrying out dominance analysis (see Budescu 1993) will help us choose a limited number of predictors which explain the differences between probability and nonprobability samples best, and could be as such suitable covariates for post-survey adjustments. Based on the results of logistic regression (dependent variable *data source*: 0 – probability panel, 1 – nonprobability panels), we selected three predictors which distinguished the probability-based panel and nonprobability-based panels the most. This was done as the literature (e.g., Dutwin & Buskirk 2017) suggests using covariates that are associated with participation in nonprobability samples, in attempt to reduce errors associated with coverage (and to lesser extent nonresponse). In the end, we matched the data and had to prune almost 50% of all units from the combined nonprobability-based sample. We then used the nonprobability sample only, raked it, calculated new weighted estimates and compared them to the benchmarks.

To carry out MDM and PSW, we selected all matching covariates from the 2015 OPBS and 2017 OPBS, a total of 17[64]. Most of them were previously discussed in the literature as so-called 'webographic' variables[65]. For comparability purposes, we pruned the same portion of cases for MDM as for CEM.

## 9.4 Results

In this section, we will firstly present the results on the accuracy of nonprobability-based online panels before assessing different post-survey approaches to improving inference. We will update the results from Kaczmirek et al. (2019) by using the Australian Census 2016 benchmarks instead of Australian

[64] We acknowledge the fact that including other matching covariates (secondary demographics or non-demographic items with corresponding benchmarks) could help reduce more bias. However, we purposely excluded them to focus on other covariates from the data, including early adopter/openness to innovation items.

[65] Webographic variables are attitudinal or lifestyle variables accounting the difference between web survey participants and those who do not do surveys online (Baker et al. 2013). Different authors considered different questions as 'webographic' questions, such as: feeling alone, eagerness to learn new things, willingness to take chances, lifestyle questions (on travelling, participation in sports, reading a book), opinions on what is a violation of privacy, knowing a 'lesbian, gay, bisexual, transgender, and queer or questioning' (LGBTQ) person (Schonlau et al. 2007), early-adopter items (DiSogra et al. 2011; Dutwin & Buskirk 2017) or media use (Baker et al. 2013). On the other hand, Mercer et al. (2018) used political attitude variables in post-survey adjustments. In our study, besides early-adopter items, we also consider internet connection, access and use, and number of surveys completed as 'webographic' variables or, simpler, 'webographics'.

Census 2011 primary and secondary demographics. Second, we will assess different post-survey methodology under different data/covariate access scenarios.

### 9.4.1 Accuracy of nonprobability online panels

The results in this section will provide updated evidence (for original results see Kaczmirek et al. 2019) regarding the accuracy of nonprobability online samples in comparison to probability samples. The identified difference in accuracy will represent a reference for assessment of effectiveness of post-survey adjustments (see Section 9.4.2).

Table 7.6 presents the results on the accuracy of OPBS 2015 and OPBS 2017 Replication surveys. The results confirm the findings from Pennay et al. (2018) and Kaczmirek et al. (2019) on the accuracy of nonprobability-based online panels in comparison to probability samples.

First, nonprobability panel samples are similarly accurate in measuring secondary demographics as probability samples (AAE for nonprobability samples, raked: 4.7-5.4, AAE for probability samples: 4.2-5.3). However, they are less accurate in measuring non-demographics than probability surveys (AAE for nonprobability samples, raked: 6.6-9.9, AAE for probability samples: 3.7-5.4). This is the bias we would particularly like to reduce with various post-survey adjustments. In addition, we introduced the RMSE measure as a variability measure, and the difference between probability and nonprobability surveys in estimating non-demographic concepts is even more apparent if looking at that measure of accuracy, largely because of nonprobability panels overestimating psychological distress. In fact, after removing psychological distress from the benchmarks, the RMSE for substantive items (raked data) reduces to 3.4 (Panel 3) and 5.1 (combined opt-in sample), which is more comparable to RMSE of probability samples excluding psychological distress item – between 2.9 (Standalone RDD) and 3.2 (RDD piggybacking).

Second, raking as a post-survey adjustment method improves the quality of estimates from probability surveys more effectively than for nonprobability-based online panels. For the nonprobability surveys, raking can even deteriorate estimates (e.g., AAE for substantive items for Panel 1: unweighted 9.2, raked 9.9).

Third, both probability-based surveys and nonprobability online panels were more accurate in measuring socio-demographic characteristics than other concepts (substantive items), such as smoking and drinking habits, and especially psychological distress.

In this study, we also looked at the accuracy of five opt-in panels combined, as the theory suggests that combining nonprobability samples can improve accuracy. Based on the results, we can argue that

the combined sample is more accurate than four out of five individual samples, and as accurate as the most accurate nonprobability-based online panel of all five.

Our results differ from both those of Pennay et al. (2018) and Kaczmirek et al. (2019), although there are broad similarities. We conjecture that the reason for this is that our results use the most recent census results of the Australian population for benchmarking, and therefore are a closer match to observed trends and behaviors.

Table 9.6: Accuracy of nonprobability online panels in comparison to probability-based samples

| | | Least accurate nonprobability panel (1) (n=601) | | Most accurate nonprobability panel (3) (n=626) | | 5 nonprobability panels combined (n=3,058) | | Life in Australia™ (n=2,580) | | A-BS (n=538) | | RDD piggybacking (n=560) | | Standalone RDD (n=601) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Survey item** | **Benchmark (%)** | UW | Basic raking | UW | Basic raking | UW | Basic raking | UW | Basic raking | UW | Basic raking | UW | Basic raking | UW | Basic raking |
| Australian citizen | **87.1** | 5.9 | 3.2 | 6.0 | 3.9 | 5.3 | 3.0 | 4.5 | 0.4 | 7.3 | 4.3 | 5.2 | 0.9 | 3.9 | -1.2 |
| Couple with dependent children | **30.3** | -3.5 | -4.5 | -3.3 | -3.3 | -2.5 | -3.1 | -6.7 | -2.8 | -9.3 | -2.9 | -6.9 | -2.3 | -7.5 | -0.7 |
| Currently employed | **61.6** | -10.5 | -11.5 | -7.6 | -7.9 | -8.5 | -9.7 | 0.2 | 4.7 | -4.2 | 4.9 | -1.1 | 7.7 | -3.4 | 7.4 |
| Enrolled to vote | **78.5** | 8.4 | 5.0 | 10.0 | 8.0 | 9.4 | 6.7 | 11.7 | 6.3 | 14.1 | 8.9 | 11.9 | 6.1 | 9.7 | 3.6 |
| Home ownership with a mortgage | **28.8** | 3.0 | 1.5 | 5.1 | 4.9 | 3.5 | 2.5 | 1.3 | 1.9 | 3.5 | 9.9 | 4.8 | 7.8 | 2.1 | 5.2 |
| Not Indigenous | **97.7** | -0.2 | -0.6 | 0.7 | 0.7 | 0.2 | 0.1 | 0.0 | -0.4 | 0.3 | 0.5 | 0.3 | 0.3 | 0.1 | -0.1 |
| Language other than English (speak only English) | **76.5** | 6.2 | 4.2 | 9.1 | 6.9 | 6.8 | 4.7 | 8.7 | 3.9 | 4.7 | -0.4 | 10.5 | 3.4 | 7.7 | 3.1 |
| Living at last address 5 years ago | **56.9** | 4.7 | 3.7 | 7.3 | 8.2 | 6.9 | 5.9 | 6.2 | -1.1 | 12.2 | 0.2 | 10.6 | 3.1 | 12.7 | 5.4 |
| Most disadvantaged quintile for area-based SES | **20.0** | -3.2 | -2.6 | -5.6 | -6.7 | -5.4 | -5.2 | -7.3 | -6.8 | -5.5 | -4.2 | -8.2 | -9.2 | -5.0 | -5.1 |
| Resident of a major city | **66.8** | 9.2 | 6.4 | 1.3 | 2.1 | 4.9 | 3.4 | 2.9 | 4.0 | 5.9 | 7.5 | 2.3 | 4.1 | 2.3 | 2.8 |
| Voluntary work (none) | **79.4** | -6.9 | -5.9 | -8.3 | -8.7 | -7.9 | -8.1 | -20.7 | -17.8 | -18.6 | -17.9 | -19.2 | -18.3 | -21.2 | -18.3 |
| Wage and salary income $1000–1249 pw | **13.8** | -6.4 | -7.5 | -4.0 | -3.9 | -4.0 | -4.4 | -1.5 | -0.8 | -0.4 | 1.3 | -0.6 | 0.1 | -0.6 | 0.7 |
| Consumed alcohol in last year | **80.6** | -1.1 | -1.4 | -3.3 | -3.0 | -2.6 | -3.4 | 4.4 | 4.4 | 1.9 | 3.8 | 3.9 | 3.6 | 1.6 | 2.4 |
| Daily smoker | **13.1** | 8.7 | 9.0 | 4.2 | 2.6 | 4.8 | 4.3 | -3.0 | -1.7 | -4.0 | -2.8 | -0.6 | 2.0 | -2.8 | -0.4 |
| General health status (very good) | **36.2** | -2.9 | -3.5 | -3.6 | -4.3 | -2.6 | -2.7 | -2.6 | -3.5 | -1.8 | -1.6 | -2.4 | -3.5 | -5.6 | -3.8 |
| Life satisfaction (8 out of 10) | **32.6** | -10.8 | -11.1 | -5.1 | -3.7 | -9.0 | -9.0 | 0.1 | -0.8 | -2.5 | -3.8 | -1.4 | -3.4 | 2.0 | 0.4 |
| Has private health insurance | **57.1** | -4.0 | -7.1 | -3.7 | -3.1 | -2.4 | -3.7 | 9.2 | 4.2 | 10.2 | 3.2 | 8.1 | 3.0 | 8.5 | 4.6 |
| Psychological distress, Kessler 6 (low) | **82.2** | -27.7 | -27.6 | -22.9 | -22.6 | -24.5 | -25.0 | -13.6 | -17.7 | -6.1 | -14.8 | -6.6 | -11.9 | -8.2 | -10.3 |
| **AAE (combined)** | | 6.9 | 6.5 | 6.2 | 5.8 | 6.2 | 5.8 | 5.8 | 4.6 | 6.3 | 5.2 | 5.8 | 5.0 | 5.8 | 4.2 |
| **RMSE (combined)** | | 9.0 | 8.8 | 7.8 | 7.5 | 8.0 | 7.8 | 7.9 | 6.8 | 7.9 | 7.0 | 7.6 | 6.7 | 7.7 | 6.0 |
| **AAE (secondary demographics)** | | 5.7 | 4.7 | 5.7 | 5.4 | 5.4 | 4.7 | 6.0 | 4.2 | 7.2 | 5.2 | 6.8 | 5.3 | 6.3 | 4.5 |
| **RMSE (secondary demographics)** | | 6.3 | 5.5 | 6.4 | 6.0 | 6.0 | 5.4 | 8.2 | 6.3 | 8.9 | 7.2 | 8.7 | 7.2 | 8.6 | 6.5 |
| **AAE (substantive items)** | | 9.2 | 9.9 | 7.1 | 6.6 | 7.6 | 8.0 | 5.5 | 5.4 | 4.4 | 5.0 | 3.8 | 4.6 | 4.8 | 3.7 |
| **RMSE (substantive items)** | | 12.8 | 13.1 | 10.0 | 9.7 | 11.0 | 11.2 | 7.1 | 7.8 | 5.3 | 6.7 | 4.7 | 5.6 | 5.6 | 5.0 |

UW – unweighted results, Basic raking (to primary demographics) – by gender, age group*education, country of birth, state*capital city in state, AAE - average absolute error, RMSE - root mean squared error

### 9.4.2 Assessment of effectiveness of post-survey adjustment methods for improving inference in nonprobability samples

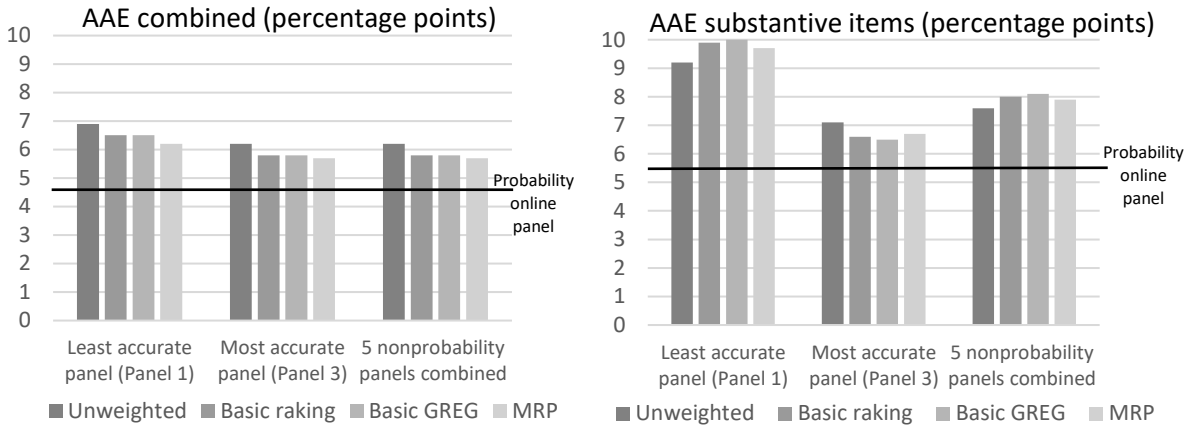
In this section, we will show if the difference in accuracy between probability and nonprobability samples, i.e., representation bias, can be reduced using different post-survey adjustment methods. The results will be presented by scenarios based on the availability of external data. For example, if only aggregated census data are available (e.g., in TableBuilder (Australian Bureau of Statistics, n.d.)), we are limited with the post-survey adjustment methodology that can be carried out with tabular data, such as calibration. If we have access to other representative data sources with available benchmarks and/or unit record data, we have additional adjustment solutions, such as matching methods and PSW. We will use Life in Australia™ Wave 2 as a reference sample for post-survey adjustment efficiency (AAE secondary demographics 4.2, AAE substantive items 5.4, AAE combined 4.6, all raked).

#### 9.4.2.1 Scenario 1: Availability of census aggregated statistics, and only primary demographics[66] were collected from the nonprobability sample

Under this scenario, the only variables that are matched in any other external data source are primary demographics from population censuses. To improve inference, primary demographics in an aggregated form can be used with different weighting/calibration methods – in this study, we are assessing the efficiency of raking, GREG and MRP. It is assumed that nonprobability opt-in panels would normally carry out post-survey adjustments under this scenario. To illustrate the effectiveness of post-survey adjustments using primary demographics, we are presenting results for unweighted and weighted data for the nonprobability online samples, comparing (1) the least accurate, (2) the most accurate, and (3) the combined sample (from all five samples) (see Figure 9.1, and Table 9.7 from the Appendix 9 provides more detailed results). For comparison, we also show in the graph the estimate from the probability online panel. Since benchmarking to substantive items generally increases the bias, we examine the AAE for combined and substantive benchmarks.

[66] Primary demographics in the Australian context are benchmarks commonly used in poststratification weighting: gender, age, education, country of birth (Australia, not Australia), and geography (state and capital city in state). In contrast to core demographics, secondary demographics (classification: Pennay et al. 2018; Kaczmirek et al. 2019) are presented in Table 9.4.

Figure 9.1: Average absolute error (AAE) for estimates, un- and weighted (raking, GREG, MRP)[67]




The results from Figure 9.91 show how basic weighting post-survey adjustments improve the quality of estimates, but the improvement is only slight on average (AAE combined reduction between 0.4 [GREG, Panel 3] and 0.7 [MRP, Panel 1]). We can confirm our previous finding on how raking improves the accuracy of nonprobability samples to a lesser extent than those from probability samples. We can also extend this finding to other calibration methods studied in this article – GREG and MRP.

The improvement in accuracy is more apparent for all 18 survey items combined than for six substantive items combined, which indicated that calibration using primary demographic more consistently improves the quality of secondary demographic estimates than non-demographic estimates. Moreover, the results from Figure 9.1 show how calibration can deteriorate substantive item estimates from nonprobability samples, especially the least accurate one, but also the combined volunteer panel sample. This is consistent across all calibration methods, with MRP performing just slightly better than GREG and raking. On the other hand, weighting improved accuracy of the most accurate nonprobability panel in a similar fashion for both secondary demographics and non-demographics.

We have to note that the differences in item-level results (not only at the AAE level, see Table 9.7) are almost non-existent for raking and GREG and very little between the first two calibration methods and MRP. Using a limited number of primary demographic covariates in weighting schemes, we cannot expect major differences in weighted estimates no matter the calibration method chosen. Taking into account that MRP requires a joint distribution of all benchmarks and is computationally intensive, raking or GREG weighting seemed to be the more optimal calibration solution in our particular case.

For more detailed results, please see Table 9.7 in the Appendix 9.

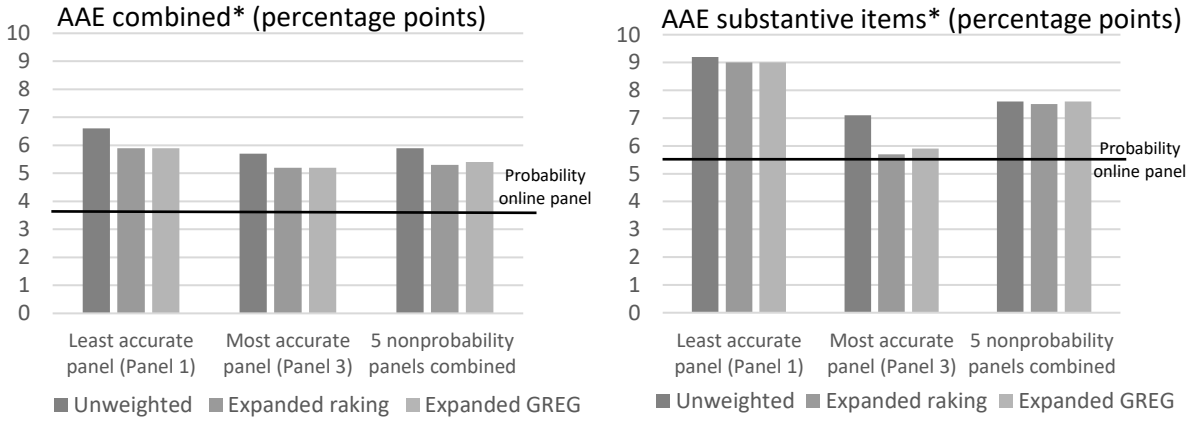
[67] AAE for secondary demographics and all RMSE calculations (combined, secondary demographics, and substantive items) are presented in the tables in the Appendix 9.

### 9.4.2.2 Scenario 2: Availability of census aggregated statistics, both primary and secondary demographics[68] were collected from the nonprobability sample

Under this scenario, there are additional socio-demographic items with corresponding benchmarks available for calibration that are matching in the nonprobability-based online panel data. Thus, post-survey adjustment methods like raking and GREG can be carried out. Due to the limitations of MRP identified in our previous analysis, we will assess the efficiency of the first two calibration methods only.

To further improve accuracy of nonprobability sample estimates, we reviewed all additional covariates from a set of secondary demographics listed in Table 9.4. We purposely selected those with the highest representation bias, i.e., items with the largest absolute errors after raking with primary demographics only. The added covariates were: employment status, language other than English, and voluntary work (see Table 9.6).

Figure 9.2: Average absolute error (AAE) for estimates, unweighted and weighted (raking, GREG)




*AAE were calculated for all items excluding the secondary demographics included in an expanded calibration scheme (employment status, language other than English (LOTE), and voluntary work, see Table 9.8 in the Appendix 9 for more information)

The results from Figure 9.2 firstly show how including new covariates in calibration further improves the accuracy of nonprobability samples. We also did not notice a significant increase of design effect. The evidence (from Figure 9.2 and from Table 9.8 in the Appendix 9) suggests that expanded raking and GREG predominantly improved secondary demographics estimates and, in some cases, estimates from substantive items. For the most and the least accurate online panel, as well as all panels combined, we can see a slight improvement in the combined AAE and RMSE. Generally speaking, we

[68] Secondary demographics in the Australian context analyzed this study are the following benchmarks: Australian citizenship, household status, employment status, enrolled to vote, home ownership with a mortgage, Indigenous status, language other than English spoken at home, living at last address 5 years ago, most disadvantaged quintile for area-based socio-economic status (SES), resident of a major city, voluntary work, wage and salary income.

can again report almost negligible differences between estimates adjusted with expanded raking and expanded GREG.

Moreover, this time calibration did not increase AAE for substantive items for the least accurate panel and five panels combined. Including three secondary demographic covariates seemed to eliminate the negative effect of raking with primary demographics only. Moreover, we can notice a notable improvement in accuracy of substantive items after using an expanded raking scheme for the most accurate nonprobability panel (AAE=7.1 unweighted and AAE=5.7 raking, AAE=5.9 GREG). The selected secondary demographic items seem to be more associated with representation bias in nonprobability online panels than our core/primary demographics.

The evidence from Figures 9.1 and 9.2 suggests that the highest-quality nonprobability online panels are not only the most accurate for unweighted estimates, but they also respond better to various calibration adjustments. Also, including additional secondary demographic covariates produced better estimates in comparison to basic calibration, albeit the overall difference is small. Often, the primary demographics are sufficient in calibration (Kalton & Flores-Cervantes 2003), similar to the Occam's razor principle that the simplest solution is usually the best.

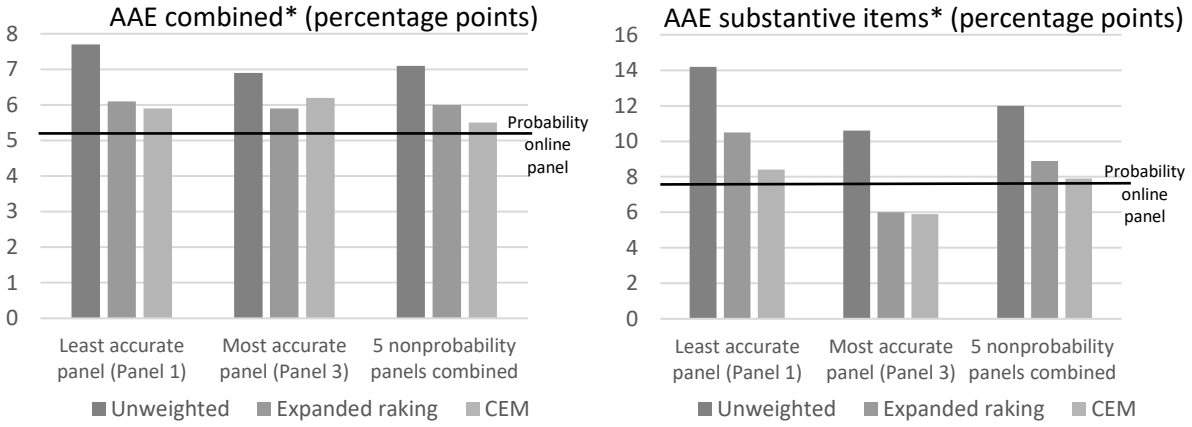
#### 9.4.2.3 Scenario 3: Availability of census aggregated statistics and one other representative source of benchmarks

Under this scenario, a survey statistician would have access to an additional external high-quality data source, either aggregated tabular data or unit record data/microdata (with matching covariates in the nonprobability data file). An example of that would be a large-scale non-population-census government survey producing representative and accurate benchmarks. An advantage of having access to a data source of that kind would be an ability to use non-demographic covariates that are somewhat associated with representation bias in nonprobability samples. If having access to unit record data, additional post-survey adjustment methods like CEM could be used.

In our case, we additionally included four covariates from the National Drug Strategy Household Survey (NDSHS) 2016 (Hewitt 2017), i.e., one secondary demographic and three substantive items, and carried out expanded raking (also with primary demographics). For CEM, we chose the same four covariates plus gender and age*education (see Table 9.4). These were essentially all matching covariates in nonprobability surveys and NDSHS 2016. In the end, primary demographics were adjusted to population totals using census benchmarks after matching with CEM. With this post-survey adjustment including the same covariates, we were able to directly compare the efficiency of calibration and matching methods. As GREG produced very similar results to raking under Scenarios 1

and 2, we decided to use expanded raking as the only calibration method. While we could use both distance-based PSW and MDM with large-scale survey microdata, the advantage of those two methods is the ability to include a high number of covariates (see Table 9.4).

Figure 9.3: Average absolute error (AAE) for estimates, unweighted and adjusted post-survey (raking, CEM)




*AAE were calculated for all items excluding the covariates in an expanded post-survey adjustment scheme (household status, frequency of smoking, and drinking alcohol, see Table 9.9 in the Appendix 9 for more information)

The results from Figure 9.3 show how including new non-demographic covariates in calibration improves the accuracy of nonprobability samples fairly similarly to including new secondary demographic covariates. However, the improvement is more substantial – an increase in accuracy measured with AAE combined ranges from 1.0 (Panel 3, raking) to 1.8 (Panel 1, CEM).

In comparison to the efficiency of calibration under Scenario 2, including non-demographic covariates improved the accuracy of substantive items[69] to a greater extent. The decrease in that AAE ranged between 3.1 (5 panels combined) and 4.6 (Panel 3, the most accurate panel). Post-survey adjustment with CEM also improved accuracy of nonprobability samples in measuring substantive items, and the decrease in AAE was even more significant – between 4.1 (5 panels combined) and 5.8 (Panel 1, the least accurate panel) for the remaining three non-demographic items.

Post-survey adjustment under Scenario 3 made nonprobability online panels almost as accurate as a probability-based online panel overall (AAE combined), especially after CEM. For three non-demographics items, the most accurate nonprobability online panel was even more accurate than the probability panel, no matter the post-survey adjustment method.

While CEM compares favorably to expanded raking using covariates from a large-scale survey, we noticed a larger design effect than for expanded raking. Since the NDSHS sample was large compared

[69] The remaining 3 substantive items were from National Health Survey 2014-15 and General Social Survey 2014 (see Table 9.4).

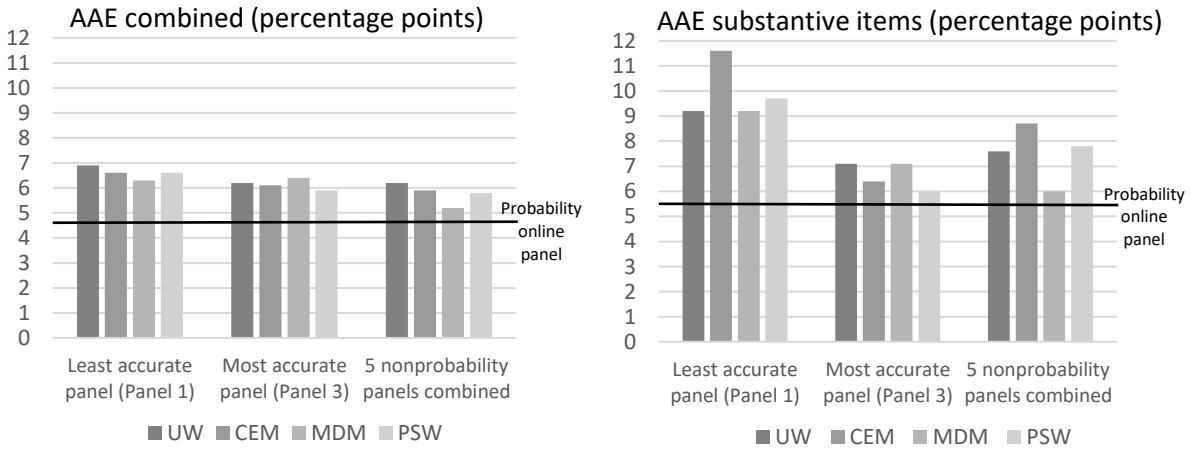
to nonprobability samples, a smaller portion of nonprobability samples was pruned (about 10%) and CEM weights primarily balanced the samples. Those weights were later used as base weights in raking, which resulted in larger design effect.

#### 9.4.2.4 Scenario 4: Availability of census aggregated statistics and a smaller-scale probability-based survey data with matching variables from nonprobability-based survey data

Under this scenario, a smaller-scale external survey data source is available in a unit record data form. Thus, a variety of methods and their combinations is possible, including calibration such as raking, GREG and MRP. However, calibration is normally carried out with benchmarks from the highest-quality large-scale surveys, and smaller-scale probability-based survey tend to introduce more error (see Table 9.6). In this post-survey adjustment exercise, we will use Life in Australia™ Wave 2 data due to its sample size (n=2,580) and with AAE and RMSE values comparable to those of the other probability-based samples. In this analysis, we will compare three methods: (1) CEM, (2) MSM, and (3) PSW.

The main difference between Scenario 3 and Scenario 4 is the range of available covariates. Under Scenario 3, we had to select all matching covariates, including four of which were initially used to compare accuracy of different samples. Under Scenario 4, we have access to a less representative external data source, but there are additional covariates which could be used to balance the samples as noted in the literature by authors such as Schonlau et al. (2007), Baker et al. (2013) or Dutwin and Buskirk (2017). In our case, 'webographics' represent early adopter score, number of surveys completed, internet connection types, internet and social media use. For that reason, no survey items with corresponding benchmarks (all listed in Table 9.10) had to be used in post-survey adjustment schemes.

Figure 9.4: Average absolute error (AAE) for estimates, unweighted and adjusted post-survey



The results from Figure 9.4 present mixed evidence on the efficiency of post-survey adjustment methods using smaller-scale external survey data with no demographics or health-associated items. First, there was a fairly moderate and inconsistent effect of post-survey adjustments on the total accuracy of nonprobability samples. In most cases, the decrease of AAE combined was less than 0.5, and no method seemed to have a clear advantage. The only exception to the rule was MDM with the data from five nonprobability-based panels combined (AAE unweighted 6.2, AAE MDM 5.2, probability panel 4.6).

Comparing AAE for substantive items, we can observe as many instances of post-survey adjustment deteriorating estimates as instances of improving estimates. The least accurate nonprobability-based panel stands out as the sample with no decrease in AAE before or after adjustment, and CEM as the method with limited efficiency for only one sample (the most accurate). The best result overall can again be attributed to MDM (AAE unweighted 7.6, MDM 6.5, probability panel 5.4), and we can also see a positive effect of PSW on the accuracy of Panel 3 (AAE unweighted 7.1, PSW 6.0, probability panel 5.4).

In comparison to the first three scenarios, there is less consistency in the effectiveness of post-survey adjustments with smaller-scale survey data and 'webographics' and other internet-related variables. This is consistent with findings from Dutwin and Buskirk (2017). One of the reasons could be that smaller-scale surveys come with some error, and a second reason could be that 'webographic' variables might not be as associated with the outcome variables analyzed in this study as required to mitigate representation bias. The same issue was reported by Mercer et al. (2018) who found mixed effects of including political attitudes on improving accuracy of non-political estimates.

## 9.5 Discussion and conclusion

This investigation into improving inference in nonprobability sample surveys supports the conclusion that the issue of improving inference in nonprobability sample surveys is a three-dimensional problem. First, the quality of post-survey adjustments is dependent on the availability of relevant high-quality covariates which are associated with either representation bias in nonprobability samples and/or outcome variables. Second, as the covariates in nonprobability samples should have matching covariates in external representative data sources, the availability and ability to access auxiliary data is a key aspect in mitigating bias. Third, the efficiency of post-survey adjustments is also dependent on the selection and combination of post-survey adjustment methods, albeit to a lesser extent.

In this study, we presented evidence that post-survey adjustment can reduce representation bias in nonprobability online samples to some extent, but cannot consistently eliminate it. These findings are in line with evidence from Tourangeau et al. (2014). However, we demonstrated a greater potential

to mitigate representation bias in nonprobability panels if having access to more external data sources and more covariates matching in nonprobability samples and auxiliary data. Ideally, we would have access to large-scale survey microdata, since smaller-scale surveys come with some nonignorable error. While those probability surveys mostly remain more accurate than nonprobability surveys even after post-survey adjustments, they are more susceptible to coverage, sampling, and nonresponse error (or even measurement mode effect) than most high-quality government surveys, and the total representation error can be carried over to post-survey adjustment results (e.g., after matching or PSW). For that reason, improving inference in nonprobability samples should be planned in the survey design stage, and relevant external data sources reviewed before data collection, if possible.

Moreover, identification of covariates from external data sources which are associated with representation bias or target outcome variables can lead to a more efficient mitigation of bias. While post-survey adjustments using primary demographics have little positive effect on the quality of nonprobability estimates, we have shown how including secondary demographics can improve the quality of other demographics and including non-demographics can decrease the error from associated non-demographics. This is consistent with findings from Bethlehem (2002). Similarly, Mercer et al. (2018) reported that including political attitude covariates in adjustment improved the quality of political engagement estimates. However, we found inconsistent evidence on the suitability of 'webographics' and other internet-associated covariates for mitigating bias in nonprobability samples. Unfortunately, we could not distinguish between the effect of those covariates and the effect of the data source on the post-survey adjustment efficiency. While auxiliary variables like early adopter items (traditionally used to mitigate bias in nonprobability samples, e.g., DiSogra et al. 2011) did not distinguish our probability online sample and nonprobability online panel samples well, we identified new covariates for post-survey adjustment (CEM) that could be considered as 'webographics', such as the number of surveys participated in. Therefore, we believe it is crucial to carry out more investigation into 'webographic' variables for post-survey adjustment, as previously suggested by Dutwin and Buskirk (2017).

On the other hand, the investigation into the suitability of post-survey adjustment methods did not highlight any particular method or a combination of them which consistently performed better parameter estimates. This supports the finding from Mercer et al. (2018). While a detailed technical investigation into calibration methods was not the focus of this article, we found little differences in efficiency between the studied methods: raking and the model-based methods (such as GREG or MRP). Therefore, we suggest the selection of calibration methods to be instead based on the availability of joint distributions of covariates weighed against the computational intensity of methods. While matching methods and PSW under limited scenarios might have a better potential for

efficient post-survey adjustment (for instance, under Scenario 3 using national health survey data), we observed less consistency in bias reduction between different samples and scenarios, as well as an increase of design effect and, consequently, confidence intervals for estimates (Kolenikov 2014).

This study has several limitations, including the availability of external data and covariates both in nonprobability surveys and high-quality government surveys. Having access to additional data sources could improve post-survey adjustments and potentially distinguish better between the efficiency of covariates and the efficiency of methods. Moreover, since estimates for only 18 items were compared to benchmarks and the majority of substantive items were more or less associated with one topic (i.e., health status), the findings would be more robust if survey items with corresponding benchmarks would be associated with other aspects of respondent's lives, not only health. We would suggest future research on improving inference in nonprobability samples to be more targeted, planned and properly designed in advance. Nonetheless, the approaches discussed in this chapter have distinct benefits in improving the inferences from surveys conducted using nonprobability samples.

## 9.6 References


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

Australian Bureau of Statistics. (2018a, March 28). *Household use of information technology*. https://www.abs.gov.au/statistics/industry/technology-and-innovation/household-use-information-technology/latest-release

Australian Bureau of Statistics. (2018b, December 12). *National Health Survey: First results.* https://www.abs.gov.au/statistics/health/health-conditions-and-risks/national-health-survey-first-results/latest-release

Australian Bureau of Statistics. (n.d.). *TableBuilder.* Retrieved November 1, 2020, from https://www.abs.gov.au/websitedbs/d3310114.nsf/home/about+tablebuilder

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

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology, 1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Barnier, J., Briatte, F., & Larmarange, J. (2020). *Questionr: Functions to make surveys processing easier* (R package version 0.7.1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=questionr

Battaglia, M. P., Hoaglin, D. C., & Frankel, M. R. (2009). Practical Considerations in Raking Survey Data. *Survey Practice, 2* (5). https://doi.org/10.29115/SP-2009-0019.

Bethlehem, J. G. (2002). *Weighting nonresponse adjustments based on auxiliary information*. Wiley.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching?. *Social Science Computer Review, 34*, 59-77.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817-848.

Bon, J. J., Ballard, T., & Baffour, B. (2019). Polling bias and undecided voter allocations: US presidential elections, 2004–2016. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 467-493.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329.

Budescu, D. V. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, *114*(3), 542.

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., de Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. Journal of Survey Statistics and Methodology, 8(1), 4–36. https://doi.org/10.1093/jssam/smz041

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, *64*(4), 464-494.

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

Couper, M. P., Gremel, G., Axinn, W., Guyer, H., Wagner, J., & West, B. T. (2018). New options for national population surveys: The implications of internet and smartphone coverage. *Social Science Research*, *73*, 221-235.

Daikeler, J., Bosnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513-539.

Datareportal. (2020, February 13). *Digital 2020: Australian.* https://datareportal.com/reports/digital-2020-australia

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376-382.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, *88*(423), 1013-1020.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social science research*, *38*(1), 1-18.

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

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

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice, 2,* 1-7.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249-264.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23,* 127–135.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153-164.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, *70*(5), 646-675.

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

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, *74*(5), 849-879.

Harrell, F. E. Jr, Dupont, C., & others (2020). *Hmisc: Harrell miscellaneous* (R package version 4.4-1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=Hmisc.

Harrison, S., Alderdice, F., Henderson, J., Redshaw, M., & Quigley, M. A. (2020). Trends in response rates and respondent characteristics in five National Maternity Surveys in England during 1995–2018. *Archives of Public Health*, *78*, 1-11.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–62.

Hewitt, M. (2017). *National Drug Strategy Household Survey 2016.* (ADA Dataverse, Version 7) [Data set]. ADA. https://doi.org/10.4225/87/JUDY2Y

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1-28.

Hug, S. (2003). Selection Bias in Comparative Research: The Case of Incomplete Data Sets. *Political Analysis, 11*(3), 255-274. https://doi.org/10.1093/pan/mpg014

Iacus, S. M., King G., & Porro, G. (2009). CEM: Coarsened Exact Matching Software. *Journal of Statistical Software, 30*. http://gking.harvard.edu/cem

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345-361.

Iacus, S. M., King, G., & Porro, G. (2020). *Cem: Coarsened exact matching (R package version 1.1.20)* [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=cem

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

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

Kaplan, J. (2020). *fastDummies: Fast creation of dummy (binary) columns and rows from Categorical Variables* (R package version 1.6.1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=fastDummies

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A comparison of web and mail survey response rates. *Public opinion quarterly*, *68*(1), 94-101.

Keiding, N., & Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 319-376.

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

Kennedy, C., & Hartig, H. (2019). *Response rates in telephone surveys have resumed their decline*. Pew Research Center.

King, G., Lucas, C., & Nielsen, R. A. (2015). *{MatchingFrontier}: {R} Package for Computing the Matching Frontier ()* [Computer software]. The Comprehensive R Archive Network. Available from http://projects.iq.harvard.edu/frontier

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal*, *14*(1), 22-59.

Leeper, T. J. (2019). Where have the respondents gone? Perhaps we ate them all. *Public Opinion Quarterly*, *83*(S1), 280-288.

Little, R. J. A., & Rubin, D. B. (2002). Single imputation methods. In R. J. A. Little, & D. B. Rubin (Eds.), *Statistical analysis with missing data* (pp. 59-74). Wiley.

Lüdecke, D. (2020). *Sjstats: Statistical functions for regression models* (version 0.18.0) () [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=sjstats

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

Matei, A. (2018). On Some Reweighting Schemes for Nonignorable Unit Nonresponse. *Survey Statistician, 77*, 21–33.

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*(S1), 250-271.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For weighting online opt-in samples, what matters most.* Pew Research Center.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, *12*(4), 375-385.

Park, D. K., Gelman, A., & Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In J. E. Cohen (Ed.), *Public opinion in state politics* (pp. 209-228). Stanford University Press.

Pasek, J. (2018). *Anesrake: Anes raking implementation* (R package version 0.80) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=anesrake

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

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

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

Pfeffermann, D., Eltinge, J. L., & Brown, L. D. (2015). Methodological issues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. *Journal of Survey Statistics and Methodology, 3*(4), 425-483.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rivers, D. (2007, July 29–August 2). *Sampling for Web Surveys* [Conference presentation]. 2007 Joint Statistical Meetings, Salt Lake City, United States of America.

Rivers, D. (2013). Comment on task force report. *Journal of Survey Statistics and Methodology, 1*(2), 111-117.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Validity in epidemiologic studies. In K. J. Rothman, S. Greenland, T. L. Lash (Eds.), *Modern epidemiology* (3rd ed.) (pp. 128-147). Lippincott Williams & Wilkins.

Schonlau, M., Soest, V. A., & Kapteyn, A. (2007). Are "Webographic" or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?. *Survey Research Methods*, *1*, 155–163.

Schonlau, M., & Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science*, *32*(2), 279-292.

Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*(522), 607-614.

Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest* (R package version 1.4) [Computer software]. The Comprehensive R Archive Network. Available from https://github.com/stekhoven/missForest

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155-176). Sage.

Stuart, E. A. (2010). The Use of Propensity Scores to Assess Generalizability. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 174*(2), 369–386.

Tourangeau, R., & Smith, W. (1985). Finding subgroups for surveys. *Public Opinion Quarterly, 49*(3), 351-365.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (2014). *Hard-to-survey populations.* Cambridge University Press.

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, *40*(1), 105-137.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231-263.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31*(3), 980-991.

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

## Appendix 9

Table 9.7: Estimates relative to the benchmarks, unweighted and weighted (raking, GREG, MRP)

| | | Least accurate nonprobability panel (1) (n=601) | | | | Most accurate nonprobability panel (3) (n=626) | | | | 5 nonprobability panels combined (n=3,058) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Survey item** | **Benchmark (%)** | **Unweighted** | **Basic raking** | **Basic GREG** | **MRP** | **Unweighted** | **Basic raking** | **Basic GREG** | **MRP** | **Unweighted** | **Basic raking** | **Basic GREG** | **MRP** |
| Australian citizen | **87.1** | 5.9 | 3.2 | 3.3 | 1.9 | 6.0 | 3.9 | 3.9 | 3.5 | 5.3 | 3.0 | 3.0 | 2.6 |
| Couple with dependent children | **30.3** | -3.5 | -4.5 | -4.4 | -4.0 | -3.3 | -3.3 | -3.2 | -2.8 | -2.5 | -3.1 | -3.1 | -2.9 |
| Currently employed | **61.6** | -10.5 | -11.5 | -11.6 | -11.5 | -7.6 | -7.9 | -7.9 | -7.9 | -8.5 | -9.7 | -9.6 | -9.5 |
| Enrolled to vote | **78.5** | 8.4 | 5.0 | 5.0 | 4.3 | 10.0 | 8.0 | 8.1 | 7.3 | 9.4 | 6.7 | 6.7 | 6.4 |
| Home ownership with a mortgage | **28.8** | 3.0 | 1.5 | 1.5 | 1.7 | 5.1 | 4.9 | 5.0 | 4.6 | 3.5 | 2.5 | 2.5 | 2.4 |
| Not Indigenous | **97.7** | -0.2 | -0.6 | -0.5 | -0.5 | 0.7 | 0.7 | 0.7 | 0.5 | 0.2 | 0.1 | 0.2 | 0.2 |
| Language other than English (speak only English) | **76.5** | 6.2 | 4.2 | 4.2 | 3.3 | 9.1 | 6.9 | 6.9 | 6.3 | 6.8 | 4.7 | 4.7 | 4.2 |
| Living at last address 5 years ago | **56.9** | 4.7 | 3.7 | 3.7 | 3.9 | 7.3 | 8.2 | 8.0 | 7.9 | 6.9 | 5.9 | 6.0 | 6.1 |
| Most disadvantaged quintile for area-based SES | **20.0** | -3.2 | -2.6 | -2.5 | -2.5 | -5.6 | -6.7 | -6.7 | -6.9 | -5.4 | -5.2 | -5.2 | -5.2 |
| Resident of a major city | **66.8** | 9.2 | 6.4 | 6.4 | 6.8 | 1.3 | 2.1 | 2.1 | 2.6 | 4.9 | 3.4 | 3.4 | 3.9 |
| Voluntary work (none) | **79.4** | -6.9 | -5.9 | -5.8 | -5.6 | -8.3 | -8.7 | -8.4 | -8.5 | -7.9 | -8.1 | -8.0 | -7.8 |
| Wage and salary income $1000–1249 per week | **13.8** | -6.4 | -7.5 | -7.4 | -7.4 | -4.0 | -3.9 | -3.8 | -3.9 | -4.0 | -4.4 | -4.4 | -4.5 |
| Consumed alcohol in last year | **80.6** | -1.1 | -1.4 | -1.4 | -1.4 | -3.3 | -3.0 | -2.9 | -3.1 | -2.6 | -3.4 | -3.4 | -3.2 |
| Daily smoker | **13.1** | 8.7 | 9.0 | 9.0 | 8.9 | 4.2 | 2.6 | 2.6 | 3.1 | 4.8 | 4.3 | 4.3 | 4.3 |
| General health status (very good) | **36.2** | -2.9 | -3.5 | -3.5 | -3 | -3.6 | -4.3 | -3.9 | -4.2 | -2.6 | -2.7 | -2.7 | -2.6 |
| Life satisfaction (8 out of 10) | **32.6** | -10.8 | -11.1 | -11.2 | -10.8 | -5.1 | -3.7 | -3.8 | -4.1 | -9.0 | -9.0 | -9.0 | -8.8 |
| Has private health insurance | **57.1** | -4.0 | -7.1 | -7.1 | -6.7 | -3.7 | -3.1 | -3.1 | -3.0 | -2.4 | -3.7 | -3.8 | -3.5 |
| Psychological distress, Kessler 6 (low) | **82.2** | -27.7 | -27.6 | -27.8 | -27.4 | -22.9 | -22.6 | -22.7 | -22.9 | -24.5 | -25.0 | -25.1 | -24.8 |
| **AAE (combined)** | | 6.9 | 6.5 | 6.5 | 6.2 | 6.2 | 5.8 | 5.8 | 5.7 | 6.2 | 5.8 | 5.8 | 5.7 |
| **RMSE (combined)** | | 9.0 | 8.8 | 8.8 | 8.6 | 7.8 | 7.5 | 7.4 | 7.4 | 8.0 | 7.8 | 7.8 | 7.7 |
| **AAE (secondary demographics)** | | 5.7 | 4.7 | 4.7 | 4.4 | 5.7 | 5.4 | 5.4 | 5.2 | 5.4 | 4.7 | 4.7 | 4.6 |
| **RMSE (secondary demographics)** | | 6.3 | 5.5 | 5.5 | 5.3 | 6.4 | 6.0 | 6.0 | 5.8 | 6.0 | 5.4 | 5.3 | 5.2 |
| **AAE (substantive items)** | | 9.2 | 9.9 | 10.0 | 9.7 | 7.1 | 6.6 | 6.5 | 6.7 | 7.6 | 8.0 | 8.1 | 7.9 |
| **RMSE (substantive items)** | | 12.8 | 13.1 | 13.2 | 12.9 | 10.0 | 9.7 | 9.7 | 9.9 | 11.0 | 11.2 | 11.3 | 11.1 |

Basic raking, GREG, MRP – by gender, age group*education, country of birth, state*capital city in state, AAE - average absolute error (percentage points), RMSE - root mean squared error (of percentage points)

Table 9.8: Estimates relative to the benchmarks, unweighted and weighted (expanded raking, expanded GREG)

| | | Least accurate nonprobability panel (1) (n=601) | | | Most accurate nonprobability panel (3) (n=626) | | | 5 nonprobability panels combined (n=3,058) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Survey item** | **Benchmark (%)** | **Unweighted** | **Expanded raking[#]** | **Expanded GREG[#]** | **Unweighted** | **Expanded raking[#]** | **Expanded GREG[#]** | **Unweighted** | **Expanded raking[#]** | **Expanded GREG[#]** |
| Australian citizen | **87.1** | 5.9 | 3.2 | 3.2 | 6.0 | 4.1 | 4.2 | 5.3 | 2.7 | 2.8 |
| Couple with dependent children | **30.3** | -3.5 | -4.0 | -4.1 | -3.3 | -3.4 | -2.9 | -2.5 | -3.1 | -3.1 |
| Currently employed | **61.6** | *-10.5* | *0.0* | *0.0* | *-7.6* | *0.0* | *0.0* | *-8.5* | *0.0* | *0.0* |
| Enrolled to vote | **78.5** | 8.4 | 4.7 | 4.8 | 10.0 | 8.6 | 8.8 | 9.4 | 6.6 | 6.7 |
| Home ownership with a mortgage | **28.8** | 3.0 | 2.7 | 2.6 | 5.1 | 7.2 | 6.9 | 3.5 | 4.0 | 4.0 |
| Not Indigenous | **97.7** | -0.2 | -0.4 | -0.5 | 0.7 | 0.8 | 0.8 | 0.2 | 0.3 | 0.3 |
| Language other than English (speak only English) | **76.5** | *6.2* | *0.0* | *0.0* | *9.1* | *0.0* | *0.0* | *6.8* | *0.0* | *0.0* |
| Living at last address 5 years ago | **56.9** | 4.7 | 3.9 | 4.0 | 7.3 | 7.2 | 6.7 | 6.9 | 5.4 | 5.4 |
| Most disadvantaged quintile for area-based SES | **20.0** | -3.2 | -2.2 | -2.0 | -5.6 | -6.8 | -6.5 | -5.4 | -5.2 | -5.4 |
| Resident of a major city | **66.8** | 9.2 | 6.4 | 6.4 | 1.3 | 2.4 | 2.6 | 4.9 | 3.3 | 3.3 |
| Voluntary work (none) | **79.4** | *-6.9* | *0.0* | *0.0* | *-8.3* | *0.0* | *0.0* | *-7.9* | *0.0* | *0.0* |
| Wage and salary income $1000–1249 per week | **13.8** | -6.4 | -7.0 | -6.8 | -4.0 | -3.6 | -3.4 | -4.0 | -3.9 | -3.8 |
| Consumed alcohol in last year | **80.6** | -1.1 | -0.5 | -0.5 | -3.3 | -1.9 | -1.6 | -2.6 | -2.6 | -2.7 |
| Daily smoker | **13.1** | 8.7 | 8.7 | 8.6 | 4.2 | 2.4 | 2.6 | 4.8 | 4.2 | 4.3 |
| General health status (very good) | **36.2** | -2.9 | -2.0 | -2.1 | -3.6 | -2.9 | -2.7 | -2.6 | -2.2 | -2.2 |
| Life satisfaction (8 out of 10) | **32.6** | -10.8 | -10.2 | -10.1 | -5.1 | -3.0 | -3.8 | -9.0 | -8.6 | -8.8 |
| Has private health insurance | **57.1** | -4.0 | -5.7 | -5.6 | -3.7 | -1.4 | -1.4 | -2.4 | -3.1 | -3.0 |
| Psychological distress, Kessler 6 (low) | **82.2** | -27.7 | -26.9 | -27.0 | -22.9 | -22.7 | -23.2 | -24.5 | -24.4 | -24.5 |
| **AAE (combined)[##]** | | 6.6 | 5.9 | 5.9 | 5.7 | 5.2 | 5.2 | 5.9 | 5.3 | 5.4 |
| **RMSE (combined)[##]** | | 9.2 | 8.6 | 8.6 | 7.7 | 7.4 | 7.4 | 8.1 | 7.6 | 7.7 |
| **AAE (secondary demographics)[##]** | | 4.9 | 3.8 | 3.8 | 4.8 | 4.9 | 4.8 | 4.7 | 3.8 | 3.9 |
| **RMSE (secondary demographics)[##]** | | 5.6 | 4.3 | 4.3 | 5.5 | 5.5 | 5.3 | 5.3 | 4.2 | 4.2 |
| **AAE (substantive items)** | | 9.2 | 9.0 | 9.0 | 7.1 | 5.7 | 5.9 | 7.6 | 7.5 | 7.6 |
| **RMSE (substantive items)** | | 12.8 | 12.5 | 12.5 | 10.0 | 9.5 | 9.8 | 11.0 | 10.9 | 10.9 |

[#] Expanded raking and GREG – calibration by core demographics (gender, age group*education, country of birth, state*capital city in state), and 3 secondary demographics with the most absolute bias in nonprobability samples (i.e., employed, Language other than English, volunteering), [##]calculated for all items but those included in weighting adjustments (colored blue)

Table 9.9: Estimates relative to the benchmarks, unweighted, weighted (raking), and matched (CEM plus raked) estimates for nonprobability-based panels

| | | Least accurate nonprobability panel (1) (n=601) | | | Most accurate nonprobability panel (3) (n=626) | | | 5 nonprobability panels combined (n=3,058) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Survey item** | **Benchmark (%)** | **UW** | **Expanded raking** | **CEM** | **UW** | **Expanded raking** | **CEM** | **UW** | **Expanded raking** | **CEM** |
| Australian citizen | **87.1** | 5.9 | 3.0 | 2.9 | 6.0 | 3.9 | 4.1 | 5.3 | 3.3 | 2.9 |
| Couple with dependent children | **30.3** | -3.5 | 0.0 | 4.9 | -3.3 | 0.0 | 4.4 | -2.5 | 0.0 | 0.1 |
| Currently employed | **61.6** | -10.5 | -10.9 | -8.2 | -7.6 | -5.6 | -4.2 | -8.5 | -7.6 | -6.9 |
| Enrolled to vote | **78.5** | 8.4 | 5.2 | 5.6 | 10.0 | 8.6 | 10.3 | 9.4 | 7.3 | 7.4 |
| Home ownership with a mortgage | **28.8** | 3.0 | 3.8 | 3.9 | 5.1 | 6.5 | 10.2 | 3.5 | 4.0 | 3.8 |
| Not Indigenous | **97.7** | -0.2 | 0.3 | -0.1 | 0.7 | 0.9 | 1.5 | 0.2 | 0.4 | 0.6 |
| Language other than English (speak only English) | **76.5** | 6.2 | 3.8 | 8.2 | 9.1 | 7.0 | 6.4 | 6.8 | 5.1 | 5.3 |
| Living at last address 5 years ago | **56.9** | 4.7 | 4.3 | 5.1 | 7.3 | 10.1 | 9.8 | 6.9 | 7.6 | 5.7 |
| Most disadvantaged quintile for area-based SES | **20.0** | -3.2 | -2.2 | -2.7 | -5.6 | -7.2 | -8.4 | -5.4 | -5.7 | -5.6 |
| Resident of a major city | **66.8** | 9.2 | 6.7 | 5.5 | 1.3 | 2.9 | 2.0 | 4.9 | 3.7 | 2.8 |
| Voluntary work (none) | **79.4** | -6.9 | -6.9 | -7.2 | -8.3 | -8.6 | -10.7 | -7.9 | -8.8 | -9.4 |
| Wage and salary income $1000–1249 per week | **13.8** | -6.4 | -6.4 | -8.1 | -4.0 | -3.2 | -2.0 | -4.0 | -4.2 | -3.1 |
| Consumed alcohol in last year | **80.6** | -1.1 | 0.0 | 8.9 | -3.3 | 0.0 | 8.5 | -2.6 | -0.1 | 4.5 |
| Daily smoker | **13.1** | 8.7 | 0.0 | -1.5 | 4.2 | 0.0 | -4.5 | 4.8 | 0.0 | -1.1 |
| General health status (very good) *(from NHS 2014-2015)* | **36.2** | -2.9 | 1.9 | 9.3 | -3.6 | 1.9 | 7.0 | -2.6 | 1.9 | 7.3 |
| Life satisfaction (8 out of 10) | **32.6** | -10.8 | -7.9 | -5.9 | -5.1 | -1.9 | 0.4 | -9.0 | -6.5 | -6.0 |
| Has private health insurance | **57.1** | -4.0 | -3.0 | -1.1 | -3.7 | -0.6 | 4.8 | -2.4 | -1.2 | -0.4 |
| Psychological distress, Kessler 6 (low) | **82.2** | -27.7 | -20.7 | -18.3 | -22.9 | -15.4 | -12.5 | -24.5 | -19.0 | -17.2 |
| **AAE (combined)*** | | 7.7 | 6.1 | 5.9 | 6.9 | 5.9 | 6.2 | 7.1 | 6.0 | 5.5 |
| **RMSE (combined)*** | | 9.9 | 7.7 | 7.3 | 8.6 | 7.1 | 7.4 | 8.9 | 7.4 | 6.8 |
| **AAE (secondary demographics)*** | | 5.9 | 4.9 | 5.2 | 5.9 | 5.9 | 6.3 | 5.7 | 5.2 | 4.9 |
| **RMSE (secondary demographics)*** | | 6.5 | 5.6 | 5.8 | 6.6 | 6.5 | 7.2 | 6.2 | 5.7 | 5.4 |
| **AAE (substantive items)*** | | 14.2 | 10.5 | 8.4 | 10.6 | 6.0 | 5.9 | 12.0 | 8.9 | 7.9 |
| **RMSE (substantive items)*** | | 17.3 | 12.9 | 11.1 | 13.7 | 9.0 | 7.7 | 15.1 | 11.6 | 10.5 |

Expanded raking - by core demographics (gender, age group*education, country of birth, state*capital city in state), and 4 substantive items from NDSHS 2016 (household composition, alcohol, smoking, general health); CEM – coarsened exact matching with all matching covariates from NDSHS 2016 (gender, age group*education, household composition, alcohol, smoking, general health), later adjusted to population totals with raking (by gender, age group*education, country of birth, state*capital city in state); *calculated for all items but those included in post-survey adjustments (colored blue);

Table 9.10: Estimates relative to the benchmarks, unweighted, weighted (PSW), and matched (MDM, CEM plus raked) estimates for nonprobability-based panels

| | | Least accurate nonprobability panel (1) (n=601) | | | | Most accurate nonprobability panel (3) (n=626) | | | | 5 nonprobability panels combined (n=3,058) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Survey item** | **Benchmark** | **UW** | **CEM** | **MDM** | **PSW** | **UW** | **CEM** | **MDM** | **PSW** | **UW** | **CEM** | **MDM** | **PSW** |
| Australian citizen | **87.1** | 5.9 | 3.6 | 4.2 | 4.6 | 6.0 | 6.4 | 6.4 | 4.9 | 5.3 | 3.9 | 3.6 | 3.5 |
| Couple with dependent children | **30.3** | -3.5 | 3.8 | -4.1 | -4.1 | -3.3 | 2.5 | -1.9 | -2.2 | -2.5 | -0.9 | -1.6 | -2.5 |
| Currently employed | **61.6** | -10.5 | -11.7 | -10.6 | -10.5 | -7.6 | -1.6 | -4.2 | -5.5 | -8.5 | -8.1 | -6.2 | -9.3 |
| Enrolled to vote | **78.5** | 8.4 | -0.2 | 6.5 | 6.2 | 10.0 | 11.7 | 9.9 | 10.0 | 9.4 | 6.0 | 8.6 | 7.1 |
| Home ownership with a mortgage | **28.8** | 3.0 | 6.7 | 1.5 | 1.4 | 5.1 | 14.0 | 6.7 | 7.9 | 3.5 | 5.3 | 3.7 | 3.1 |
| Not Indigenous | **97.7** | -0.2 | -0.3 | 2.1 | 0.3 | 0.7 | 0.4 | 1.5 | 0.7 | 0.2 | 0.1 | 0.9 | 0.4 |
| Language other than English (speak only English) | **76.5** | 6.2 | 3.2 | 5.1 | 6.3 | 9.1 | 0.0 | 8.8 | 6.8 | 6.8 | 2.0 | 7.6 | 4.7 |
| Living at last address 5 years ago | **56.9** | 4.7 | 3.0 | 4.0 | 6.2 | 7.3 | 10.0 | 9.2 | 8.2 | 6.9 | 5.0 | 6.1 | 6.6 |
| Most disadvantaged quintile for area-based SES | **20.0** | -3.2 | -1.7 | -1.7 | -3.2 | -5.6 | -8.2 | -10.3 | -8.8 | -5.4 | -5.3 | -6.6 | -5.5 |
| Resident of a major city | **66.8** | 9.2 | 5.1 | 5.2 | 6.0 | 1.3 | 0.5 | 1.4 | 1.4 | 4.9 | 2.5 | 2.1 | 2.9 |
| Voluntary work (none) | **79.4** | -6.9 | -4.6 | -4.9 | -5.4 | -8.3 | -12.5 | -9.2 | -9.1 | -7.9 | -10.4 | -7.4 | -8.3 |
| Wage and salary income $1000–1249 per week | **13.8** | -6.4 | -5.2 | -8.3 | -6.0 | -4.0 | -4.2 | -3.3 | -4.5 | -4.0 | -5.1 | -3.8 | -4.2 |
| Consumed alcohol in last year | **80.6** | -1.1 | 0.4 | 0.7 | 1.4 | -3.3 | -1.3 | 2.7 | -0.7 | -2.6 | -1.5 | -0.3 | -2.2 |
| Daily smoker | **13.1** | 8.7 | 7.4 | 9.8 | 9.3 | 4.2 | 1.0 | 4.0 | 2.2 | 4.8 | 4.0 | 3.8 | 4.0 |
| General health status (very good) | **36.2** | -2.9 | 4.2 | -4.4 | -1.6 | -3.6 | -0.1 | -3.6 | -3.7 | -2.6 | -1.4 | -1.5 | -2.5 |
| Life satisfaction (8 out of 10) | **32.6** | -10.8 | -17.0 | -6.5 | -13.4 | -5.1 | 1.4 | -2.8 | -2.0 | -9.0 | -9.7 | -7.4 | -8.4 |
| Has private health insurance | **57.1** | -4.0 | -12.1 | -7.7 | -6.3 | -3.7 | -2.2 | 3.9 | -1.4 | -2.4 | -3.9 | -0.1 | -3.6 |
| Psychological distress, Kessler 6 (low) | **82.2** | -27.7 | -28.5 | -25.8 | -26.3 | -22.9 | -32.4 | -25.5 | -25.7 | -24.5 | -31.4 | -22.6 | -25.9 |
| **AAE (combined)** | | 6.9 | 6.6 | 6.3 | 6.6 | 6.2 | 6.1 | 6.4 | 5.9 | 6.2 | 5.9 | 5.2 | 5.8 |
| ***RMSE (combined)*** | | 9.0 | 9.5 | 8.3 | 8.8 | 7.8 | 10.0 | 8.4 | 8.2 | 8.0 | 9.0 | 7.2 | 7.9 |
| **AAE (secondary demographics)** | | 5.7 | 4.1 | 4.8 | 5.0 | 5.7 | 6.0 | 6.1 | 5.8 | 5.4 | 4.6 | 4.8 | 4.8 |
| ***RMSE (secondary demographics)*** | | 6.3 | 5.0 | 5.5 | 5.6 | 6.4 | 7.8 | 6.9 | 6.6 | 6.0 | 5.3 | 5.4 | 5.4 |
| **AAE (substantive items)** | | 9.2 | 11.6 | 9.2 | 9.7 | 7.1 | 6.4 | 7.1 | 6.0 | 7.6 | 8.7 | 6.0 | 7.8 |
| ***RMSE (substantive items)*** | | 12.8 | 14.8 | 12.1 | 12.9 | 10.0 | 13.3 | 10.9 | 10.7 | 11.0 | 13.6 | 9.9 | 11.4 |

UW – unweighted results, PSW – propensity score weighting, CEM – coarsened exact matching, MDM – Mahalanobis distance matching; CEM covariates selected with dominance analysis: number of surveys completed in past 4 weeks (grouped), type of internet connection - a dial-up, frequency of posting to blog/forums/interest groups

# Chapter 10 Survey response in RDD-sampling SMS-invitation web-push study

## 10.1 Introduction

Survey data collection underpins a large proportion of social science research across multiple disciplines, but is increasingly difficult. Surveys are facing unprecedented challenges, response rates have declined steadily over the years, and methods to sample national populations are growing more expensive and complex (Couper 2017). Generally speaking, there are three things more valued than anything else in survey research in practice: low cost, data quality, and time-efficiency, but you can only have two of them at the same time (Keeter 2019). Many surveys that could normally be conducted time-efficiently and for relatively low costs, are generally considered of lesser quality compared to large-scale nationally representative surveys; an example of lower-quality surveys could be convenience samples or volunteer opt-in panel surveys, which are not based on probabilistic principles (Baker et al. 2010). In this study, I will test a survey data collection approach that provides an option for quick data collection at a relatively low cost. In contrast to volunteer panels, it is a probability-based online survey. However, due to its simplistic design, it might be conducted with some loss of quality, similarly to nonprobability surveys. The issue of data quality will be briefly addressed at the end of this article.

Primarily, this article focuses on the response dimension of the approach to a particular kind of data collection. Nonresponse is an issue in both probability and nonprobability surveys, and this project is also meant to provide some evidence on the connection between nonresponse and bias in a survey with low response rates. More specifically, I will study response rates in an RDD-sampling SMS-invitation web-push survey. This study is one of the first to combine these three approaches to sampling (RDD), recruitment (SMS), and data collection methodology (web-push). To the best of my knowledge, it was previously tested only in Germany (Bucher and Sand 2021). Applications of this approach in practice would not be possible without the rise of mobile internet devices, such as smartphones, and/or an increase of internet coverage (Couper 2017). For example, in 2019 in Australia, internet penetration rate was approximately 88% (Datareportal 2020), 91% of adult Australians used mobiles to go online (Australian Communications and Media Authority, n.d.-a), and the smartphone penetration rate was approximately 91% (Deloitte, n.d.). Smartphone surveys might be a promising tool to collect data, but more work should be done to improve response and decrease nonresponse bias in this, to some degree, intrusive task (Elevelt et al. 2019).

Thus, I will investigate how different data collection solutions and maximization approaches affect response, and how researchers can reduce nonresponse to mitigate potential representation bias. We have to have in mind that low response rates might not result in representation bias, as the link between response rates and nonresponse bias has been reported as weak at best (Groves & Peytcheva 2008). However, the decline in response to RDD surveys increased the potential for bias in estimates (Brick 2008). I can argue that there has to be a threshold below which surveys with extremely low response rates fail to sufficiently capture the socio-demographic, attitudinal, behavioral, or factual variability of the population; that is, in combination with coverage bias.

## 10.2 Literature review

### 10.2.1 Probabilistic sampling and undercoverage in web and smartphone surveys

Probability-based sampling requires each unit of the population to have a known non-zero chance of being selected to the sample (Neyman 1938). Generally speaking, there is a missing link between probabilistic sampling and web surveys, as there are no general population sampling frames of email addresses. To address this issue, either offline recruitment (F2F, CATI, or postal) combined with a web-push approach, or non-probabilistic approaches such as river sampling are used in practice (Callegaro et al. 2015).

In CATI surveys, random digit dialing (RDD) is often used as a set of techniques. The advantage of RDD is the probabilistic nature of selection into the sample. It is a method for generating telephone numbers randomly, either landline or mobile numbers. In survey methodology, it has been predominantly used in telephone surveys (Brick 2008). With a decreasing percentage of landlines in developed countries like Australia, RDD methodology has had to adjust to these changes and sample more mobile numbers than landline numbers. This approach to sampling has previously been effectively used in recruitment to online surveys, such as probability-based online panels. In 2016, the Social Research Centre (Melbourne, Australia), managing the only national probability-based online panel, included 30% of landline numbers and 70% of mobile numbers in their sampling design. In 2018, the refreshment sample was recruited exclusively via mobile phones (Kaczmirek et al. 2019), showing a trend towards mobile-only recruitment in the future. However, sampling of only mobile phone numbers for cross-sectional general population surveys has been quite rare and not well documented in the literature.

It is still unclear whether smartphone surveys are a promising alternative to web surveys. They might suffer from coverage and nonresponse, possibly more than web surveys (Antoun et al. 2019) that are generally known for undercoverage of people without access to the internet (Couper 2000). This might

result in undercoverage bias for a number of non-demographic items (Hsia et al. 2020). However, web surveys are not the only survey mode subject to coverage error. For example, in telephone surveys, some people are without landlines and mobile phones. The coverage bias can increase further if only mobile phone numbers are sampled. On the other hand, in countries with high internet penetration rates and high smartphone penetration rates like Australia (Australian Communications and Media Authority, n.d.-a; Deloitte, n.d.), this might be less of a problem than in other countries.

### 10.2.2 Factors affecting response in web surveys

In their systematic review, Fan and Yan (2010) conceptualized factors affecting response in web surveys into those affecting response rates in survey development, survey delivery (such as contact delivery modes, design of invitations, pre-notifications, reminders, and incentives), survey completion (from socio-demographics, psychographics to participation theories), and survey return. This research investigates survey response maximization strategies, and is thus predominantly focused on survey delivery factors.

In practice, more recent evidence shows that survey delivery factors such as survey structure, assurance of privacy and confidentiality, interests of participants, and communication method, highly influence response in web surveys (Saleh & Bista 2017). At the questionnaire design level, factors such as survey length, question difficulty, the content of the first question, and usage of a progress bar are related to completion rates (Liu & Wronsky 2018). Moreover, Van Mol (2017) reported the effectiveness of extra reminders in an online survey among over-surveyed populations (regardless of the reminder content). On the other hand, Saleh and Bista (2017) reported that email reminders and incentives are effective in only particular socio-demographic groups.

Texting mobile numbers has previously been used as a survey invitation mode, a response maximization technique, and a pre-recruitment method to other survey modes. For example, American Trends Panel panellists are sent either email or SMS invitations if they have previously consented; all initially offline respondents who later received tablets, receive only text message invitations to their device (Keeter 2019). Under certain conditions, texting is more effective than sending emails. De Bruijne and Wijnant (2014) confirmed that text messaging is more efficient for invitations when considering a response via a smartphone, and equally efficient as email invitations when considering total response in an online panel (but this can lead to a faster response). Moreover, Phillips and Compton (2019) reported that SMS reminders were associated with an increase in response rate in an online survey; in comparison to telephone reminders, they were less efficient, but more cost-effective. On the other hand, SMS reminders can have a positive impact on response in comparison to e-mail reminders (Sala et al. 2018), and Bosnjak et al. (2008) reported that sending an

advance SMS was more effective when compared to email pre-notifications in an opt-in online panel survey. These results were interpreted as SMS being both attention-grabbing and effective for establishing legitimacy.

Offering rewards is one of the most common response maximization strategies in survey delivery. There has been extensive research on the effectiveness of conditional and unconditional monetary/coupon incentives in web surveys, as well as prize draws, but the evidence has been mixed. In the last decade, some authors report that offering unconditional incentives (e.g., Parsons & Manierre 2014), conditional incentives (e.g., Dykema et al. 2011), and a chance to enter a prize draw/lottery (e.g., Laguilles et al. 2011; Morgan et al. 2017) increase response rates in web surveys. Other research did not find an effect of unconditional (e.g., Dykema et al. 2011) or conditional (e.g., Knowles & Stahlmann-Brown 2021) incentives, or an increase in the conditional incentive amount (Neal et al. 2020; Spreen et al. 2020). However, Mavletova and Couper (2016) reported that offering conditional differential incentives increased response rates of mobile-completers more than of PC web-completers.

### 10.2.3 SMS and text-to-web surveys

Dillman (2018) argues that the 2020s will be the age of smartphones, with most telephone communication being no longer voice conversation but rather texts and emails. This creates problems for RDD telephone surveys, and opportunities for text message surveys and text-to-web surveys. SMS surveys can be considered as a form of mass messaging and are known to have a low response, selection bias, and low data quality (Kongsgard et al. 2014). However, we have to acknowledge regulatory environments that could make SMS and text-to-web surveys, as well as texting to increase response, quite limited in particular contexts. In the US (Fordyce et al. 2020) and some European countries (Kongsgard et al. 2014), prior consent to text messages is required, even for research purposes. On the other hand, it is not required in countries like Germany (Bucher & Sand 2021) and Australia.

Texting as an interview mode can be defined as pre-sending survey questions via SMS and also receiving answers from a respondent via SMS. Some of the advantages of this approach to data collection are a quick turnaround, collecting responses close in time to behaviors as the subject of survey research, and the ability for behavioral intervention, i.e., sending both information and reminders (Conrad et al. 2017). As well as SMS surveys, data can be collected with text-to-web surveys. Fordyce et al. (2020) compared synchronous text message surveys, i.e., questions and answers are exchanged in text messages, and asynchronous text message surveys, i.e., a text-to-web survey with a URL in the invitation; they reported no significant relationship between completion rates and the

type of text message survey. Lastly, Balabanis et al. (2007) who examined the use of SMS to recruit respondents to web and telephone surveys, concluded that SMS can be used effectively with mixed-mode methods or as a pre-recruitment method to panels of respondents.

### 10.2.4 Aims of this research

The literature explains that telephone surveys based on RDD sampling are generally known for their relatively low response rates (Keeter et al. 2017), and the same conclusion can be made for text message surveys (Conrad et al. 2017; Kongsgard et al. 2014). In terms of errors of representation (see the Total Survey Error framework in Groves et al. 2009), nonresponse errors are not the only errors prevalent in internet and smartphone surveys – there is also the issue of undercoverage of people with no internet access in their household or on their mobile devices (Antoun et al. 2019; Couper 2000). This could lead to a notable representation bias as a result of combining both types of error of representation. It is often challenging to distinguish between nonresponse and coverage errors in mobile surveys, and discussing selection bias is more appropriate (Couper et al. 2017). Thus, in this article I will test this fairly new approach with a focus on nonresponse and, as a result of both nonresponse and undercoverage, socio-demographic representation bias. With an empirical analysis, I will answer the following research questions.

RQ1) What response rates can be expected in an RDD-sampling SMS-invitation web-push survey?

RQ2) What data collection characteristics in survey delivery, such as incentives, text message content, or time of sending SMS invitations, affect response rates in a survey of this type?

RQ3) What level of socio-demographic representation bias is present in a survey of this type?

The evidence presented in this study can be extended to other survey research approaches in similar contexts (e.g., in Australia, with a similar topic) using text messaging. This includes but is not limited to: SMS surveys, SMS recruitment, and SMS pre-notifications and reminders.

## 10.3 Methods

### 10.3.1 Data

The data from this methodological project were collected and compiled by the Centre for Social Research and Methods at the Australian National University, with the main aim to explore ways to replace existing, expensive survey methods with cheaper, more flexible ones. The RDD-sampling SMS-invitation web-push survey was purposely designed to enable the study of not only response, but also nonresponse/representation bias, and accuracy relative to a number of demographic and non-demographic benchmarks from large-scale government-funded surveys in Australia.

To collect survey data in this project, the online 'Survey on Wellbeing, Health and Life in general 2020' questionnaire was programmed (see Appendix 10). Besides items measuring health and wellbeing, the questionnaire included items on the use of internet and technology, satisfaction with different dimensions of life, personality traits, primary demographic items like gender, age, and education, and secondary demographics such as country of birth, citizenship, employment, and income. The questionnaire consisted of 35 questions, out of which there was one multiple answer question and no grid or open-ended questions. The median response time was about 8 minutes.

To study factors affecting survey response, the analyzed data file consisted of all randomly generated mobile numbers receiving an SMS invitation (n=38,512) as cases. For each telephone number, the following data collection characteristics were coded prior to data collection: *time of day* and *day of week the SMS was sent, text message content, reminder, incentives offered, type of invitation, appended geo-demographics,* and *stratification information*. After the data collection was completed, the survey data file (n=631) was used to identify all mobile numbers belonging to survey participants and derive the response variable *survey response* (1=unit response, 0=unit nonresponse). Section 'Data analysis' describes the dependent variables, independent variables, and statistical modeling in more detail.

### 10.3.2 Sampling and sampling frame

Since there is no real connection between probabilistic sampling and web survey collection data from the general population (Callegaro et al. 2015), I combined a sampling approach generally used in telephone surveys and telephone recruitment with online data collection. Thus, I carried out random digit dialing (RDD) generation of Australian mobile numbers. Each Australian mobile number consists of the leading numbers 04 and 8 more digits that can be randomly generated – format 04XX XXX XXX. Hence, there are 100 million possible combinations, and not all of them are used as of 2020. There are spare numbers, e.g., 0440 000 000 - 0444 300 000, numbers allocated to satellite phones, e.g., 0420 100 000 – 0420 109 999 (Pivotel), and rail corporations, e.g., 0420 000 000 – 0420 019 999 (Rail Corporation New South Wales) (Australian Communications and Media Authority, n.d.-b). The remaining numbers, 62 million or 62% of all possible combinations, are allocated to Vodafone, Optus, and Telstra mobile service providers and can be used in general population surveys. However, knowing that there are about 24 million people living in Australia (Australian Bureau of Statistics 2016) and

about 20 million mobile numbers[70], only about one out of three combinations represent a valid/live mobile number.

To remove invalid mobile numbers and to decrease the cost of texting invitations, i.e., not sending SMS messages to mobile numbers that are not live, I used the service provided by SamplePages (n.d.). They matched my randomly generated mobile numbers[71] to the mobile numbers from the general population in their database for validation and to append geo-demographics (with approximately 7% matching rate, 34,734 numbers[72]). They estimated that 90% of all those numbers were live at the time of matching and appending. For the top-up sample (Stage 2, please see 'Survey experiment' subsection), they also validated 3,778 additional numbers[73] with no matches in their database. 100% of those numbers were active on the day of validation, i.e., 1-3 days before the text messages were sent to the validated numbers. In the end, the total sample consisted of:

- 12,302 numbers with full geo-demographics available (gender, age group, Statistical Area Level 4 (SA4));
- 22,432 numbers with partial geo-demographics available (about 96% of those without age [SA4 only, gender only, or SA4 and gender] and about 4% with age information [age only, age and gender, or age and SA4]);
- 3,778 numbers with no geo-demographics available (added in Stage 2).

For Stage 1, the sample was stratified[74], and for Stage 2, there was just one sample with no stratification carried out. Stratification was primarily used to study its effect on data accuracy, i.e., for the purpose of a separate study.

[70] Estimation based on: 18.5 million Australians aged 18+ (Australian Bureau of Statistics 2016), 6% of them with two mobile phones, 32% of Australian aged 6-13 and 91% of Australian aged 14-17 own a mobile phone (Roy Morgan 2015; Roy Morgan 2016; Roy Morgan 2018).

[71] 500,000 mobile numbers were generated using RDD; the number was determined based on the estimated match rate and the required sample size of mobile numbers with a positive match.

[72] While generation of mobile phone numbers was carried out by the researcher and was random (following RDD principles), validation of mobile numbers provided by SamplePages in Stage 1 excluded Australian mobile numbers not in their database. As such, the selection closely resembled drawing random numbers from the SamplePages database, but with more control over sampling on the researcher's end (and for lower cost). In practice, there is a trade-off between (1) coverage of mobile numbers not in SamplePages database, (2) the ability to carry out stratified sampling of mobile numbers in the Australian context. This potential undercoverage bias is, combined with nonresponse bias, addressed in the Representation bias subsection of the Results.

[73] Mobile number validation also known as 'pinging'; 12,000 RDD mobile numbers (with no matches in SamplePages database in Stage 1 selection), 3,778 active, 31% live number validation rate, fairly consistent with '20 million mobile owners / 62 million possible combinations' ratio. This type of selection using random generation of mobile numbers with subsequent 'pinging' can be considered as a true RDD in the Australian context. The 'pinged' sample of mobile numbers was added in Stage 2 to study response rate conditional on the type of validation of mobile numbers (see Table 10.4 for results).

[74] Based on availability of stratification information: 8,000 numbers by gender*age group*SA4 and 19,000 by state only.

### 10.3.3 Survey experiment

In this study, I collected data with a sophisticated survey experiment, including two associated stages in a responsive survey design.

- Stage 1: initial stage to study various data collection characteristics

To study factors affecting and improving response in a survey applying an RDD-sampling SMS-invitation web-push approach to data collection, I divided the sample of 27,000 numbers into 48 experimental groups. There were a number of experimental variables I intended to test the survey (non)response against: (1) survey reminder; (2) day of the week initial SMS was sent; (3) time of day initial SMS was sent; (4) incentives; and (5) SMS invitation text (information on the topic of the survey, information on benefits of participation). Please see Table 10.1 for more information.

Table 10.1: Experimental design (Stage 1, n=27,000)

| Reminder | Day of the week 1st SMS sent | Time of the day | Type of incentives | SMS invitation text | % of the combined sample in Stage 1 |
|---|---|---|---|---|---|
| Survey reminder *(5 days after first SMS)* | Weekday (Mon-Fri) | Afternoon | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |
| | | Evening | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |
| | Weekend (Sat-Sun) | Afternoon | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |
| | | Evening | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |
| No SMS reminder | Weekday (Mon-Fri) | Afternoon | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |
| | | Evening | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |
| | Weekend (Sat-Sun) | Afternoon | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |
| | | Evening | No incentives | Benefits | 3.750% |
| | | | | Survey topic | 3.750% |
| | | | $5 conditional | Benefits | 0.625% |
| | | | | Survey topic | 0.625% |
| | | | Prize draw | Benefits | 1.875% |
| | | | | Survey topic | 1.875% |

While all other conditions were split 50:50, I had to adjust the sizes of the experimental groups based on the types of incentives, which depended on the available budget. With a total budget of $3000, I intended to spend $200 on incentives for the prize draw (target sample size n=400), $1000 on incentives for the $5 conditional incentives group (target sample size n=200), approximately $700 on

mobile number validation and appending geo-demographics, and approximately $1100 on text messaging. Assuming a lower response in groups not being offered a (monetary) reward, the no-incentives group of sampled potential respondents was larger (60%) than the lottery (30%), and the $5 incentives (10%) groups. I purposely collected data on all days of the week in an attempt to minimize the effect on the response of particular events on certain days of the week. However, this was done with an intention to aggregate the days into 'weekdays' and 'weekends', to keep sufficient statistical power for all experimental groups.

- Stage 2: a sample top-up stage, also testing for two other response maximization approaches

The response rate in Stage 1 was about 50% lower than initially expected, and to increase the sample size for a separate benchmarking component of the project, the decision was made to use a top-up sample. For this reason, I decided to use all remaining numbers with appended geo-demographics (n=7,734, with predominantly partial information for stratification), and to validate new numbers with no geo-demographic information (n=3,778, pinged but not matched).

In Stage 2, I did not replicate the experimental design from Stage 1. Instead, I analyzed the existing preliminary data from Stage 1 and selected the optimal combination of approaches in terms of response and costs (a responsive approach). If there was no notable difference in rates between experimental groups, I decided to go with the cheapest data collection approach. In the end, I standardized the following conditions: weekday evening invitations (Tuesday-Thursday), text message communicating benefits, no incentives offered, and no survey reminders.

I introduced two alternative approaches to SMS recruitment as follows.

(1) Advance SMS/pre-notification. I attempted to increase response by pre-notifying respondents about the upcoming SMS invitation; the literature explains that advance SMS can help establish trust and legitimacy (Bosnjak et al. 2008). At the same time, I only texted a link to the survey to respondents who did not opt-out by responding 'STOP'.

(2) 'Responsive' text message survey invitation. I attempted to increase response by not sending the link to the survey in the first text message, but only if the respondents agreed to receive an invitation by responding 'YES'. This can be considered as a less intrusive approach to survey invitations, similarly to an advance SMS.

The third type of invitation was a 'standard' single-invitation SMS including a URL to an online questionnaire. This was the only type of invitation from Stage 1. Please see Table 10.2 for more information.

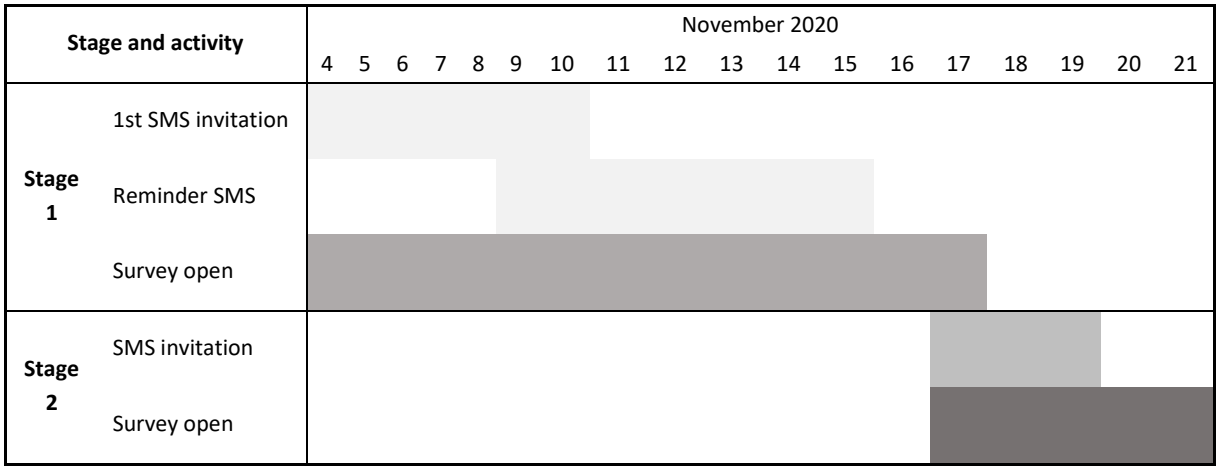
Table 10.2: Experimental design (Stage 2, n=11,512)

| Reminder | Day of the week 1st SMS sent | Time of the day | Type of incentives | SMS invitation text | Type of invitation | % of the full sample |
|---|---|---|---|---|---|---|
| No reminder | Weekday (Tue-Thu) | Evening | No incentives | Benefits | Advance SMS invitation | 33.33% |
| | | | | | 'Responsive' SMS invitation | 33.33% |
| | | | | | 'Standard' SMS invitation | 33.33% |

### 10.3.4 Data collection

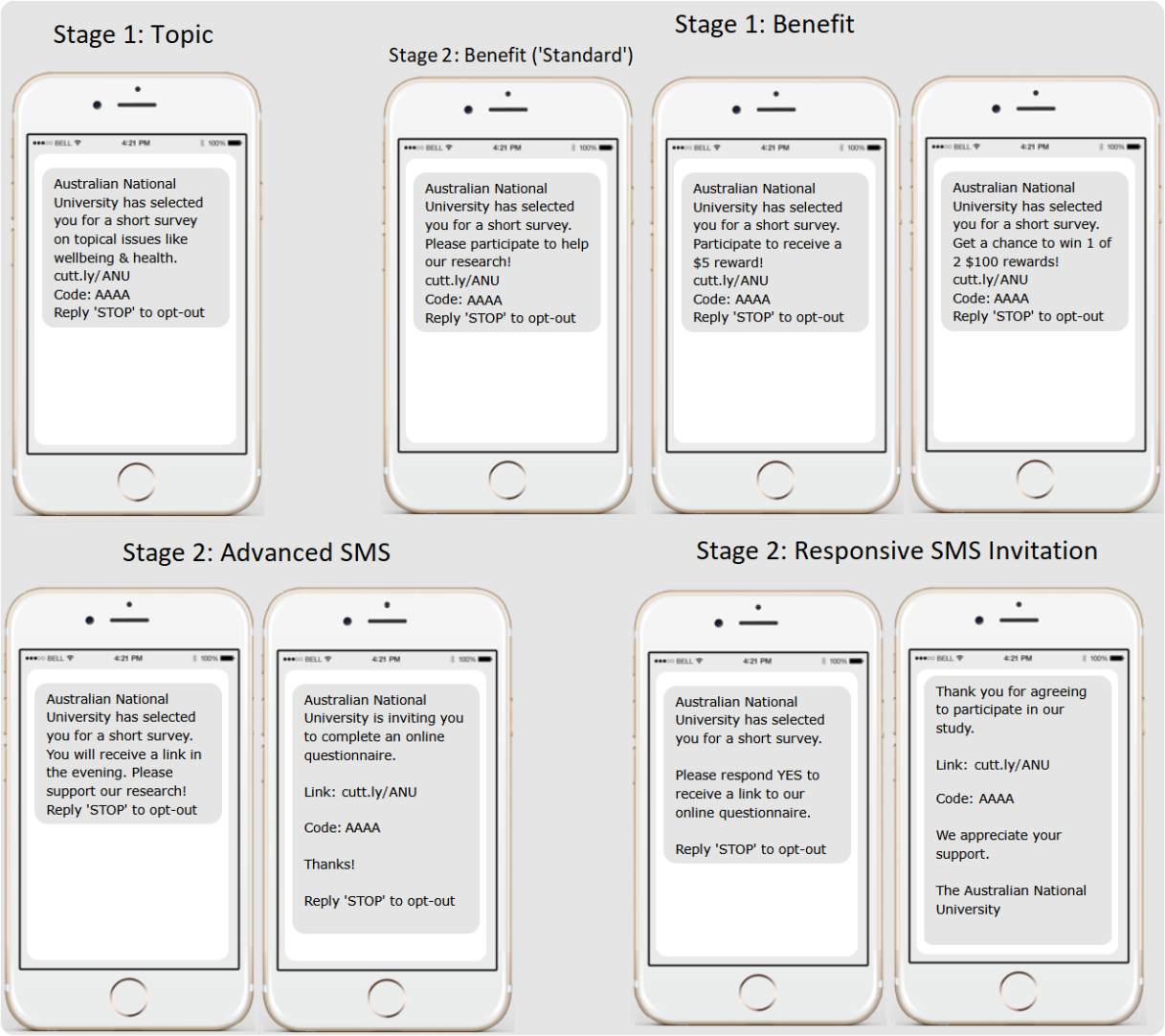
Data collection took place between Wednesday November 4th and Saturday November 21st 2020. The initial stage, Stage 1, took place between November 4th and November 17th, and Stage 2 took place between November 17th and November 21st 2020. Due to the experimental design, invitations were sent during most of these periods; the online survey/questionnaires were deactivated two days after the last invitation in a stage was sent. Reminders sent to one half of the Stage 1 sample, excluding those who opted-out of receiving future SMS after the first SMS invitation, were sent between November 9th and November 15th using a unified approach that avoids adding extra variability: SMS was sent 5 days after the first one, in the evening, offering the same type of incentives as in the first SMS, and communicating benefits. For more information about the timeline, please see Figure 10.1.

Figure 10.1: Data collection timeline




Text messages were designed based on the results of a qualitative study by the Social Research Centre, who ran focus groups to test the design of their advance SMS that was later used in recruitment to their online panel (Kellard 2017). Moreover, the communicated benefits of participation without receiving a reward (i.e., asking to help our research) was based on survey participation theories, such as social exchange or cognitive dissonance theories (for more information, see Keusch 2015). The text message content is shown in Figure 10.2.

Figure 10.2: SMS content




Stage 2 was conducted in much less time and was built on an optimized survey design heavily based on the results of Stage 1. It was finished in 5 days and included elements of rapid data collection, which could be considered as a key advantage of the proposed RDD-sampling SMS-invitation web-push approach.

### 10.3.5 Data analysis and statistical modeling

With the division into experimental survey groups (see Tables 10.1 and 10.2), I aimed to create a binary logistic regression model with *survey response* (0=unit nonresponse, 1=unit response) as the dependent variable and characteristics of data collection as the predictor variables. After combining and coding data from Stages 1 and 2, those predictors were:

- time of day (afternoon, evening);
- day of the week (weekday, weekend);

- SMS invitation text (topic, benefit);
- reminder (reminder sent, reminder not sent);
- type of invitation for maximizing response ('standard' single-SMS invitation with no incentives offered (reference group), 'standard' single-SMS invitation with $5 incentives offered, 'standard' single-SMS invitation with an offer to enter a $100 prize draw, advance SMS with no incentives offered, responsive SMS invitation with no incentives offered).

After reviewing the results of the preliminary analysis and receiving feedback from SamplePages, I decided to use *different information of appended geo-demographics to a mobile number* as a predictor as well. The following groups of mobile numbers were coded: (1) geo-demographics including age; (2) partial geo-demographics not including age; and (3) no geo-demographics appended/available. The reason for this was the different probabilities of a mobile number being live, and due to eligibility of respondents dependent on their age information (people aged younger than 18 were ineligible).

Taking these probabilities into account, response rate calculation adjusting for unknown eligibility (*e* value) would be considerably affected. To calculate the rates, I used AAPOR Response Rate Standard definitions, and their proposed calculations of response rates RR2 and RR4 (The American Association for Public Opinion Research 2016, pp. 61-62). In both cases, the RR were calculated by counting partial interviews[75] as survey respondents. To calculate the *e* value for RR4, I used estimates from SamplePages in combination with the estimates from Roy Morgan Young Australians Survey 2018 (Roy Morgan 2018) and Roy Morgan Single Source Australia 2016 on phone ownership of minors (Roy Morgan 2016). More details about the estimates and calculation of response rates are available in the Results section.

Data analysis was carried out in the statistical software Stata 13 (StataCorp 2013). As well as descriptive analysis, as previously discussed, I conducted binary logistic regression modeling.

## 10.4 Results

In this section, I will present the results to answer research questions RQ1-RQ3. Firstly, I will discuss response rates in a survey combining online data collection with RDD sampling and SMS invitation, comparing responses between different fundamental approaches, such as offering incentives and an advance SMS. Secondly, I will dig deeper into what affects response and analyze the data for all experimental groups using binary logistic regression modeling. Finally, I will present nonresponse bias by comparing the distribution of socio-demographic variables between this study, the Australian

[75] Partial interviews were those respondents who provided enough information for post-stratification weighting, i.e., reached at least question 28 out of 35 (see the questionnaire in the Appendix 10), but did not complete the questionnaire.

Census 2016 benchmarks (Australian Bureau of Statistics 2016), and three probability-based samples from the Online Panels Benchmarking Study (Pennay et al. 2018).

### 10.4.1 Response rates

To answer the first research question RQ1, I will present the results on response rates for a few different response maximization approaches from data collection Stages 1 and 2. In this analysis, I will not control for other characteristics of data collection (e.g., with logit regression). For this reason, I calculated AAPOR RR2 and RR4. The difference between these two calculations is the estimate of a portion of the sample with unknown eligibility that are ineligible (represented by the *e* value). In an RDD-sampling SMS-invitation web-push survey, respondents with unknown eligibility are those who did not respond and who did not break off. Respondents who confirmed that they were at least 18 years of age by starting the survey were considered eligible. Unfortunately, those who clicked on the link but did not start the survey, did not provide enough information to assume their eligibility status.

With three 'types' of mobile numbers based on the availability of appended information, three different *e* values had to be estimated to calculate AAPOR RR4. Based on the information I received from SamplePages, about 10% of their database contains mobile numbers that are not live, hence the coefficient of 0.9 for the 'Complete geo-dem' group with only respondents aged 18+. Without having appended information about mobile owners' age, I estimated the portion of Australians with mobile phones who were not yet 18. I consulted two reports from Roy Morgan (2016: 91% of teenagers aged 14-17 have a mobile phone, 2018: 32% of children aged 6-13 have a mobile phone) and the Australian Census 2016 (Australian Bureau of Statistics 2016) distribution by age, and estimated that about 9% of all Australian mobile owners were not yet 18. Combined with the portion of inactive numbers, the final *e* value for 'Partial geo-dem group' was 0.82. With pinged (all live) numbers, I only had to adjust the *e* value for ineligible respondents not yet 18 years of age.

Table 10.3: Response rates by different response maximization approaches

| AAPOR Survey Outcomes | | Stage 1[a] | | | Stage 2[b] | | |
|---|---|---|---|---|---|---|---|
| | | **No incentives** | **$5 incentives** | **Lottery** | **Standard invitation[c]** | **Advance SMS** | **Responsive SMS** |
| All invited | | 16,184 | 2,724 | 8,092 | 3,838 | 3,837 | 3,837 |
| Participated | Complete | 267 | 44 | 126 | 36 | 79 | 43 |
| | Partial | 14 | 2 | 14 | 3 | 3 | 0 |
| Known eligibility | Breakoff | 33 | 1 | 10 | 10 | 12 | 4 |
| Unknown eligibility (for mobile numbers with different appended information) | Geo-dem including age | 6,894 | 1,117 | 3,417 | 542 | 510 | 534 |
| | Partial geo-dem not including age | 8,976 | 1,560 | 4,525 | 1,997 | 1,997 | 2,014 |
| | No geo-dem (pinged) | 0 | 0 | 0 | 1,250 | 1,236 | 1,242 |
| | Total | 15,870 | 2,677 | 7,942 | 3,789 | 3,743 | 3,790 |
| e value | Geo-dem including age | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 |
| | Partial geo-dem not including age | 0.82 | 0.82 | 0.82 | 0.82 | 0.82 | 0.82 |
| | No geo-dem (pinged) | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 |
| | Combined | 0.855 | 0.853 | 0.854 | 0.861 | 0.861 | 0.861 |
| **Response rate** | **RR2** | **1.74%** | **1.69%** | **1.73%** | **1.02%** | **2.14%** | **1.12%** |
| | **RR4** | **2.02%** | **1.97%** | **2.02%** | **1.18%** | **2.47%** | **1.30%** |

[a] about 45% of the three samples received a reminder SMS (50% randomly assigned minus those who opted-out after receiving first SMS invitation [about 1 of 10])

[b] no incentives or reminders

[c] 'standard' invitation was a single SMS invitation including a URL to the questionnaire; in case of 'standard' invitations, no advance SMS or SMS asking potential respondent to text back 'YES' to receive a questionnaire URL ('responsive SMS invitation'), were sent; all Stage 1 invitations were 'standard'

The results from Table 10.3 show how different actions to increase response in a mobile survey are more or less efficient. Interestingly, there was a very little and statistically not significant difference between respondents who were not offered incentives (1.74% RR2, 2.02% RR4), those who were offered $5 incentives (1.69% RR2, 1.97% RR4), and those who were offered to enter a lottery for a $100 eGift card (1.73% RR2, 2.02% RR4) in Stage 1.

It seems that it is not worth investing money into offering potential respondents a compensation for their participation in an SMS invitation survey, but rather into other approaches like sending an advance SMS (2.14% RR2, 2.47% RR4). When using the approach with an introductory SMS but no reminders, the response rate was more than twice as high than without an introductory SMS (1.02%

RR2, 1.18% RR4 in Stage 2), and still notably higher than for the Stage 1 groups with about 45% of respondents receiving a reminder SMS.

Standard invitation group and responsive SMS group (RR2 1.12%, RR4 1.30%) from Stage 2 had a lower response but they also did not receive a reminder SMS in contrast to the groups from Stage 1 and, therefore, response rates cannot be compared directly. It is possible that response rate in a responsive SMS survey with a reminder would be comparable to response rates from the first three approaches. This will be estimated with further analysis.

### 10.4.2 Factors affecting response

In the first subsection, I showed how response rates differ between different SMS texting approaches and how unknown eligibility, which I was able to estimate using external data, affects the final response rates as a conditional indicator of data quality. In this subsection, I will show the results of binary logistic regression to provide answers to RQ2. Regression modeling was conducted to showcase how different approaches I tested with the experimental randomized design affect or do not affect (non)response. In this section, only RR2 numbers are presented and discussed. In the unit record file, I cannot assume which numbers were eligible and ineligible, hence no RR4 can be calculated.

The results from Table 10.4 show that there are a number of ways to maximize response, but there are also unnecessary (and quite expensive) measures that do not successfully convince potential survey respondents to participate. Thus, many of my findings on response maximization methods and techniques were not in line with theoretical expectations. First of all, offering incentives in a single-invitation SMS had no positive or negative effect on response rates in comparison to a single-invitation SMS with no incentives offered. This is in line with findings on response rates (RQ1). Inefficiency of incentives comes as a surprise, as it was anticipated in the survey design phase to give double the response rate for the lottery experimental group, and three times the response rate in the $5 conditional incentives experimental group. This inefficiency resulted in a much lower total response rate in Stage 1, encouraging me to make a decision to use a top-up sample and Stage 2 of data collection. In this stage, I used an opportunity to test two different approaches as an alternative to offering incentives: an advance SMS invitation; and a responsive SMS invitation. While responsive SMS invitations did not improve response in comparison to a standard single-SMS invitation, an advance SMS invitation statistically significantly boosted response. The improvement in response was quite similar to an improvement if potential respondents were sent a reminder 5 days after receiving the first SMS invitation.

Table 10.4: Binary logistic regression with survey response (RR2) as dependent variable

| **Predictors** | **Coef** | **Std error** |
|---|---|---|
| **Type of invitation** | | |
| No incentives offered, standard invitation | 0 | |
| $5 conditional incentives, standard invitation | -0.01 | 0.16 |
| Prize draw for $100 coupons, standard invitation | 0.01 | 0.10 |
| No incentives, advance SMS invitation | 0.67** | 0.15 |
| No incentives, responsive SMS invitation | 0.02 | 0.18 |
| **Stratification information** | | |
| Pinged numbers, no info | 0 | |
| Complete stratification info or incomplete info with age | 0.69** | 0.18 |
| Incomplete stratification info (with no age info) | 0.38* | 0.18 |
| **Survey SMS reminder** | | |
| No SMS reminder sent | 0 | |
| Yes, 5 days after first SMS | 0.59** | 0.09 |
| **Day of the week** | | |
| Weekday | 0 | |
| Weekend | -0.05 | 0.09 |
| **Time of the day** | | |
| Evening | 0 | |
| Early afternoon | -0.13 | 0.09 |
| **SMS invitation text** | | |
| Benefit | 0 | |
| Topic | -0.01 | 0.09 |
| *Constant* | -4.82** | 0.19 |
| *Pseudo R Squared* | 0.013 | |

*** p<0.01, * p<0.05, Coef = binary logistic regression coefficient*

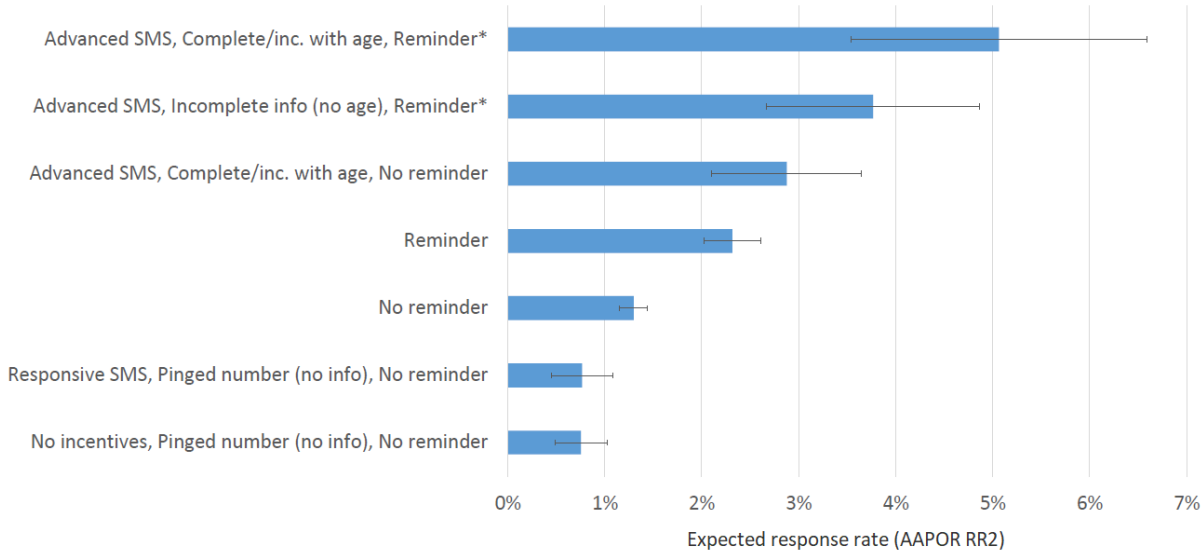
Moreover, I noticed statistically significant differences in response rates between mobile numbers with different levels of appended demographics. The findings are partially in line with findings in the first subsection, where I calculated e values (estimates of unknown eligibility). The numbers with the highest response rate were those with *complete stratification info or incomplete info with age*, which comes as no surprise because they do not include mobile owners that are younger than 18. They are followed by the numbers with *partial stratification information with no age information* and 'pinged' numbers with no stratification information appended, although the difference between these two groups is statistically significant at the $p<0.05$ level but not at the $p<0.01$ level. One possible explanation could be that people with mobile numbers with some stratification information available are less concerned with privacy, since they agreed to have their mobile number listed.

On the other hand, I did not notice any statistically significant differences between the time of day I sent SMS invitations, the days I sent SMS invitations, and the content of the text message (communicating topic or benefits, such as offering incentives). In Stage 1 I noticed a slightly higher

response rate if texting invitations in the evening, but the difference after combining data is not statistically significant. With days of the week, I noticed that there might be differences between certain days (e.g., Saturday at first seemed to be a bad day for sending the first SMS, but a suitable day for sending a reminder), but the aggregation into weekdays and weekends eliminated these differences. Some of these data collection characteristics should be explored further in future research using larger samples.

In Figure 10.3, I am extending the analysis, and presenting the predictive margin results for the best and worst combinations of response maximization approaches for SMS invitation web-push data collection. The results show it is as important to have a high-quality list of validated mobile numbers, as it is to select the right approach to SMS invitation. While we did not observe large differences between nonresponse in the results from Table 10.3, the presented predictive margins showcase how different (hypothetical) strategies can result in quite different response outcomes.

Figure 10.3: Predictive margins for the best and worst combinations of approaches based on RR2 response rates (binary logistic regression model, see Table 10.4 for coefficients)




*those were not experimental groups in my study; the predictive margins for these combinations were calculated with Stata 13 based on response rates in other similarly structured experimental groups from my study

The maximization approach with the highest predicted RR2 is one that I did not test in our study, as I did not send reminders in Stage 2. This combination is an advance SMS using mobile numbers with appended age, and a reminder (RR2 95% confidence interval (CI) [3.5%,6.6%]). It is followed by the same approach except using mobile numbers with no age information (RR2 95% CI [2.7%,4.9%]), but the difference is not statistically significant. Of all approaches I combined and tested in practice, the best seemed to be 'advance SMS using mobile numbers with appended age and no reminder' (RR2 95% CI [2.1%,3.6%]).

In terms of using reminders – the difference between sending a reminder SMS five days after the initial text message invitation, and not sending a reminder, is statistically significant. In practice, the difference is about 1%-point, *ceteris paribus*. This is consistent with the results presented in Table 10.3.

Out of all the different maximization approaches I tested, the least effective were shown to be using pinged numbers with no reminders and either a responsive or a single-SMS invitation with no incentives offered. In these cases, we could not realistically expect RR2 of more than 1.1% (95% CI upper interval).

### 10.4.3 Representation bias

In the following paragraphs, I will address the issue of representation bias, which may or may not be related to low response rates; the issue identified in the existing literature on the topic (e.g., Conrad et al. 2017). As I discussed previously, extremely low response rates can represent a bigger problem than just low response rates, which are an issue for most survey research nowadays. Also, 90% of the sampled mobile numbers in this study were randomly selected from SamplePages database containing estimated 22% of all mobile numbers from the general population of Australians. Thus, there was a potential for undercoverage bias.

To estimate representation bias, I will compare the distribution of socio-demographics variables, most of which are commonly used in post-stratification weighting like raking, between different probability-based samples. Estimates from these samples will be compared to the benchmarks from the Australian Census 2016 (for the general population, 18+ years of age [Australian Bureau of Statistics 2016]) as the highest quality data source for Australia. Representation bias from the RDD-sampling SMS-invitation web-push survey will then be compared to representation bias from the following: a standalone RDD telephone sample; an RDD end of a telephone survey ("piggybacking") sample; and an address-based sample[76]. In the end, I will answer the research question RQ3.

[76] The samples are from the Online Panels Benchmarking Study (OPBS) 2015; estimates are taken from Pennay et al. (2018, pp. 39-40).

Table 10.5: Differences in distributions of key primary socio-demographic variables (unweighted)

| | Benchmark AU Census 2016 | RDD SMS Web-push | Standalone RDD | A-BS sampling recruitment | RDD End of survey recruitment |
|---|---|---|---|---|---|
| **Sex** | | | | | |
| Male | 48.8 | 44 | 46 | 39 | 42 |
| Female | 51.2 | 56 | 54 | 61 | 58 |
| **Age (years)** | | | | | |
| 18–24 | 11.8 | 6 | 7 | 4 | 6 |
| 25–34 | 18.5 | 11 | 9 | 10 | 9 |
| 35–44 | 17.3 | 13 | 15 | 13 | 15 |
| 45–54 | 17.1 | 18 | 15 | 16 | 19 |
| 55–64 | 15.1 | 24 | 20 | 22 | 21 |
| 65-74 | 11.4 | 23 | 18 | 22 | 21 |
| 75+ | 8.8 | 6 | 14 | 13 | 9 |
| **Education** | | | | | |
| Secondary Education or Cert I/II | 43.4 | 25 | 38 | 40 | 35 |
| Cert III/IV, (Advanced) Diploma | 29.8 | 34 | 29 | 24 | 29 |
| Bachelor's degree or higher | 26.7 | 41 | 33 | 37 | 35 |
| **Birthplace** | | | | | |
| Australia | 66.3 | 76 | 75 | 73 | 75 |
| Other | 33.7 | 24 | 25 | 27 | 25 |
| **Region** | | | | | |
| New South Wales | 32.0 | 32 | 29 | 34 | 32 |
| Victoria | 25.5 | 25 | 24 | 26 | 25 |
| Queensland | 19.9 | 18 | 22 | 17 | 20 |
| South Australia | 7.3 | 8 | 9 | 8 | 8 |
| Western Australia | 10.5 | 9 | 10 | 9 | 9 |
| Tasmania | 2.2 | 3 | 3 | 2 | 3 |
| Northern Territory | 0.9 | <1 | 1 | 1 | 1 |
| Australian Capital Territory | 1.7 | 4 | 4 | 3 | 1 |
| **Response rate (RR2)** | | **1.6%** | **14.7%** | **26.2%** | **9.8%** |

The results in Table 10.5 show socio-demographic differences between the general adult population (from Australian Population Census 2016) and different sample surveys. These can be interpreted as a combination of nonresponse and coverage bias. In practice, the differences are corrected with post-stratification weighting, which often does but sometimes does not improve the accuracy of non-demographic estimates (Groves et al. 2009).

We can see that my sample is closer to the benchmarks than other samples for some variables and their categories, and more biased for some other items. Generally speaking, all probability-based samples, i.e., my RDD-sampling SMS-invitation web-push sample and OPBS 2015 samples, are different to the population distributions of the target primary demographics. Firstly, surveys tend to attract more females than males, which was confirmed by my data. Secondly, surveys attract older respondents; in my survey, there was a larger than usual portion of those aged 55-74, but people aged 75+ were less overrepresented than in the other surveys, probably due to a lack of digital literacy. On the other hand, a slightly higher portion of those younger than 35 (18–24 and 25–34 age groups combined) was included in our study, and they were less underrepresented. Thirdly, relatively more educated than less educated respondents participate in surveys, and in my survey, the education-

related nonresponse bias was even more severe. The differences could potentially be attributed to people with a university degree being more likely to own a mobile phone (coverage bias); however, due to a high smartphone penetration rate (only about 9% of Australians are without a smartphone), education-related coverage bias can only explain a portion of the total representation bias. Moreover, all surveys in Australia seem to underestimate the portion of foreign-born residents of Australia with a similar magnitude. Finally, my sample was quite accurate in estimating distribution by state, and the difference between the Australian Census 2016 and my estimates was larger than 2%-points for the Australian Capital Territory only.

To sum up, low response rates in the RDD SMS-invitation web-push study and undercoverage of people without internet or smartphones seemed to lead to some demographic representation bias, but this can be reported for other probability surveys as well. The only notable difference between my sample and all other samples was in estimating education distribution, which could potentially be affected by an increasing differential nonresponse five years after the OPBS was carried out.

## 10.5 Discussion

The 2020s brings both good news and bad news for survey methodology. The bad news is that response is declining in most survey research, and low response rates in my survey were an example of that. Nevertheless, the good news is that the existing literature (e.g., Groves & Peytcheva 2008) explains how nonresponse is not necessarily associated with representation bias and does not affect data accuracy in survey research based on probabilistic principles. In this study, being prepared to deal with high nonresponse, my focus was on exploring both the bad news and good news perspectives of errors of representation in a relatively new approach to probabilistic sampling with an SMS invitation directing to an online survey. While nonresponse was even higher than initially predicted in the survey design phase (potential nonresponse bias) and selection was somewhat limited to SamplePages database (potential undercoverage bias), in the end they translated into representation bias comparable to the socio-demographic nonresponse bias in three probability-based surveys with significantly higher response rates. In comparison to a similar text-to-web study with a low response rate conducted by Bucher and Sand (2021), my study experienced similar representation bias for education (overrepresentation of the most educated), but not for age (overrepresentation of the oldest instead of the youngest). Moreover, studying non-demographic bias, namely attitudinal, behavioral, knowledge, and other factual bias, will be an important step for further determining the overall representation bias of the proposed approach.

I firstly showed that the expected response rates in a web survey with text message invitations can be very low. Based on a number of text message responses from recipients of my SMS invitations, it was

clear that potential respondents were skeptical that the SMS really came from the Australian National University, and that my survey was a legitimate academic research project. While the first page of the online questionnaire thoroughly explained the practical and ethical aspects of the research, the link-click numbers showed that only about one in ten recipients clicked on the link in the survey invitation. The solution to this problem could be sending a longer SMS and providing more information in these messages, but that could considerably increase data collection costs, especially if there was little increase in link-click rates.

Further, I presented evidence on how response in a study of this kind is difficult to increase substantially by using many approaches known to be effective in other survey modes, e.g., offering conditional incentives. One possible explanation of this could be that in a country with a high standard of living like Australia, $5 incentives in a form of a supermarket gift card do not represent enough value for respondents. At the same time, administering incentives to all respondents in my survey would double the costs per completed survey. It appears that the majority of respondents in this study participated for other reasons, which can be better explained with social exchange or cognitive dissonance theories (see Keusch 2015 for more information). This is further supported by the fact that only 8 of 44 '$5 conditional incentives' respondents decided to accept the coupon after filling out the survey. The other 36 respondents either did not wish to provide their details to be sent a $5 eGift card, or decided to enter the lottery for a $100 eGift card instead, which was an option given after they completed the questionnaire (and not in the introduction or the invitation SMS). It would be quite interesting to see the potential effect of providing $10 or $15 incentives on response, which are standard gift card amounts used by the only Australian probability-based panel Life in Australia™ for surveys of similar length (Kaczmirek et al. 2019), although this could significantly increase costs and attract more 'professional respondents'. On the other hand, a text-to-web survey offering higher incentives might attract more of the less educated (with lower income), which would help reduce some of representation bias.

Due to inefficiency in offering incentives, I later explored a couple of different potential response maximization solutions reported in the literature (e.g., Bosnjak et al. 2008). My findings on data collection characteristics affecting (and not affecting) response were consistent with findings from Andreadis (2020) on the use of pre-notification and reminders, and the day and time of SMS invitations. Sending an advance SMS proved to be the best solution to increase response by supposedly building trust with some respondents, even if being sent from a mobile number not associated with the University and only two hours in advance. The other fairly effective approach was a reminder SMS, which showed a similar increase in response rates. Since sending two or more text messages to each mobile number would exceed my budget, I did not test sending more than one SMS

reminder. For experimental groups receiving a reminder, the number of completes in two days after the first SMS (initial invitations) were fairly comparable to the number of completes in two days after the second SMS (reminders). This indicates that sending the second reminder SMS a couple of days later could result in a similar increase of response as the first reminder SMS, which was previously reported by Andreadis (2020). Moreover, I did not have a chance to combine two of the best approaches, texting an advance SMS and sending a reminder. Combining the best survey maximization approaches, as well as texting mobile numbers with appended geo-demographics, should be a subject of future research on this topic.

In this study, I was bounded by the survey budget. Thus, there was a limit to how many mobile numbers I could validate, text, pre-notify or remind, and I could only offer small conditional incentives. With a larger budget, my experimental groups would be of a sufficient size to work with more statistical power in order to identify the best approaches to this kind of data collection (e.g., the best day for texting invitations). However, my study adds value by presenting results of a sophisticated survey experiment, including many experimental groups. In the future, investigating response maximization approaches should be more targeted, while building on the results of this study; e.g., carrying out an individual study on the number of reminders offering the best cost-benefit balance for the data collector.

## 10.6 References


Andreadis, I. (2020). Text Message (SMS) Pre-notifications, Invitations and Reminders for Web Surveys. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2020-00019

Antoun, C., Conrad, F. G., Couper, M. P., & West, B. T. (2019). Simultaneous estimation of multiple sources of error in a smartphone-based survey. *Journal of Survey Statistics and Methodology, 7*(1), 93-117.

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

Australian Communications and Media Authority. (n.d.-a). Mobile-only Australia: living without a fixed line at home. Retrieved January 10, 2020, from https://www.acma.gov.au/publications/2020-12/report/mobile-only-australia-living-without-fixed-line-home

Australian Communications and Media Authority. (n.d.-b). *the Numbering System*. Retrieved December 1, 2020, from https://www.thenumberingsystem.com.au/#/smartnumbers-login

Balabanis, G., Mitchell, V. W., & Heinonen-Mavrovouniotis, S. (2007). SMS-based surveys: Strategies to improve participation. *International Journal of Advertising*, *26*(3), 369-385.

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

Bosnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., & Kaczmirek, L. (2008). Prenotification in web-based access panel surveys: The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review, 26*(2), 213-223.

Brick, J. M. (2008). Random-digit dialing. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 675-678). Sage.

Bucher, H., & Sand, M. (2021). Exploring the Feasibility of Recruiting Respondents and Collecting Web Data Via Smartphone: A Case Study of Text-To-Web Recruitment for a General Population Survey in Germany. *Journal of Survey Statistics and Methodology (2021) 00, 1–12.*

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

Conrad, F., Schober, M. F., Antoun, C., Hupp, A., & Yan, H. (2017). Text Interviews on Mobile Devices. In P. P. Biemer, E. D. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 299-318). John Wiley & Sons.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, *64*(4), 464-494.

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology, 43*, 121-145.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. In P. P. Biemer, E. D. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 133-154). John Wiley & Sons.

Datareportal. (2020, February 13). *Digital 2020: Australian.* https://datareportal.com/reports/digital-2020-australia

De Bruijne, M., & Wijnant, A. (2014). Improving response rates and questionnaire design for mobile web surveys. *Public Opinion Quarterly, 78*(4), 951-962.

Deloitte. (n.d.). *Mobile Consumer Survey 2019*. Retrieved December 1, 2020, from https://www2.deloitte.com/au/en/pages/technology-media-and-telecommunications/articles/mobile-consumer-survey.html

Dillman, D. A. (2018, October 2). *How Web-Push Surveys are Changing Survey Methodology* [Seminar presentation]. NatCen/ESS/City University methods seminar, City University of London, London, United Kingdom.

Dykema, J., Stevenson, J., Day, B., Sellers, S. L., & Bonham, V. L. (2011). Effects of incentives and prenotification on response rates and costs in a national web survey of physicians. *Evaluation & the health professions, 34*(4), 434-447.

Elevelt, A., Lugtig, P., & Toepoel, V. (2019). Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process?. *Survey Research Methods, 13*(2), 195-213.

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in human behavior*, *26*(2), 132-139.

Fordyce, E., Bilgen, I., & Stern, M. J. (2020, May 11-14). *The Use of Synchronous and Asynchronous Text Messaging During Survey Recruitment and Screening* [Conference presentation]. 75th annual conference of the American Association for Public Opinion Research, Virtual.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167-189.

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

Hsia, J., Zhao, G., & Town, M. (2020). Estimating Undercoverage Bias of Internet Users. *Preventing Chronic Disease 2020*, 17:200026. http://dx.doi.org/10.5888/pcd17.200026

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

Keeter, S., Hatley, N., Kennedy, C., & Lau, A. (2017). *What low response rates mean for telephone surveys.* Pew Research Center.

Keeter, S. (2019). *Growing and Improving Pew Research Center's American Trends Panel.* Pew Research Center.

Kellard, K. K. (2017, July 17-21). *What should (and shouldn't) we tell them? Qualitative pre-testing of different approaches for recruitment of a new probability panel* [Conference presentation]. 7th biennial conference of the European Survey Research Association, Lisbon, Portugal.

Keusch, F. (2015). Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, *65*(3), 183-216.

Knowles, S., & Stahlmann-Brown, P. (2021). Cash is not king in incentivizing online surveys. *Applied Economics Letters*, *28*(2), 105-108.

Kongsgard, H. W., Syversen, T., & Krokstad, S. (2014). SMS phone surveys and mass-messaging: promises and pitfalls. *Epidemiology: Open Access*, *4*(4). http://dx.doi.org/10.4172/2161-1165.1000177

Laguilles, J. S., Williams, E. A., & Saunders, D. B. (2011). Can lottery incentives boost web survey response rates? Findings from four experiments. *Research in Higher Education*, *52*(5), 537-553.

Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, *36*(1), 116-124.

Mavletova, A., & Couper, M. P. (2016). Device use in web surveys: The effect of differential incentives. *International Journal of Market Research, 58*(4), 523-544.

Morgan, A. J., Rapee, R. M., & Bayer, J. K. (2017). Increasing response rates to follow-up questionnaires in health intervention research: Randomized controlled trial of a gift card prize incentive. *Clinical Trials, 14*(4), 381-386.

Neal, Z., Neal, J. W., & Piteo, A. (2020). Call me maybe: using incentives and follow-ups to increase principals' survey response rates. *Journal of Research on Educational Effectiveness*, *13*(4), 784-793.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*(201), 101-116.

Parsons, N. L., & Manierre, M. J. (2014). Investigating the relationship among prepaid token incentives, response rates, and nonresponse bias in a web survey. *Field Methods, 26*(2), 191-204.

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

Phillips, B., & Compton, S. (2019). Text messages and reminder calls in student and alumni web surveys. *CSRM and SRC Methods Paper, 2019* (4).

Roy Morgan. (2015, January 22). *Over a million Australians now use two mobile phones*. http://www.roymorgan.com/findings/6024-a-million-australians-use-two-mobile-phones-november-2014-201501160100

Roy Morgan. (2016, August 22*). 9 in 10 Aussie teens now have a mobile (and most are already on to their second or subsequent handset)*. http://www.roymorgan.com/findings/6929-australian-teenagers-and-their-mobile-phones-june-2016-201608220922

Roy Morgan. (2018, November). *Kids and Mobiles, How Australian children are using mobile phones.* https://app.powerbi.com/view?r=eyJrIjoiNzA4Njc4YWMtZDk5OC00M2EwLWI3MDktMjkzMGRmOGNhOThkIiwidCI6IjBkYWM3ZjM5LWQyMGMtNGU3MS04YWYzLTcxZWU3ZTI2OGEyYiJ9&pageName=ReportSectionf48b9bb17d51b38a13e1

Sala, E., Respi, C., & Decataldo, A. (2018). Non response in web surveys. The role of SMS reminders. *Rassegna italiana di Sociologia*, *59*(1), 5-24.

Saleh, A., & Bista, K. (2017). Examining factors impacting online survey response rates in educational research: Perceptions of graduate students. *Online Submission*, *13*(2), 63-74.

SamplePages. (n.d.). *What we do*. Retrieved November 15, 2020, from https://samplepages.com.au/#service

Spreen, T. L., House, L. A., & Gao, Z. (2020). The impact of varying financial incentives on data quality in web panel surveys. *Journal of Survey Statistics and Methodology, 8*(5), 832-850.

StataCorp. (2013). *Stata Statistical Software: Release 13*. StataCorp LP.

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

Van Mol, C. (2017). Improving web survey efficiency: the impact of an extra reminder and reminder content on web survey response. *International Journal of Social Research Methodology, 20*(4), 317-327.

## Appendix 10

### Survey on Wellbeing, Health and Life in general 2020 online questionnaire (from *RDD-sampling SMS-invitation web-push study*)

In the last week, did you do any exercise which caused a moderate increase in your heart rate or breathing, that is, **moderate exercise**? (e.g., gentle swimming, social tennis, golf)

- ○ Yes (1)
- ○ No (2)

How many serves of vegetables do you usually eat each day?

- ○ 1 serve (1)
- ○ 2 serves (2)
- ○ 3 serves (3)
- ○ 4 serves (4)
- ○ 5 serves (5)
- ○ 6 serves or more (6)
- ○ less than one serve (7)
- ○ do not eat vegetables (8)

Have you had an alcoholic drink of any kind in the last 12 months?

- ○ Yes (1)
- ○ No (2)

How often do you now smoke cigarettes, pipes or other tobacco products?

- ○ Daily (1)
- ○ At least weekly (but not daily) (2)
- ○ Less often than weekly (3)
- ○ Not at all, but I have smoked in the last 12 months (4)
- ○ Not at all and I have <u>not</u> smoked in the last 12 months (5)

In general, would you say your health is...?

- ○ Excellent (1)
- ○ Very good (2)
- ○ Good (3)
- ○ Fair (4)
- ○ Poor (5)

---

How often do you feel rushed or pressed for time?

- ○ Always (1)
- ○ Often (2)
- ○ Sometimes (3)
- ○ Rarely (4)
- ○ Never (5)

---

In the past 4 weeks, about how often did you feel **hopeless**?

- ○ None of the time (1)
- ○ A little of the time (2)
- ○ Some of the time (3)
- ○ Most of the time (4)
- ○ All of the time (5)

---

Do you have access to the Internet at home?

- ○ Yes (1)
- ○ No (2)

---

Which, if any, of the following devices do you own or have ready access to (i.e., that is readily available for you to use)?

- [ ] Smartphone (1)
- [ ] Laptop (2)
- [ ] Desktop/tower computer (3)
- [ ] Tablet (4)

Thinking about new brands or technology, do you agree or disagree with the following statement?

*I like to be the first among my friends and family to try something new.*

- ○ Strongly agree (1)
- ○ Agree (2)
- ○ Disagree (3)
- ○ Strongly disagree (4)

In the last twelve months did you spend any time doing voluntary work through an organisation or group?

- ○ No, did not do voluntary work (1)
- ○ Yes, did voluntary work (2)

In the last two weeks did you spend time providing unpaid care, help or assistance to family members or others because of a disability, a long term health condition or problems related to old age?

- ○ No, did not provide unpaid care, help or assistance (1)
- ○ Yes, provided unpaid care, help or assistance (2)

How well do the following words describe you? For each word, indicate how well that word describes you. There are no right or wrong answers.

How well does **Warm** describe you?

- 1 - Does not describe me at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 - Describes me very well (7)

---

What about **Orderly**?

- 1 - Does not describe me at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 - Describes me very well (7)

---

How well does **Moody** describe you?

- 1 - Does not describe me at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 - Describes me very well (7)

---

What about **Quiet**?

- 1 - Does not describe me at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 - Describes me very well (7)

And finally, how well does **Philosophical** describe you?

- 1 - Does not describe me at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 - Describes me very well (7)

I am now going to ask you some questions about how satisfied or dissatisfied you are with some of the things happening in your life.

Please pick a number between 0 and 10 to indicate how satisfied you are with **your health.**

| | 0 - Totally dissatisfied (1) | 1 (2) | 2 (3) | 3 (4) | 4 (5) | 5 - Neither satisfied nor dissatisfied (6) | 6 (7) | 7 (8) | 8 (9) | 9 (10) | 10 - Totally satisfied (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Your health (1) | | | | | | | | | | | |

How satisfied are you with **your financial situation?**

| | 0 - Totally dissatisfied (1) | 1 (2) | 2 (3) | 3 (4) | 4 (5) | 5 - Neither satisfied nor dissatisfied (6) | 6 (7) | 7 (8) | 8 (9) | 9 (10) | 10 - Totally satisfied (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Your financial situation (1) | | | | | | | | | | | |

How satisfied are you with **how safe you feel?**

| | 0 - Totally dissatisfied (1) | 1 (2) | 2 (3) | 3 (4) | 4 (5) | 5 - Neither satisfied nor dissatisfied (6) | 6 (7) | 7 (8) | 8 (9) | 9 (10) | 10 - Totally satisfied (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| How safe you feel (1) | | | | | | | | | | | |

How satisfied are you with **your life as a whole**?

| | 0 - Totally dissatisfied (1) | 1 (2) | 2 (3) | 3 (4) | 4 (5) | 5 - Neither satisfied nor dissatisfied (6) | 6 (7) | 7 (8) | 8 (9) | 9 (10) | 10 - Totally satisfied (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Your life as a whole (1) | | | | | | | | | | | |

Now some questions about yourself to help me analyse results.

Are you…?

- ○ Male (1)
- ○ Female (2)

Which of the following age groups do you belong to…?

- ○ 18-24 (1)
- ○ 25-34 (2)
- ○ 35-44 (3)
- ○ 45-54 (4)
- ○ 55-64 (5)
- ○ 65-74 (6)
- ○ 75 and over (7)

---

What is the highest year of primary or secondary school you have completed?

- ○ Year 12 or equivalent (1)
- ○ Year 11 or equivalent (2)
- ○ Year 10 or equivalent (3)
- ○ Year 9 or equivalent (4)
- ○ Year 8 or below (5)
- ○ Did not go to school (6)

---

Have you completed any educational qualification (including a trade certificate)?

- ○ Yes (1)
- ○ No (2)

---

What is the level of the highest qualification you have completed?

- ○ Certificate I or Certificate II (1)
- ○ Certificate III or Certificate IV (2)
- ○ Associate Diploma (3)
- ○ Undergraduate Diploma (4)
- ○ Bachelor Degree (5)
- ○ Master’s Degree, Postgraduate Degree, or Postgraduate Diploma (6)
- ○ Doctorate (7)

---

What state do you live in?

- New South Wales (1)
- Victoria (2)
- Queensland (3)
- South Australia (4)
- Western Australia (5)
- Tasmania (6)
- Northern Territory (7)
- Australian Capital Territory (8)

Do you live in the capital city of your state or territory?

- Yes (1)
- No (2)

Are you an Australian citizen?

- Yes (1)
- No (2)

Were you born in Australia?

- Yes (1)
- No (2)

Do you speak a language other than English at home?

- Yes (4)
- No (5)

Are you of Aboriginal or Torres Strait Islander origin?

- No (1)
- Yes, Aboriginal (2)
- Yes, Torres Strait Islander (3)

Did you live at your current address 5 years ago (in November 2015)?

- ○ Yes (1)
- ○ No (2)

---

We're interested in whether your home is owned by you or members of your household. Is the dwelling that you live in…?

- ○ Owned outright (1)
- ○ Owned with a mortgage (2)
- ○ Being purchased under a rent/buy scheme (3)
- ○ Being rented (4)
- ○ Being occupied rent free (5)
- ○ Being occupied under a life tenure scheme (6)
- ○ Other (8)

---

Which category best describes this household?

- ○ Person living alone (1)
- ○ Couple living alone (2)
- ○ Couple with non-dependent child(ren) (3)
- ○ Couple with dependent child(ren) (4)
- ○ Couple with dependent and non-dependent child(ren) (5)
- ○ Single parent with non-dependent child(ren) (10)
- ○ Single parent with dependent child(ren) (7)
- ○ Single parent with dependent and non-dependent child(ren) (8)
- ○ Non-related adults sharing house/apartment/flat (9)
- ○ Other household type (11)

---

Last week, did you have a job of any kind?

- ○ Yes, worked for payment or profit (1)
- ○ Yes, but absent on holidays, on paid leave, on strike, or temporarily stood down (4)
- ○ Yes, unpaid work in a family business (3)
- ○ Yes, other unpaid work (5)
- ○ No, did not have a job (6)

---

What is the total of all income you usually receive?

- ○ 3,000 or more per week / $156,000 or more per year (1)
- ○ $2,000 - $2,999 per week / $104,000 - $155,999 per year (2)
- ○ $1,750 - $1,999 per week / $91,000 - $103,999 per year (3)
- ○ $1,500 - $1,749 per week / $78,000 - $90,999 per year (4)
- ○ $1,250 - $1,499 per week / $65,000 - $77,999 per year (5)
- ○ $1,000 - $1,249 per week / $52,000 - $64,999 per year (6)
- ○ $750 - $999 per week / $39,000 - $51,999 per year (7)
- ○ $500 - $749 per week / $26,000 - $38,999 per year (8)
- ○ $250 - $499 per week / $13,000 - $25,999 per year (9)
- ○ $1 - $249 per week / $1 - $12,999 per year (10)
- ○ Nil income (11)
- ○ Negative income (12)

---

# Chapter 11 Accuracy in RDD-mobile-sampling SMS-invitation web-push survey: Empirical evidence and a TSE-based methodological framework for benchmarking analysis

## 11.1 Introduction

Widespread access to the internet on mobile devices as well as a growing market of smartphones and tablets, offer both new opportunities and challenges for survey methodology. Simultaneously, response rates have been gradually decreasing and, in turn, academic, government, social, and market research have had to face increasing costs of data collection (Couper 2017; Stedman et al. 2019).

This survey project aims to evaluate a rather new approach to smartphone data collection (Bucher & Sand 2021, p. 10) from the Total Survey Error (TSE) framework perspective (see Groves et al. 2009). Survey data has been collected to investigate errors of representation of a less common approach to survey data collection, combining RDD-mobile sampling with SMS invitation and web-push, and is as such based on probabilistic sampling principles. The major sources of that combined survey error could be undercoverage as some people do not have access to the internet and/or a smartphone (Couper 2000), differential unit nonresponse as a result of a digital divide and a relatively high nonresponse to web surveys (e.g., Daikeler et al. 2020), and sampling bias as a result of a specific approach to sampling in this survey research project. These sources then combine errors into representation bias which can be ultimately reflected in less accurate survey estimates.

Considering the TSE framework as the basis, this article presents evidence on data accuracy of the proposed approach by comparing the estimates of this survey to the nationally representative benchmarks measuring health, wellbeing, technology, life satisfaction, personality traits, and primary and secondary demographics (see the questionnaire in the Appendix 10). This study will be a benchmarking study in which the relative accuracy of the sample will be compared to three probability and one nonprobability sample for several items, and answer the following research question:

*RQ1: How accurate are data collected with an RDD-mobile-sampling SMS-invitation web-push survey?*

Further, as there are notable methodological differences in how benchmarking studies have been conducted until now, four other surveys using the same questionnaire will be used to critically evaluate the approach to this benchmarking analysis, and outline a TSE-based framework for carrying out benchmarking analysis as a tool for assessment of data quality in survey research with imperfect survey design. The different conventions reported in the literature will be tested in terms of how

benchmarking analysis should be carried out in a methodologically efficient manner to not introduce nonignorable bias. The following research question will be answered:

*RQ2: How do different decisions made in the design of a benchmarking study, such as selection of representative data sources, benchmarks and their benchmark categories, sample sizes, and weighting, affect the results of the accuracy of a survey in a benchmarking study?*

## 11.2 Background

### 11.2.1 Errors of representation in web, smartphone, and SMS surveys

The most commonly used framework for studying survey errors and data quality perspectives of surveys is the TSE paradigm. In this study, the TSE is used to define the most notable potential sources of representation bias in RDD-mobile-sampling SMS-invitation, web-push survey (shorter: RDD SMS web-push), which is fundamentally a hybrid of SMS surveys and web surveys on smartphones. Although measurement error plays an important role in mobile surveys (Couper et al. 2017), studying it in detail is beyond the scope of this study.

#### 11.2.1.1 Coverage error

Survey researchers are concerned about how well the sampling frame covers the target population, and undercoverage is the weakness of the frame with the highest threat of coverage error, which is defined as the effect of problems of a sampling frame on a survey estimate (Groves et al. 2009). Coverage bias can be estimated as follows (see Equation 11.1):

$$\overline{Y_C} - Y = \frac{U}{N} (\overline{Y_C} - \overline{Y_U}) \qquad (11.1)$$

where $\overline{Y_C}$ is the mean of the covered population, $\overline{Y_U}$ is the mean of the population not covered with a size U, and $\bar{Y}$ is the mean of the total population with a size of N (Groves et al. 2009). Web surveys are known for undercoverage of people without access to or use of the internet (Couper 2000), and the bias is more severe if the difference in estimates between online ($\overline{Y_C}$) and offline populations ($\overline{Y_U}$) is larger and if the offline population of size U represents a larger portion of the total population of size N (associated with internet penetration). However, Groves et al. (2009) also explained that bias $\overline{Y_C} - Y$ is a property of survey statistics $\bar{Y}$, which means that not all survey items are affected in the same way by undercoverage and other sampling frame problems.

In practice, coverage error appears to be the prevalent component of representation bias. Hsia et al. (2020) found different levels of undercoverage bias for their health-related items from their web survey (overall bias), but they could report bias within some demographics and not all of them (relative

bias). Moreover, Fuchs and Busse (2009) reported considerably large socio-demographic coverage bias in a smartphone survey, consistent with the findings of Antoun et al. (2019), who concluded that while smartphone surveys might represent an advanced data collection opportunity, they can suffer from more undercoverage than web surveys. Similar claims have been made for text message surveys, which suffer from selection bias and data quality (Kongsgard et al. 2014), and not every person has a cell phone or a smartphone (McGeeney & Kennedy 2015). In RDD text-to-web surveys, one of the issues can be excluding people with cellphones that are not smartphones as the recipients of an SMS cannot access the online questionnaire (Bucher & Sand 2021).

#### 11.2.1.2 Sampling error

Due to the gap between the sampling frame and the sample selected from the frame, survey researchers have to deal with sampling error; this study distinguishes between sampling bias affected by the assignment of probabilities of selection to elements in the frame, and sampling variance, affected by the sample size and type of sampling (simple random sampling, stratified sampling, clustering, etc.; Groves et al. 2009). In contrast to nonprobability sampling, which is (very) often associated with web surveys (Callegaro et al. 2015), each unit of the population has to have a known non-zero chance of selection in probability sampling (Neyman 1938). Due to the unavailability of sampling frames with e-mail addresses required for probability-based sampling in general population surveys, or an ability to assemble random e-mail addresses (Fricker 2008), “offline” random sampling is predominantly carried out in probability-based online survey recruitment (Callegaro et al. 2015) to mitigate sampling bias, such as (single frame mobile) random digit dialing (Kennedy et al. 2018).

#### 11.2.1.3 Nonresponse error

Nonresponse bias is associated with members of the chosen sample not responding to any questions (unit nonresponse) or particular questions (item nonresponse) in a survey (Merkle 2008). Nonresponse error can be observed when the value of statistic(s) computed with all members of the chosen sample differ from those computed with the respondent data only (Groves et al. 2009). Generally, higher response rates can reduce the risk of bias, but Groves and Peytcheva (2008) also identified some surveys with low nonresponse rates and high relative response bias, and those with high nonresponse and low bias in survey estimates. This indicates that a higher response does not guarantee lesser nonresponse bias, which can be expressed as follows in Equation 11.2:

$$\overline{y_r} - \overline{y_s} = \frac{m_s}{n_s} (\overline{y_r} - \overline{y_m}) \qquad (11.2)$$

where $\bar{y}_s$, $\bar{y}_r$, and $\overline{y_m}$ are the means of the entire specific sample ($s$ of sample size $n_s$), respondents ($r$), and nonrespondents ($m$ of sample size $m_s$), respectively (Groves et al. 2009).

Web surveys are generally known for lower response rates than other survey modes (Cook et al. 2000; Daikeler et al. 2020; Manfreda et al. 2008), and since nonrespondents represent a larger portion of the entire sample ($\frac{m_s}{n_s}$), bias could be more severe. However, if the difference in estimates between nonrespondents ($\bar{y}_r$) and the entire sample ($\overline{y_m}$) is small, nonresponse bias can be negligible. In comparison to web surveys, nonresponse and the associated bias can be more significant in mobile web surveys (Antoun et al. 2019; Couper et al. 2017), which can also be attributed to higher breakoff rates in such types of surveys (Mavletova & Couper 2016) or the inability of respondents to participate via a smartphone and choosing not to respond on a PC (Peterson et al. 2017). Finally, in their RDD text-to-web survey, Bucher and Sand (2021) reported a response rate of less than 1% (AAPOR RR1).

#### 11.2.1.4 Adjustment error

Adjustment error is tied to post-survey adjustments, such as post-stratification weighting, carried out to improve sample estimates affected by the other three sources of representation bias (Groves et al. 2009) predominantly by noncoverage and nonresponse. While post-survey adjustments are meant to decrease bias, weighting at the same time generally inflates variance of estimates, particularly in the case of large weighting adjustments (Kalton & Flores-Cervantes 2003). In particular scenarios, weighting might have no positive effect on accuracy, while negatively affecting the precision of the results; however, survey statisticians routinely weigh data for all analyses (Groves et al. 2009).

Post-stratification weighting, which is adjusting the sampling weights to the known population totals and is one of the most commonly used techniques, does not always improve the accuracy of estimates in practice and it works better with probability than nonprobability online samples. For example, Loosveldt and Sonck (2008) reported a minor impact of weighting adjustment to the differences in results between nonprobability and probability surveys, and Yeager et al. (2011) and Pennay et al. (2018) both concluded that post-stratification increases the accuracy of probability samples more consistently than nonprobability online samples. Currently, nonprobability samples are often adjusted post-survey using other methods such as propensity weighting or matching methods (Mercer et al. 2018). The accuracy of results can also depend on the selection of weighting variables. For example, Kennedy et al. (2016) found that online samples only balanced on age, while gender and region were less accurate than those also weighted by education and income. All the studies mentioned above compared the accuracy of survey samples and weighting approaches using benchmarking analysis.

### 11.2.2 Benchmarking as an accuracy estimation method

In survey methodology, benchmarking is a method used in the estimation of accuracy of different surveys, survey samples, survey modes, and post-survey adjustments to name a few. It can be argued that it is directly linked to the TSE framework as benchmarking analysis is fundamentally a calculation of either: (1) the combined survey error, or (2) a particular type of bias for a survey item. Having known that nationally representative sample survey estimates also come with a (smaller) error (Pennay et al. 2018; Yeager et al. 2011), the distance between a benchmark and a survey item statistic is used in that case as the best estimate of the total error.

In practice, researchers carry out benchmarking to determine the quality of their surveys (e.g., Bialik 2018) to compare surveys based on different types of sampling (e.g., Yeager et al. 2011), and to study measurement mode effects, although those studies are not called benchmarking studies per se (Vannieuwenhuyze & Loosveldt 2013). The most common benchmarking analysis that has been used in the literature is the accuracy of nonprobability samples in comparison to probability samples. While the results of several benchmarking studies of that type have been published (e.g., Chang & Krosnick 2009; Dutwin & Buskirk 2017; Kaczmirek et al. 2019; MacInnis et al. 2018; Malhotra & Krosnick 2007; Pennay et al. 2018; Yeager et al. 2011), and the basic principles of benchmarking analysis have been described and discussed, there is no commonly used benchmarking methodological framework advising survey researchers on how to perform benchmarking in a methodologically sound and unbiased manner.

Generally, some principles on how to design and carry out benchmarking have been used consistently across all studies, and there is less consistency for some other methodological aspects. Further, particular methodological decisions made in those benchmarking studies could be considered arbitrary and might have introduced nonignorable bias. First, there appears to be a consensus that population benchmarks should come from high-quality data sources, such as large-scale national government/federal surveys (e.g., Pennay et al. 2018; Yeager et al. 2011), censuses (Kaczmirek et al. 2019; Pennay et al. 2018), or government administrative data such as data on driver's licenses (Yeager et al. 2011) or data on electoral enrollment (Pennay et al. 2018). Second, benchmarks from "gold standard" surveys are predominantly selected based on their availability and/or convenience, which is often a result of designing a benchmarking study after having already collected survey data. Further, in Pennay et al. (2018) and Kaczmirek et al. (2019), similar benchmarks (and categories) were selected as in Yeager et al. (2011) as they replicated the US benchmarking study in Australia. Yeager et al. (2011) also indicated that in the best-case scenario, benchmarks would be selected from the available high-quality sources randomly. Further, Dutwin and Buskirk (2017) suggested avoiding items likely to be

subject to measurement error, such as satisficing or social desirability. Third, there is no consensus on the survey item categories to choose for benchmarking analysis. While this problem does not exist for binary variables, two different approaches have been used in practice for nominal and ordinal variables with 3+ categories: (1) using all categories in the calculation of error (e.g., Chang & Krosnick 2009), (2) using the modal category (e.g., MacInnis et al. 2018). To address this issue, Yeager et al. (2011) applied the second approach, but confirmed that the first approach would have led to the same conclusions. Four, the majority of benchmarking studies approached weighting in a similar manner by presenting both weighted (post-stratification) and unweighted results, while Dutwin & Buskirk (2017) also presented results with nonprobability samples after propensity scoring and matching. Finally, a number of different benchmarking measures have been used to compare accuracy, such as average absolute error (AAE; e.g., Yeager et al. 2011), standard deviation of the AAE, maximum AAE (Dutwin & Buskirk 2017), ranking of AAE, number of statistically significant differences from benchmarks (e.g., Pennay et al. 2018), and root mean squared error (RMSE) (MacInnis et al. 2018).

## 11.3 Methods

In this section, the methodological aspects of this benchmarking study are presented, including the studied population, sampling, data collection specifics, selected benchmarks, and empirical analysis carried out to evaluate the existing approaches commonly applied in benchmarking studies.

### 11.3.1 Data

As this is a benchmarking study, both survey sample data and different high-quality representative sources of benchmarks are used. This means that the data are analyzed from the following:

- RDD SMS Web-push Survey (2020) – the main subject of this research;
- Online Panels Benchmarking Study collected by the Social Research Centre (2015; Pennay et al. 2018) – four reference samples; generally, these were less convenient and more expensive surveys with samples of similar size to the RDD SMS Web-push Survey;
- National Health Surveys 2014–2015 and 2017–2018; National Drug Strategy Household Surveys 2013, 2016, and 2019; Australian Censuses 2011 and 2016; General Social Survey 2014 and 2019; and Household, Income and Labour Dynamics Australia Release 18 data (Waves 13, 17 and 18[77]) – high-quality nationally representative surveys as sources of benchmarks.

[77] Household, Income and Labour Dynamics Australia (HILDA) is a the highest-quality Australian household-based panel study. However, as it is a longitudinal study and a subject to panel attrition (and possibly panel conditioning), it might be less representative of the Australian general population than the other government surveys used as data sources of benchmarks in this study.

The RDD SMS Web-push Survey data are the main subject of this study as the aim of this project was to assess the accuracy of this less commonly used mixture of approaches to data collection. Data were collected in November 2020 using the online survey mode. Randomly selected respondents (single frame RDD) received a text message (SMS) to their mobile devices with an URL to an online questionnaire programmed in Qualtrics (web-push). The online *Survey on Wellbeing, Health and Life in general 2020* questionnaire included a variety of questions about people's lives with a total of 26 secondary demographics and non-demographic variables[78] with corresponding benchmarks from nationally representative sources listed above. The median response time was 8 minutes, the total sample size was 632 respondents, and the final AAPOR RR4 (The American Association for Public Opinion Research 2016) response rate was 1.9%.

The methodological information about the Online Panels Benchmarking Study (OPBS) data and on the high-quality government surveys as sources of benchmarks is available in the Appendix 11 (also in Table 11.2, Table 11.3, and Pennay et al. 2018).

### 11.3.2 Population and sampling

The population in the RDD SMS Web-push Survey and OPBS surveys was defined as "Australian residents aged 18 or older." The benchmarks from high-quality surveys were also calculated for the adult population of Australia, excluding residents younger than 18 years of age for comparability purposes. In the case of the Australian Censuses, every person in Australia on Census Night was enumerated, excluding Australian residents out of the country on that night (Australian Bureau of Statistics, n.d.).

The sample was in principle selected using random digit dialing of mobile numbers (i.e., single frame RDD mobile sampling). In Australia, 62 million mobile number combinations are allocated to Vodafone, Optus, and Telstra as the commercial mobile service providers. Taking into account that the population size was approximately 25.7 million in June 2020 (Australian Bureau of Statistics, n.d.) and that some Australian residents do not own a smartphone (Roy Morgan 2018), only approximately one-third of all possible combinations are active/live mobile numbers. To validate mobile numbers and to append geo-demographics of mobile owners where available (i.e., age, gender, and information on statistical area), services provided by SamplePages (n.d.) who matched the randomly generated

[78] 14 benchmarked matched those from OPBS; 5 secondary demographic benchmarks, such as those requiring postcodes to be derived (e.g., major city) or those from administrative sources (enrolment to vote), were purposely replaced with substantive items measuring other lifestyle dimensions; one indicator of Kessler 6 (K-6) psychological distress was used instead of the whole index.

numbers to the records in their database were used. The total sample size of the validated mobile numbers was 34,734.

For sampling information about the OPBS surveys, please see Table 11.2 in the Appendix 11 (as well as Pennay et al. 2018). For sampling information about the government-funded nationally representative data sources of benchmarks, please see Table 11.3 in the Appendix 11.

### 11.3.3 Benchmarking analysis

To analyze the accuracy of the RDD SMS Web-push Survey and answer the research question 1 (RQ1), the most commonly used principles on benchmarking analysis are followed, which were carefully planned as early as in the questionnaire design stage of the study. As the sample required reference samples to compare its accuracy relative to similar sampling and data collection approaches in Australia, many of those principles closely resembled the benchmarking design in the OPBS. First, all benchmarks were from "gold standard" data sources funded by the Australian Government, and 18 out of 26 have corresponding benchmarks from more than one data source (e.g., *general health status:* National Health Survey 2017–18 and National Drug Strategy Household Survey 2019, see Table 11.4). Second, benchmarks were not chosen randomly – 14 out of 26 survey items matched the OPBS survey items, and the rest of them were carefully chosen from the high-quality data sources to expand the range of studied dimensions of peoples' lives (i.e., personality traits, time stress, eating habits, exercising, and caring responsibilities). The number of benchmarked items was limited to not prolong the questionnaire and risk an even lower response rate. The modal category for non-binary categorical variables was used to present both weighted (post-stratification[79]) and unweighted results. Among all the benchmarking measures proposed in the literature, the average absolute error (AAE) from Yeager et al. (2011) was used as it is the most commonly used measure in practice, and root mean squared error (RMSE) from MacInnis et al. (2018) was calculated as presented below (see Equations 11.3 and 11.4):

$$AAE = \sum_{j=1}^{k} \frac{|\widehat{y_j} - y_j|}{k} \qquad (11.3)$$

and

$$RMSE = \sqrt{\frac{\sum_{j=1}^{k} (\widehat{y_j} - y_j)^2}{k}} \qquad (11.4)$$

[79] RDD SMS Web-push: raking by gender, age*education, country of birth (Australia, abroad), state; OPBS probability-based surveys: design weight and raking by gender, age*education, country of birth (Australia, English speaking country, Non-English-speaking country), state*capital city; OPBS Online panel 2: raking by gender, age*education, country of birth (Australia, English-speaking country, non-English speaking country), state*capital city

where $\hat{y}_j$ is the j-th estimate from the sample surveys and $y_j$ is the value for a corresponding benchmark. The decision on adding an RMSE measure was based on the fact that it is more sensitive to larger errors from the benchmark and better illustrates potentially higher variability or errors.

The second part of the analysis targeted to answer research question 2 (RQ2) and demonstrate how different decisions in the design of a benchmarking study, associated with the TSE, may or may not lead to different results, interpretations, or even conclusions. Thus, this study demonstrates and discusses the potential effects of the following:

- the selection of data sources for non-demographic items not available in Censuses: in particular, the effect of time gap between the analyzed and government surveys, and the effect of methodological differences between different high-quality sources (e.g., mode, sample size);
- the selection of benchmarks: in particular, the effect of the survey item topic, the effect of the fast-changing results over time in combination with the time gap defined above, and the effect of social desirability and satisficing;
- the selection of survey item categories: in particular, the effect of benchmarking across all categories of categorical variables, and the effect and item modal category proportion on the random error;
- size differences of the analyzed samples (demonstrated with Monte Carlo simulation in R (R Core Team 2020));
- differences in post-stratification weighting adjustments.

## 11.4 Results

### 11.4.1 Accuracy in RDD SMS Web-push Survey

In this section, the results of the benchmarking analysis with the RDD SMS Web-push Survey sample and four OPBS samples are presented. To compare the overall accuracy of different samples, AAE and RMSE scores were calculated for all samples for unweighted and weighted (raked) data.

The results from Table 11.1 show the survey estimates for individual survey items. The errors were calculated separately for RDD SMS Web-push and OPBS surveys using different sets of benchmarks as there was a 5-year gap in data collection periods. Reviewing changes in estimates from nationally representative data sources (e.g., *Australian citizen* from Australian Censuses 2011 (83.9%) and 2016 (87.1%)) proves that this was necessary to decrease bias. The errors were then combined into sample-level measures AAE and RMSE scores for the items appearing in all surveys to estimate the Total Survey

Error (TSE). For the RDD SMS Web-push sample, AAE and RMSE were also calculated for all 26 items with corresponding benchmarks.

The analysis offered some very interesting evidence on the accuracy of different surveys. First, RDD SMS Web-push appears to be a survey approach collecting data of a similar quality to other cross-sectional probability-based surveys as well as the most accurate nonprobability online panel. The AAE for "OPBS" items and unweighted estimates was approximately 1% point higher for the RDD SMS Web-push Survey and identical to the AAE of the Opt-in panel 2 (no statistically significant differences). After weighting, which slightly improved accuracy, the differences in errors between probability-based samples remained, and again, there were no statistically significant differences in AAE. Since raking did not improve the accuracy of the Opt-in panel 2 data and only slightly improved the accuracy of RDD SMS Web-push estimates, the AAE scores after weighting resulted in being higher than AAE scores of OPBS probability samples (statistically significant difference).

Table 11.1: Accuracy of RDD SMS Web-push sample in comparison to OPBS samples (and benchmarks)

| Survey item with available benchmark(s) | RDD SMS Web-push Survey (2020) | | | Online Panels Benchmarking Study (OPBS) (2015) *(taken from Pennay et al. 2018, 16-22)* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bench[a] | Survey estimates | | Bench[b] | Survey estimates | | | | | | | |
| | | | | | RDD | | A-BS | | ANUpoll | | Opt-in panel 2 | |
| | | UW | W | | UW | W | UW | W | UW | W | UW | W |
| Australian citizen | 87.1 | 92.2 | 85.0 | 83.9 | 91.0 | 86.6 | 94.4 | 92.0 | 92.3 | 86.6 | 90.5 | 88.1 |
| Couple with dependent children | 28.4 | 16.6 | 22.4 | 38.4 | 22.8 | 27.9 | 21.0 | 28.2 | 23.4 | 27.0 | 25.8 | 24.0 |
| Currently employed | 61.6 | 60.3 | 61.7 | 59.4 | 58.2 | 69.3 | 57.4 | 64.6 | 60.5 | 66.4 | 54.3 | 53.3 |
| Home ownership with a mortgage | 28.8 | 29.7 | 29.3 | 29.6 | 31.0 | 33.8 | 32.3 | 40.0 | 33.6 | 37.4 | 30.2 | 30.9 |
| Not Indigenous | 97.7 | 97.9 | 97.6 | 98.1 | 98.8 | 98.8 | 98.0 | 98.4 | 98.4 | 98.4 | 96.5 | 96.5 |
| Language other than English (*Speak only English*) | 76.5 | 81.7 | 76.5 | 75.7 | 84.2 | 85.5 | 81.2 | 80.4 | 87.0 | 84.5 | 84.2 | 85.4 |
| Living at last address 5 years ago | 56.9 | 58.8 | 48.0 | 54.8 | 69.6 | 62.1 | 69.1 | 54.7 | 67.5 | 58.4 | 61.0 | 58.1 |
| Voluntary work (N*one*) | 79.4 | 61.3 | 62.8 | 74.2 | 58.2 | 62.7 | 60.8 | 63.0 | 60.2 | 62.6 | 73.8 | 77.1 |
| Wage and salary income (*$1000–1249 per week*) | 14.6 | 11.8 | 15.5 | 13.8 | 10.0 | 11.8 | 14.1 | 12.8 | 14.3 | 15.0 | 12.1 | 12.8 |
| Consumed alcohol in last year | 81.0 | 83.1 | 82.7 | 81.9 | 82.2 | 85.9 | 82.5 | 85.5 | 84.5 | 84.8 | 75.8 | 76.6 |
| Daily smoker | 11.6 | 11.6 | 13.3 | 13.5 | 10.3 | 15.1 | 9.1 | 9.4 | 12.5 | 17.0 | 20.2 | 20.2 |
| General health status (*Very good*) | 35.5 | 28.9 | 29.5 | 36.2 | 30.6 | 33.6 | 34.4 | 36.6 | 33.8 | 34.2 | 32.7 | 30.4 |
| Psychological distress, Kessler 6 - Hopeless (*Never*) | 77.7 | 43.9 | 38.6 | 80.0* | 70.9 | 69.9 | 69.3 | 64.6 | 68.2 | 65.0 | 48.0 | 50.3 |
| Satisfaction with: Life as a whole (*8 out of 10*) | 33.0 | 24.3 | 23.3 | 32.6 | 34.6 | 34.5 | 30.1 | 30.6 | 31.3 | 30.6 | 20.2 | 21.0 |
| Satisfaction with: Health (*8 out of 10*) | 27.5 | 18.0 | 19.0 | | | | | | | | | |
| Satisfaction with: Financial situation (*8 out of 10*) | 21.8 | 19.3 | 15.0 | | | | | | | | | |
| Satisfaction with: Safety (*8 out of 10*) | 28.6 | 18.4 | 18.5 | | | | | | | | | |
| Carer status (*Yes*) | 12.3 | 35.9 | 33.1 | | | | | | | | | |
| Moderate exercise in the last week | 30.0 | 71.3 | 70.1 | | | | | | | | | |
| Number of serves of vegetables each day (*2 serves*) | 28.2 | 25.5 | 29.1 | | | | | | | | | |
| Time stress (*Always/often*) | 39.6 | 38.4 | 40.4 | | | | | | | | | |
| Big 5 personality traits, Agreeableness - Warm (*5 out of 7*)* | 30.3 | 31.4 | 31.0 | | | | | | | | | |
| Big 5 pers. traits, Conscientiousness - Orderly (*5 out of 7**) | 24.6 | 26.9 | 26.0 | | | | | | | | | |
| Big 5 personality traits, Emotional stability - Moody (*2 out of 7**) | 24.5 | 23.3 | 22.5 | | | | | | | | | |
| Big 5 personality traits, Extroversion - Quiet (*4 out of 7**) | 22.0 | 20.1 | 18.3 | | | | | | | | | |
| Big 5 personality traits, Openness - Philosophical (*4 out of 7**) | 21.3 | 19.9 | 22.0 | | | | | | | | | |
| **AAE (OPBS items)** | | 7.04 | 6.67 | | 6.37 | 5.63 | 6.17 | 5.47 | 6.18 | 5.71 | 7.42 | 7.40 |
| **RMSE (OPBS items)** | | 11.34 | 12.13 | | 8.38 | 6.81 | 8.38 | 7.19 | 8.24 | 7.21 | 10.74 | 10.36 |
| **AAE (all items)** | | 7.59 | 7.30 | | | | | | | | | |
| **RMSE (all items)** | | 12.84 | 12.93 | | | | | | | | | |

[a] from data collected as closely in time to 2020 as possible: Australian Census 2016, National Health Survey 2017-18, National Drug Strategy Household Survey 2019, HILDA Release 18 (2018), and General Social Survey 2019; [b] from Pennay et al. (2018): Australian Census 2011, National Health Survey 2014-15, National Drug Strategy Household Survey 2013, and General Social Survey 2014; UW – unweighted, W – weighted (post-stratification), Bench – Benchmark; *added from the National Health Survey 2014-15 (K-6 score from OPBS was replaced with a single indicator)

In terms of the effects of post-survey adjustment and adjustment error, the results indicate that RDD SMS Web-push Survey is slightly more similar to probability than nonprobability surveys. As the post-survey adjustment, raking improved accuracy for 18 out of 26 items, the overall improvement was smaller in comparison to OPBS probability samples.

Further, the RMSE measure for OPBS items, both weighted and unweighted, provided slightly different evidence. As the measure penalizes samples with larger single-item errors, the RDD SMS Web-push sample was no longer as accurate as the OPBS probability samples; it was about as accurate as the opt-in panel sample. A detailed review of survey estimates as contributors to the total RMSE highlighted *Psychological distress – Hopeless*[80] as the item significantly increasing RMSE for RDD SMS Web-push and opt-in samples. As the sample was the most accurate for 7 out of 14 survey items, excluding the Kessler 6 indicator would result in RDD SMS Web-push having the lowest average errors among all the samples. There could be two explanations for the very inaccurate measurement of psychological distress with this sample: (1) the survey was, in that regard, more similar to nonprobability-based surveys, which generally over-estimate psychological distress in the general population; (2) an effect of COVID-19 pandemic on mental health in 2020[81]. Most probably, these two and the adjustment error were combined. Unfortunately, the results can only be interpreted in terms of the total survey error, and one cannot disentangle between representation error (coverage and nonresponse bias), changes of population statistics over time, or even measurement error.

Moreover, the "non-OPBS" estimates for RDD SMS Web-push provided further evidence on representation bias. While *satisfaction* items appear to be measured with some error, the whole distribution should be observed more closely, not only the modal category. A similar conclusion can be made for *personality items* with a difference that those survey estimates were very close to their benchmarks (i.e., within the margin of error). Further, two items are prominent for very large errors. The first one is *carer status*. In all OPBS probability samples and the RDD SMS Web-push sample, *volunteering* was over-estimated by approximately 100% (relative to the Australian Census 2016 benchmark), and in the sample used in this study, *caring* is over-estimated by more than 150% (raked estimate). Altruistic motivation is common in those concepts, which is evident in activities helping other people, and survey participation for no or very small financial benefits can be categorized under the umbrella of altruistic acts. Consequently, it can be argued that non-government probability-based

[80] This Kessler 6 item was chosen as it was an indicator with the highest correlation with the combined K-6 score index (in the OPBS study).

[81] The data collection for RDD SMS Web-push Survey was planned for April/May 2020, but it had to be postponed until easing of restrictions in all states in Australia to decrease the effect of a pandemic on people's lives and consequently, survey estimates.

surveys with low response rates can introduce large nonresponse bias for items measuring any type of altruistic behavior.

The second item with an even larger error is *moderate exercise in the last week*. After a thorough review of potential sources of error, it can be concluded that the term "moderate" exercise was interpreted incorrectly by a large portion of respondents, although the wording of the original question had not been changed. The estimate from the RDD SMS Web-push Survey is much closer to an estimate from the National Health Survey (NHS) 2017–2018 for both walking and moderate exercise (such as playing golf) combined. In the NHS questionnaire, respondents were first asked about walking, including its frequency and time spent, which was followed by instructions to exclude any walking when answering the following questions on moderate exercise. This is something that would be challenging to incorporate fully in a short smartphone survey. However, including a question without the required introduction and a proper definition (except for "e.g., gentle swimming, social tennis, golf"), the result was a large measurement error related to the questionnaire and question design as well as potential interviewer effect. This is a textbook example of how much attention should be paid to the design of a benchmarking study, which will be discussed more elaborately in the next paragraphs.

### 11.4.2 Effect of benchmarking methodology on results

In this section, different methodological conventions and decisions in benchmarking studies are critically evaluated with a focus on empirical results from this study. Evidence will be presented to answer research question 2 (RQ2).

#### 11.4.2.1 Selection of data sources

To assess the accuracy of the RDD SMS Web-push sample, data sources were selected as closely in time to survey data collection periods as possible, while choosing the most reliable source of a benchmark for a survey item. Comparing data sources for the same benchmarks (see Table 11.4), it can be concluded that:

- estimates from the same surveys/censuses change over time as a result of changes in society, which can add (or reduce) absolute item errors in benchmarking studies – for example, eight benchmarks from the Australian Census on average changed by 1.9%-point between 2011 and 2016 (see Table 11.4) and as much as 5.2%-points (*volunteering*);
- estimates from sample surveys might come with more error than expected – for example, assuming that the Australian Census is the most accurate source of primary and secondary demographic benchmarks, the General Social Survey (GSS) samples significantly over-

estimated volunteer work (25.8% and 20.6% in Censuses, 30.9% and 28.8% in GSS 2014 and 2019[82]).

#### 11.4.2.2 Selection of survey items as benchmarks

Reviewing the literature, the results from Table 11.1, and the list of benchmarks and their sources in Table 11.4, the following potential issues related to the selection of benchmarks were identified:

- non-random selection or even purposive selection of benchmarks;
- survey items are associated with survey participation/nonresponse (e.g., *volunteering*);
- survey items are strongly associated with mode of data collection in a study primarily focusing on estimation of representation bias (e.g., *internet access at home* in OPBS (Pennay et al. 2018));
- survey items are very sensitive to social desirability, satisficing, are associated with interview administration, and the data as a source of benchmarks were collected in a different mode than the analyzed sample; in this study, potential social desirability bias could be observed in the same benchmarks from different high-quality sources (i.e., potential overreporting of *volunteer work* in the GSS face-to-face surveys in comparison to (predominant) self-completion in the Census);
- survey items are strongly affected by societal or any other changes in times of a crisis (e.g., health including mental health – *Psychological distress* in this study), but benchmark data were collected before the crisis;
- measurement equivalence is challenging to or cannot be achieved in different survey modes (e.g., *moderate exercise*).

None of these issues is associated with actual data quality and data accuracy (or weakly at best). However, they can introduce large measurement bias in studies on representation error or vice versa, representation bias in studies on measurement error, such as measurement mode effect analysis with benchmarks.

#### 11.4.2.3 Sampling error and its effect on benchmarking results

One aspect of total survey error estimation in benchmarking not considered so far in the literature is the sampling effect on the observed total error for a single survey item (i.e., a net effect of sampling

[82] The over-estimation of altruistic activities in the GSS survey might as well be a result of differential nonresponse in sample surveys, although to a lesser extent than in smaller-scale non-government probability surveys, such as the RDD SMS Web-push.

variance). With a simple data simulation exercise in R, how two parameters could have effects on the results will be shown.

Data were simulated using a Monte Carlo method and samples were drawn from the same population using sampling without replacement. Two parameters were randomized: sample size of "smaller-scale surveys" (n=between 500 and 3000), and proportion of respondents answering with the modal category of a survey item (p=between 0.1 and 0.5). The sample size of the "high-quality benchmark survey" was maintained at a constant of n=20.000. The results from Table 11.5 demonstrate that:

- by chance, smaller samples have estimates with a greater average distance to the benchmarks in comparison to larger samples, although they were drawn from the same population (e.g., mean error for p=0.5 and n=500 equals 1.9%, but for p=0.5 and n=3000 only 1%); thus, in benchmarking analysis of samples with high accuracy, surveys with larger samples would be favored;
- by chance, low proportion modal category estimates (e.g., p=0.1) are on average closer to the benchmarks than medium proportion modal category estimates (e.g., p=0.5), although the samples were drawn from the same population; while this is a problem consistent across all the samples compared, survey items with proportions in the modal categories closer to p=0.5 can thus have a greater influence on the total benchmarking measure scores, particularly on RMSE.

Knowing that estimates from different surveys are mostly on the same side of the benchmark (i.e., all over-estimating or under-estimating a particular concept (see Table 11.1)), the effect of sampling variance is not negligible in benchmarking studies. However, it is less of a problem when comparing less accurate samples with a total representation, nonresponse, and measurement bias exceeding sampling variance.

#### 11.4.2.4 Selection of item categories

Considering the same reason that non-random selection of benchmarks can introduce error in benchmarking studies, purposive selection of item categories can introduce bias. To avoid this issue, benchmarking studies selected the modal category for all benchmarks or carried out benchmarking across all categories of variables. However, there is not sufficient evidence to assume that the approaches generally lead to the same findings on accuracy.

Besides, several other issues were identified:

- for ordinal variables with four or more categories, is it sufficient to compare proportions for one (modal) category only?

The evidence from Tables 11.6 and 11.7 suggests that carrying out benchmarking across all categories[83] should be strongly considered as it might change the conclusions. While the differences in the total error and for individual items remain fairly consistent, they decrease in size. Consequently, the differences between the weighted samples are no longer statistically significant. In other words, comparing only the modal category can over-estimate differences in accuracy. As the modal category has the largest p, this is related to sampling variance discussed above.

- for linear numeric scales with only the endpoints labeled, should categories be combined to resemble real-life reporting of results (e.g., aggregating 8–10 on an 11-point linear numeric scale), or should averages be analyzed instead?

  The results from Table 11.7 reveal how comparing the modal category only over-estimates the error for some samples and under-estimates the error for other samples, in comparison to the other approaches (all categories, mean, top 3 categories). Comparing only one category can also be sensitive to satisficing (primacy, recency).

- should more numeric variables be included in benchmarking studies by comparing averages?

  The evidence from Table 11.7 suggests that there is an effective approach to integrate numeric variables into benchmarking studies.

#### 11.4.2.5 Post-survey adjustment techniques and schemes

The results from Table 11.1 show how weighting mostly improves the accuracy of probability samples. However, considering the RDD SMS Web-push sample, raking also introduced adjustment error for the already inaccurate estimates, which is reflected in an increased RMSE score. To test the assumption that certain weighting schemes perform better than some others, which was already discussed in Kennedy et al. (2016), weights were re-calculated by carrying out raking with only three weighting variables (gender, age group, and state).

Comparing the results from Table 11.8, a significant effect of a different post-survey adjustment scheme was not noticed. In contrast to raking with five variables, it appears that raking with three variables did not affect the overall accuracy relative to unweighted data. Among 26 variables, raking with all five primary demographic variables performed slightly better with 15 variables, slightly worse with 10 of them, and the average distance in weighted estimates is only 1.2%-points between the two raking schemes.

[83] Either by reporting the mean difference or standardized mean difference.

## 11.5 Discussion and conclusion

The findings on the accuracy of the RDD SMS Web-push Survey offer valuable insight on the usability of cost-efficient probability-based online survey data collection. With only 1.9% AAPOR 4 response rate (high nonresponse was previously reported in Bucher & Sand 2021), the effects of a digital divide (Couper 2017), excluding the offline population and the majority of people without smartphones, one can expect a more severe total representation error. The sample proved to be slightly more accurate than the most accurate nonprobability panel from OPBS (based on calculations by Kaczmirek et al. 2019) and slightly less accurate than the probability-based OPBS surveys, which is consistent with findings of Couper et al. (2017) and Antoun et al. (2019) on undercoverage and nonresponse bias in mobile/smartphone surveys. However, taking into account cost- and time-efficiency (possible rapid data collection) of the proposed approach, and considering an absence of interviewer effect in comparison to telephone or face-to-face recruitment, this article presents a suitable alternative to a more traditional cross-sectional survey recruitment based on probabilistic principles. It also has the potential to be developed further (e.g., by increasing the response and identifying other topics with lower accuracy).

Moreover, most of the gaps in accuracy between the RDD SMS Web-push sample and OPBS probability samples could be potentially explained with the effect of a social, economic, and health crisis on one particular estimate. Higher levels of self-reported psychological distress were consistent with findings from Biddle and Gray (2021), who saw an increase in average psychological distress in Australia in 2020, which also fluctuated substantially over time. An increase in the average K-6 score was even more substantial among respondents younger than 45 years of age. This age group is generally under-represented in most probability samples (see Chapter 10), including this survey, and thus received larger weights, which resulted in more deteriorated estimates of psychological distress after raking.

The major influence of only one survey item on conclusions shows how important is the design of a benchmarking study. This was later confirmed by two other items that contained a significant error, but except the benchmarks, this study could not compare the estimates to the estimates of other probability and nonprobability surveys in attempts to disentangle coverage and nonresponse bias from measurement bias. Including a bigger sample and a wider range of (randomly selected) benchmarks with different topics could make my findings more robust.

Further, using empirical evidence, this study critically evaluated the approach to benchmarking and commonly used methodological conventions in other benchmarking studies. The investigation on the adequacy of benchmarking methodologies supports the following conclusions:

1. *Benchmarks should be taken from data sources with the highest-quality.*

   This is already an established practice (e.g., Dutwin & Buskirk 2017; Pennay et al. 2018; Yeager et al. 2011). In most countries, the preferred source of socio-demographic benchmark, from a representation error perspective, would be their national census. For non-demographic benchmarks, data from other high-quality sample surveys should be used. The overall accuracy of both census and other high-quality sample survey estimates could be assessed by comparing the matching socio-demographic estimates between different representative data sources.

2. *The more randomness in the selection of benchmarks, the better it is.*

   As Groves et al. (2009) explained, bias is a property of survey statistics and not all survey items are affected by undercoverage in the same manner. A random selection of benchmarks was first proposed by Yeager et al. (2011), but it is not always possible to do this as benchmarking studies are often designed after the data are already collected. If a benchmarking study is a replication of a different benchmarking study (such as Pennay et al. 2018, or this RDD SMS Web-push Survey), benchmarks should not be selected purposely and new benchmarks would ideally be introduced with some level of randomness.

3. *If available, values of benchmarks measuring the same concept should be compared across different data sources.*

   The review and comparative evaluation are even recommended for benchmarks varying over time due to social or other types of changes. In case of a large time gap between data collection periods ("gold standard" survey/census – analyzed survey) and substantial changes over time (for example, due to an exogenous shock), excluding that benchmark should be considered.

4. *Benchmarks sensitive to measurement error (e.g., satisficing and social desirability) affected by a social crisis or strongly associated with the mode of data collection, should be avoided.*

   This is in line with advice from Dutwin and Buskirk (2017) and this issue was identified in my study as well (data on psychological distress were collected during the COVID-19 pandemic). This is a recommendation for benchmarking studies focusing on errors of representation.

5. *It is recommended to carry out benchmarking with all categories of nominal and ordinal survey items, not only modal categories.*

   It is often not convenient to present the results for all categories. In that case, researchers should at least confirm if benchmarking with all categories changed any conclusions, similar to Yeager et al. (2011, p. 10). This study also recommends including more numeric variables in benchmarking studies.

6. *Weighting should be done consistently across different samples.*

   In contrast to the findings of Kennedy et al. (2016), the accuracy of the sample of this study was limitedly affected by different raking schemes. However, not introducing adjustment bias, all samples being compared for accuracy in a benchmarking study should be weighted with a consistent technique and weighting scheme. It is already an established practice to present both weighted and unweighted results.

7. *Measures such as average absolute error (AAE) and root mean squared error (RMSE) should be used.*

   While a number of different measures have been used in the literature, the AAE + RMSE combination captures both the average error and variability or errors.

8. *Ideally, benchmarking studies would be carried out with survey samples of sufficient size, with survey samples of a similar size, and with samples of benchmarks of sufficient size.*

   As shown with data simulation in this study, the effect of sampling variance on benchmarking results is not negligible in fairly accurate samples. Also, a larger sample of benchmarks works similar to a larger sample of survey participants – it helps reduce the effect of outliers with unwanted bias, and it improves precision.

These general recommendations outline a methodological framework for benchmarking. While each imperfect or questionable methodological decision in a benchmarking study might introduce limited bias, it can combine to become a large total bias with a significant effect on conclusions. In the future, other benchmarking solutions should be considered. For example, as data analytics is rarely limited to univariate analysis in practice, bivariate benchmarking analysis should be considered in benchmarking studies.

## 11.6 References

Antoun, C., Conrad, F. G., Couper, M. P., & West, B. T. (2019). Simultaneous estimation of multiple sources of error in a smartphone-based survey. *Journal of Survey Statistics and Methodology*, *7*(1), 93-117.

Australian Bureau of Statistics. (n.d.). *2016 Census QuickStats.* Retrieved January 15, 2021, from https://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/036

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

Biddle, N., & Gray, M. (2021). *Tracking outcomes during the COVID-19 pandemic (January 2021) – Cautious optimism*. ANU Centre for Social Research and Methods.

Bucher, H., & Sand, M. (2021). Exploring the Feasibility of Recruiting Respondents and Collecting Web Data Via Smartphone: A Case Study of Text-To-Web Recruitment for a General Population Survey in Germany. *Journal of Survey Statistics and Methodology (2021) 00, 1–12.*

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

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641-678.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and psychological measurement*, *60*(6), 821-836.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, *64*(4), 464-494.

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology, 43*, 121-145.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. In P. P. Biemer, E. D. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 133-154). John Wiley & Sons.

Daikeler, J., Bosnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513-539.

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

Fricker, R. D. (2008). Sampling methods for web and e-mail surveys. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 195-216). Sage.

Fuchs, M., & Busse, B. (2009). The coverage bias of mobile web surveys across European countries. *International Journal of Internet Science*, *4*(1), 21-33.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167-189.

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

Hsia, J., Zhao, G., & Town, M. (2020). Estimating Undercoverage Bias of Internet Users. *Preventing Chronic Disease 2020*, 17:200026. http://dx.doi.org/10.5888/pcd17.200026

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

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

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

Kennedy, C., McGeeney, K., Keeter, S., Patten, E., Perrin, A., Lee, A., & Best, J. (2018). Implications of moving public opinion surveys to a single-frame cell-phone random-digit-dial design. *Public Opinion Quarterly*, *82*(2), 279-299.

Kongsgard, H. W., Syversen, T., & Krokstad, S. (2014). SMS phone surveys and mass-messaging: promises and pitfalls. *Epidemiology: Open Access*, *4*(4). http://dx.doi.org/10.4172/2161-1165.1000177

Loosveldt, G., & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, *2*(2), 93-105.

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

Malhotra, N., & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. *Political Analysis*, *15*(3), 286-323.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79-104.

Mavletova, A., & Couper, M. P. (2016). Device use in web surveys: The effect of differential incentives. *International Journal of Market Research, 58*(4), 523-544.

McGeeney, K., & Kennedy, C. (2015). *Advances in Telephone Survey Sampling: Balancing efficiency and coverage using several new approaches.* Pew Research Center.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For weighting online opt-in samples, what matters most.* Pew Research Center.

Merkle, D. M. (2008). Nonresponse bias. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 531-533). Sage.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*(201), 101-116.

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

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

Peterson, G., Griffin, J., LaFrance, J., & Li, J. (2017). Smartphone participation in web surveys. In P. P. Biemer, E. D. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 203-233). John Wiley & Sons.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roy Morgan. (2018, November). *Kids and Mobiles, How Australian children are using mobile phones.* https://app.powerbi.com/view?r=eyJrIjoiNzA4Njc4YWMtZDk5OC00M2EwLWI3MDktMjkzMGRmOGNhOThkIiwidCI6IjBkYWM3ZjM5LWQyMGMtNGU3MS04YWYzLTcxZWU3ZTI2OGEyYiJ9&pageName=ReportSectionf48b9bb17d51b38a13e1

SamplePages. (n.d.). *What we do*. Retrieved November 15, 2020, from https://samplepages.com.au/#service

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The end of the (research) world as we know it? Understanding and coping with declining response rates to mail surveys. *Society & Natural Resources*, *32*(10), 1139-1154.

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

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

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

## Appendix 11

Online Panel Benchmarking study data

The Online Panels Benchmarking Study (OPBS) data are not the focus of this study, but are used as reference samples as the accuracy of different probability and nonprobability surveys, as part of this study, have already been thoroughly investigated (see Pennay et al. 2018; Kaczmirek et al. 2019). Instead, the OPBS data are used to make comparisons in accuracy between the RDD SMS Web-push Survey data and four OPBS samples, relative to the nationally representative benchmarks. The four OPBS surveys are as follows: RDD Standalone, RDD end-of-survey recruitment, and Address-Based Sampling survey as the probability surveys, and Panel 2 data as the most accurate nonprobability survey data from OPBS (Pennay et al. 2018). The online panel 2 sample will work as a nonprobability reference to the RDD SMS Web-push sample used in this study. All OPBS surveys used the same *Health, Wellbeing and Technology Questionnaire* with a total of 19 secondary demographics and non-demographic benchmarks.

Table 11.2: Online Panels Benchmarking Study (OPBS 2015) samples used as reference samples

| Survey sample | Sampling | Data collection mode(s) | Sample size |
|---|---|---|---|
| Address-based sampling (A-BS) sample | A-BS using Geocoded National Address File (GNAF) | online, telephone, mail | 538 |
| End of telephone survey "piggybacking" sample | Dual frame RDD sampling | online, telephone, mail | 560 |
| Standalone telephone sample | Dual frame RDD sampling | telephone | 601 |
| Opt-in online panel 2 | Nonprobability convenience sampling | online | 600 |

Table 11.3: Government funded nationally representative data sources of benchmarks

| Study | Sampling | Data collection mode(s) | Sample size |
|---|---|---|---|
| National Health Survey 2014-15 | a stratified multistage area sample of private dwellings | F2F | n=19,259 persons |
| National Health Survey 2017-18 | a stratified multistage area sample of private dwellings | F2F | n=21,315 persons |
| General Social Survey 2014 | Stratified sampling (oversampling in low socio-economic areas) | F2F | n=12,932 dwellings |
| General Social Survey 2019 | Stratified sampling (oversampling in low socio-economic areas) | F2F and online self-completion | n=3,535 households |
| Household, Income and Labour Dynamics in Australia (HILDA) Survey (Release 18) | Multi-stage stratified random sample of households (a panel study with top-up samples) | F2F (telephone interviews as the last resort) | Between n=23,237 (Wave 18) and n=23,299 (Wave 13) persons |
| National Drug Strategy Household Survey 2013 | A multi-stage stratified random sample design | self-administered paper based | n=23,855 persons |
| National Drug Strategy Household Survey 2016 | A multi-stage stratified random sample design | self-administered paper-based or online, telephone interviews | n=23,749 persons |
| National Drug Strategy Household Survey 2019 | A multi-stage stratified random sample design | self-administered paper-based or online, telephone interviews | n=22,013 persons |
| Australian Census 2011 | total coverage/census | self-administered online, F2F | N=21,507,717 people, N= 9,117,033 dwellings |
| Australian Census 2016 | total coverage/census | self-administered online, F2F | N=23,401,892 people, N=9,901,496 private dwellings |

Table 11.4: Data sources of benchmarks with benchmark values (1/2)

| No | Survey item with available benchmark(s) | Included in OPBS | Data source 1 | Data source 2 | Data source 3 | Data source 4 | Data source 1 benchmark value | Data source 2 benchmark value | Data source 3 benchmark value | Data source 4 benchmark value |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Australian citizen | Yes | Australian Census 2011* | Australian Census 2016 | | | 83.9 | 87.1 | | |
| 2 | Couple with dependent children | Yes | National Drug Strategy Household Survey 2013* | National Drug Strategy Household Survey 2016 | National Drug Strategy Household Survey 2019 | | 38.4 | 30.3 | 28.4 | |
| 3 | Currently employed | Yes | Australian Census 2011* | Australian Census 2016 | | | 59.4 | 61.6 | | |
| 4 | Home ownership with a mortgage | Yes | Australian Census 2011* | Australian Census 2016 | | | 29.6 | 28.8 | | |
| 5 | Not Indigenous | Yes | Australian Census 2011* | Australian Census 2016 | | | 98.1 | 97.7 | | |
| 6 | Language other than English (*Speak only English*) | Yes | Australian Census 2011* | Australian Census 2016 | | | 75.7 | 76.5 | | |
| 7 | Living at last address 5 years ago | Yes | Australian Census 2011* | Australian Census 2016 | | | 54.8 | 56.9 | | |
| 8 | Voluntary work (*None*) | Yes | Australian Census 2011* | Australian Census 2016 | General Social Survey 2014 | General Social Survey 2019 | 74.2 | 79.4 | 69.1 | 71.2 |
| 9 | Carer status (*Yes*) | No | Australian Census 2011 | Australian Census 2016 | | | 12.7 | 12.3 | | |
| 10 | Wage and salary income (*$1000–1249 per week*) | Yes | National Health Survey, 2014–15* | National Health Survey, 2017–18 | | | 13.8 | 14.6 | | |
| 11 | Consumed alcohol in last year | Yes | National Drug Strategy Household Survey 2013* | National Drug Strategy Household Survey 2016 | National Drug Strategy Household Survey 2019 | | 81.9 | 80.6 | 81.0 | |
| 12 | Daily smoker | Yes | National Drug Strategy Household Survey 2013* | National Drug Strategy Household Survey 2016 | National Drug Strategy Household Survey 2019 | | 13.5 | 13.1 | 11.6 | |
| 13 | General health status (*Very good*) | Yes | National Health Survey 2014-15* | National Health Survey 2017-18 | National Drug Strategy Household Survey 2016 | National Drug Strategy Household Survey 2019 | 36.2 | 35.5 | 38.1 | 38.6 |
| 14 | Psychological distress, Kessler 6 - Hopeless (*Never*) | Yes | National Health Survey 2014-15 | National Health Survey 2017-18 | | | 80.0 | 77.7 | | |

Table 11.4: Data sources of benchmarks with benchmark values (2/2)

| No | Survey item with available benchmark(s) | Included in OPBS | Data source 1 | Data source 2 | Data source 3 | Data source 4 | Data source 1 benchmark value | Data source 2 benchmark value | Data source 3 benchmark value | Data source 4 benchmark value |
|---|---|---|---|---|---|---|---|---|---|---|
| 15 | Moderate exercise in the last week (*Yes*) | No | National Health Survey 2014-15 | National Health Survey 2017-18 | | | 29.4 | 30.0 | | |
| 16 | Number of serves of vegetables each day (*2 serves*) | No | National Health Survey 2014-15 | National Health Survey 2017-18 | | | 28.2 | 28.2 | | |
| 17 | Time stress (*Always/often*) | No | General Social Survey 2014 | General Social Survey 2019 | | | 40.9 | 39.6 | | |
| 18 | Satisfaction with: Life as a whole (*8 out of 10*) | Yes | HILDA Wave 15 (2015) | HILDA Wave 18 (2018) | General Social Survey 2014* | | 33.0 | 33.0 | 32.6 | |
| 19 | Satisfaction with: Health (*8 out of 10*) | No | HILDA Wave 15 (2015) | HILDA Wave 18 (2018) | | | 27.0 | 27.5 | | |
| 20 | Satisfaction with: Financial situation (*8 out of 10*) | No | HILDA Wave 15 (2015) | HILDA Wave 18 (2018) | | | 21.8 | 21.8 | | |
| 21 | Satisfaction with: Safety (*8 out of 10*) | No | HILDA Wave 15 (2015) | HILDA Wave 18 (2018) | | | 28.6 | 28.6 | | |
| 22 | Big 5 personality traits, Agreeableness - Warm (*5 out of 7***) | No | HILDA Wave 13 (2013) | HILDA Wave 17 (2017) | | | 30.5 | 30.3 | | |
| 23 | Big 5 personality traits, Conscientiousness - Orderly (*5 out of 7***) | No | HILDA Wave 13 (2013) | HILDA Wave 17 (2017) | | | 25.0 | 24.6 | | |
| 24 | Big 5 personality traits, Emotional stability - Moody (*2 out of 7***) | No | HILDA Wave 13 (2013) | HILDA Wave 17 (2017) | | | 23.6 | 24.5 | | |
| 25 | Big 5 personality traits, Extroversion - Quiet (*4 out of 7***) | No | HILDA Wave 13 (2013) | HILDA Wave 17 (2017) | | | 22.6 | 22.0 | | |
| 26 | Big 5 personality traits, Openness - Philosophical (*4 out of 7***) | No | HILDA Wave 13 (2013) | HILDA Wave 17 (2017) | | | 22.9 | 21.3 | | |

*Benchmarks taken from Pennay et al. (2018);

**Scale: 1 – Does not describe me at all, 7 – Describes me very well

Table 11.5: Mean estimated sampling variance, absolute distance benchmark-estimate, benchmark from a 'high-quality survey', data simulation (n=20,000)

| | Sample size of a smaller-scale survey | | |
|---|---|---|---|
| **Proportion of modal category responses** | **n=500** | **n=1500** | **n=3000** |
| | **mean [95% CI]** | **mean [95% CI]** | **mean [95% CI]** |
| **p=0.10** | 0.0104 [0.0004,0.0304] | 0.0062 [0.0002,0.0166] | 0.0046 [0.0002,0.0133] |
| **p=0.25** | 0.0163 [0.0008,0.0463] | 0.0095 [0.0003,0.0275] | 0.0068 [0.0003,0.0198] |
| **p=0.50** | 0.0189 [0.009,0.0553] | 0.0107 [0.0003,0.0310] | 0.0078 [0.0003,0.0228] |

Table 11.6: Benchmarking analysis with all categories, not only modal

| **Survey item with available benchmark(s)** | **RDD SMS Web-push Survey (2020)** | **Online Panels Benchmarking Study (OPBS) (2015)** ***(analysis of the data file (Pennay et al. 2016))*** | | | |
|---|---|---|---|---|---|
| | | **RDD** | **A-BS** | **ANUpoll** | **Opt-in panel 2** |
| | **Mean error**** | **Mean error**** | **Mean error**** | **Mean error**** | **Mean error**** |
| Australian citizen | 2.1 | 2.7 | 8.1 | 2.7 | 4.2 |
| Household status* | 2.6 | 3.4 | 3.7 | 3.6 | 3.9 |
| Employment status | 0.1 | 9.9 | 5.2 | 7.0 | 6.1 |
| Home ownership* | 2.6 | 3.1 | 2.7 | 2.5 | 2.6 |
| Indigenous status | 0.1 | 0.7 | 0.3 | 0.3 | 1.6 |
| Language other than English | 0.0 | 9.8 | 4.7 | 8.8 | 9.7 |
| Living at last address 5 years ago | 8.9 | 7.3 | 0.1 | 3.6 | 3.3 |
| Voluntary work | 16.6 | 11.6 | 11.2 | 11.6 | 2.9 |
| Wage and salary income* | 2.8 | 2.8 | 1.8 | 1.0 | 3.3 |
| Consumed alcohol in last year | 1.7 | 4.0 | 3.6 | 2.8 | 5.3 |
| Smoking habits* | 2.2 | 0.8 | 2.5 | 3.6 | 4.3 |
| General health status* | 7.7 | 3.8 | 2.1 | 3.2 | 7.5 |
| Psychological distress, Kessler 6 - Hopeless* | 15.6 | 4.2 | 6.2 | 6.0 | 11.9 |
| Satisfaction with: Life as a whole* | 5.3 | 1.1 | 2.7 | 2.8 | 5.5 |
| **AAE (OPBS items)** | 4.88 | 4.65 | 3.92 | 4.26 | 5.15 |
| **RMSE (OPBS items)** | 7.17 | 5.78 | 4.88 | 5.19 | 5.84 |

*Categorical variables with 3+ categories

**Errors were calculated for each category of categorical variables with 3+ categories separately and then averaged into a mean

Table 11.7: Comparison of benchmarking results for a numeric scale item with different descriptive analysis approaches

| Item | Measure/ category | RDD SMS Web-push | | OPBS | | | | | RDD SMS Web-push | RDD | A-BS | ANUpoll | Panel 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Benchmark | W | Benchmark | RDD W | A-BS W | ANUpoll W | Panel 2 W | Error | Error | Error | Error | Error |
| Satisfaction with life as a whole | Mean | 7.90 | 6.96 | 7.89 | 7.61 | 7.35 | 7.27 | 6.61 | 9.4* | 2.8* | 5.4* | 6.2* | 12.8* |
| | 8 | 33.0 | 23.3 | 34.4 | 34.5 | 30.6 | 30.6 | 21.0 | 9.7 | 0.1 | 3.9 | 3.8 | 13.4 |
| | 8-10 | 67.5 | 51.1 | 67.9 | 62.3 | 54.7 | 55.4 | 36.3 | 16.4 | 5.6 | 13.2 | 12.5 | 31.6 |
| | 0 | 0.1 | 2.5 | 0.1 | 0.3 | 0.5 | 1.1 | 0.0 | 5.3 | 1.1 | 2.7 | 2.8 | 5.5 |
| | 1 | 0.2 | 1.2 | 0.2 | 0.7 | 0.5 | 0.2 | 1.5 | | | | | |
| | 2 | 0.4 | 2.8 | 0.3 | 1.1 | 0.6 | 1.6 | 3.2 | | | | | |
| | 3 | 0.7 | 3.1 | 0.8 | 1.1 | 1.2 | 1.8 | 3.8 | | | | | |
| | 4 | 1.3 | 3.4 | 1.3 | 1.2 | 2.5 | 2.5 | 4.5 | | | | | |
| | 5 | 3.6 | 13.1 | 3.8 | 7.2 | 9.8 | 12.0 | 15.7 | | | | | |
| | 6 | 6.3 | 11.1 | 5.9 | 6.6 | 11.7 | 8.2 | 10.5 | | | | | |
| | 7 | 20.0 | 11.8 | 19.7 | 19.4 | 17.7 | 16.9 | 21.2 | | | | | |
| | 8 | 33.0 | 23.3 | 34.4 | 34.5 | 30.6 | 30.6 | 21.0 | | | | | |
| | 9 | 22.6 | 11.6 | 21.7 | 15.8 | 14.6 | 13.9 | 9.5 | | | | | |
| | 10 | 11.8 | 16.2 | 11.7 | 12.0 | 9.5 | 10.9 | 5.7 | | | | | |

*relative error calculated as the distance between the means divided by the maximum possible distance, i.e., 10

Table 11.8: Comparing accuracy in RDD SMS Web-push data after two different post-survey adjustment schemes

| Survey item with available benchmark(s) | RDD SMS Web-push Survey (2020) | | | |
|---|---|---|---|---|
| | Benchmark | Survey estimates | | |
| | | UW | Full raking[a] | Partial raking[b] |
| Australian citizen | 87.1 | 92.2 | 85.0 | 90.1 |
| Couple with dependent children | 28.4 | 16.6 | 22.4 | 21.3 |
| Currently employed | 61.6 | 60.3 | 61.7 | 63.2 |
| Home ownership with a mortgage | 28.8 | 29.7 | 29.3 | 30.8 |
| Not Indigenous | 97.7 | 97.9 | 97.6 | 97.6 |
| Language other than English (*Speak only English*) | 76.5 | 81.7 | 76.5 | 78.9 |
| Living at last address 5 years ago | 56.9 | 58.8 | 48.0 | 49.3 |
| Voluntary work (N*one*) | 79.4 | 61.3 | 62.8 | 62.0 |
| Wage and salary income (*$1000–1249 per week*) | 14.6 | 11.8 | 15.5 | 13.5 |
| Consumed alcohol in last year | 81 | 83.1 | 82.7 | 83.4 |
| Daily smoker | 11.6 | 11.6 | 13.3 | 12.0 |
| General health status (*Very good*) | 35.5 | 28.9 | 29.5 | 30.1 |
| Psychological distress, Kessler 6 - Hopeless (*Never*) | 77.7 | 43.9 | 38.6 | 41.4 |
| Satisfaction with: Life as a whole (*8 out of 10*) | 33.0 | 24.3 | 23.3 | 23.7 |
| Satisfaction with: Health (*8 out of 10*) | 27.5 | 18.0 | 19.0 | 19.2 |
| Satisfaction with: Financial situation (*8 out of 10*) | 21.8 | 19.3 | 15.0 | 18.3 |
| Satisfaction with: Safety (*8 out of 10*) | 28.6 | 18.4 | 18.5 | 18.3 |
| Carer status (*Yes*) | 12.3 | 35.9 | 33.1 | 33.5 |
| Moderate exercise in the last week (*Yes*) | 30.0 | 71.3 | 70.1 | 71.4 |
| Number of serves of vegetables each day (*2 serves*) | 28.2 | 25.5 | 29.1 | 28.3 |
| Time stress (*Always/often*) | 39.6 | 38.4 | 40.4 | 40.9 |
| Big 5 personality traits, Agreeableness - Warm (*5 out of 7**) | 30.3 | 31.4 | 31.0 | 32.2 |
| Big 5 pers. traits, Conscientiousness - Orderly (*5 out of 7**) | 24.6 | 26.9 | 26.0 | 27.3 |
| Big 5 personality traits, Emotional stability - Moody (*2 out of 7**) | 24.5 | 23.3 | 22.5 | 22.9 |
| Big 5 personality traits, Extroversion - Quiet (*4 out of 7**) | 22.0 | 20.1 | 18.3 | 18.6 |
| Big 5 personality traits, Openness - Philosophical (*4 out of 7**) | 21.3 | 19.9 | 22.0 | 21.1 |
| **AAE (OPBS items)** | | 7.04 | 6.67 | 6.86 |
| **RMSE (OPBS items)** | | 11.34 | 12.13 | 11.57 |
| **AAE (all items)** | | 7.59 | 7.30 | 7.38 |
| **RMSE (all items)** | | 12.84 | 12.93 | 12.79 |

[a] raking by gender, age group*education (degree, no degree), state, country of birth

[b] raking by gender, age group, state

# Chapter 12 Discussion

In this last chapter, I present and generalize the most important findings related to dealing with survey errors in online panels and discuss practical implications identified in this thesis for online panel research. This includes presenting the theoretical contributions and the limitations of the research, discussing the cost dimension of online panel recruitment and data collection, contesting the requirement to collect the data from the offline population by mixing modes, and presenting the overview of the state-of-the-art of online panel research at the beginning of the 2020s. In these last paragraphs, I expand the academic focus of this methodological research to commercial social and market research perspectives.

## 12.1 Theoretical contributions and the limitations of this research

Besides contributions for the practice of online panel survey research, there are a number of theoretical contributions of this thesis. Many of these are directly related to applying and extending theories in survey methodology to online panel and longitudinal research. First, in Chapters 3 and 10 we provide additional comprehensive evidence on survey response maximization approaches to the 'attributes of the survey design' dimension of the theory of survey participation. The meta-analysis of recruitment rates is the first comprehensive study to generalize evidence on recruitment outcome maximization strategies to probability-based online panel research worldwide. Also, the study on response maximization in RDD SMS web-push data collection is the first of its kind to fully apply a theory of survey participation on this new approach to probability-based online data collection. Second, this thesis is making contributions to theory of the response process by introducing a new source of measurement error – panel measurement mode effects. This type of mode effects is specific to mixed-mode longitudinal research and can introduce error only if panellists are able to and/or are encouraged to respond in different survey modes. In the same study presented in Chapter 8, panel conditioning has been for the first time conceptualized as a factor of the frequency of repeated measurements and the time gap between measurements, which is also quite specific to the longitudinal approach in online panel research. Third, this thesis extends social-psychological theories of survey participation and leverage-salience theory to a longitudinal (online panel) context. We show how these theoretical and conceptual theories can explain outcomes in different stages of a panel lifecycle and not just in recruitment to a study, which is the only stage in a cross-sectional survey design. Additionally, we showed how social-psychological theories and leverage-salience can explain changes in panel response behavior over time.

There are also notable limitations to the research presented in this thesis and many of them are more or less related to the availability of relevant data to study survey errors in online panel research. First,

since eight out of nine research papers (Chapters 4–11) are based on the analysis of Australian data and there is only one national probability-based online panel, findings are somewhat limited to the Australian context and Life in Australia™. Also, a RDD SMS web-push approach to data collection could not be carried out in countries requiring respondents' consent to be texted an SMS invitation to an online survey.

Second, the amount of methodological information provided by organizations managing probability-based online panels exceeds the information that commercially-oriented volunteer panel providers are willing to release publicly. In the meta-analysis, we show how even probability-based panels that are more commercial in nature were not willing to share with us much methodological details on their recruitment. In order to make our findings more robust and to study the previously discussed relationship between data quality and data collection costs, we would require access to more data and more information about commercially-oriented online panels. This could be beneficial for certain online panel providers as well, as subsequent research could provide evidence required to optimize recruitment and data collection costs.

Lastly, while this thesis addresses various research gaps and answers several research questions, it also generates a number of new research questions and methodological solutions to be tested in practice. The evidence from the meta-analysis could provide a good basis for a research organization to build a probability-based online panel in a new country, but their recruitment maximization strategies would have to be tested in their particular context due to unobserved country specifics. While we identified an approach for a fairly accurate identification of nonrespondents in a subsequent online panel wave/survey, further research is required on how to treat potential nonrespondents to reduce nonresponse in practice. Moreover, some of the findings should be tested in online panel research settings. For example, while I show how representation bias in RDD SMS web-push survey does not represent a notable issue, there has not been extensive study of the same bias if a similar approach is used in recruitment to a probability-based online panel. All these offer new opportunities for methodological research on online panels in the future.

## 12.2 Online panel research and the cost dimension

Out of three things that are valued in the survey research profession above all, this thesis mostly discussed the quality aspect and to a lesser extent the speed aspect. However, there is the third aspect which plays a key role in commercial, polling, government and academic research sectors – affordability. The cost dimension should evidently be considered in practice, as survey methodology generally focuses on improving quality without affecting costs, or reducing costs without affecting quality (Groves 2004). With declining response rates resulting in increasing costs of data collection,

the challenge for survey methodology space is to mitigate inferential risks and to do more with less (Couper 2017). In the era of limited budgets, organizations have to take into consideration how important inference to the population is, what they would like to estimate, and the cost/quality trade-off. Besides censuses, large-scale probability-based surveys are considered the most accurate in measuring social phenomena in the population, but they require large budgets, especially if data are collected face-to-face (Couper 2017). At the other extreme, nonprobability panels can provide data for low costs, but at the expense of quality. Probability-based panels are somewhere in between, from both cost and quality perspectives. However, if a volunteer panel (see Chapter 9) or an RDD SMS web-push survey (see Chapter 11) produce estimates with only slightly larger absolute error, it is not always possible to justify spending at least five times more for a marginally more accurate probability-based online panel sample[84]. Some of the findings presented in this thesis point in that direction, but there is still limited evidence on the relationship between data collection costs and data quality. From a scientific perspective, making data collection/research cost data available in addition to data on sample accuracy would represent added value in the study of survey errors. Synthesizing evidence on the cost/accuracy trade-off would certainly be of scientific interest, and future research should further explore the relationship between those two important dimensions in more detail. This general recommendation should not be limited to online panel research.

Furthermore, the trends in probability-based online panel research indicate that organizations managing those panels nowadays tend to choose more affordable strategies to recruitment. Those are the previously discussed shift from telephone to postal recruitment, end-of-survey approach to recruitment to panels (e.g., GESIS panel or panels from the UK), and a noncoverage of the offline population. The analysis in Chapter 3 shows how recruitment rates decreased slightly over time, but meta-regression models explain that trend with methodological changes, some of which are associated with more cost-efficient recruitment solutions. I could speculate that the recruitment-expenses saving efforts are a result of increasing costs of survey data collection in general, and/or a harsh competition against cheaper nonprobability-based panels.

The evidence from this thesis supports a few other potential solutions to decrease costs of data collection in probability-based online panel research. First, we show how the guaranteed recruitment incentives amount did not have an effect on recruitment rates, which means that a 'symbolic' amount in the form of a mixture of unconditional and conditional incentives could result in as efficient recruitment as more expensive recruitment incentives schemes. This evidence was later supported in

[84] For example, in the RDD SMS web-push study (see Chapter 10), I collected responses for about 5 AUD per participant (with potential to decrease costs based on the results of a survey experiment); Garrow et al. (2020) collected responses for about 0.5 USD (Mechanical Turk) and 7.5 USD (Qualtrics) per respondent; both studies excluded personnel time in the calculation of costs.

Chapter 4, as only about one in eight Life in Australia™ panellists mentioned *receiving incentives* or *donating to charity* as motivational factors for completing panel surveys. Second, the evidence from Chapters 10 and 11 suggests that RDD SMS recruitment is not only a fairly accurate online survey data collection approach, but it could potentially be used as a cost-effective mode to recruitment to probability-based or mixed-sampling online panels. Lastly, we presented evidence on how different probability-based online panels do not use the offline mode or provide the required technology to the offline population to respond online, which can be seen as a cost-saving measure. The survey error implications of that are discussed in more detail in the next paragraphs.

## 12.3 Mixing-modes in online panel research and its effect on survey errors

As covering the offline population by mixing modes in probability-based online panel research normally increases the costs of data collection, there has to be valid data quality-related reasons (including the perception of quality by clients) for either using the offline mode or offering technology to potential panellists. In Life in Australia™, the panel largely investigated in this study, the telephone mode is used to collect survey data from those who cannot or do not want to respond online and as a reminder/data collection mode for those respondents who initially do not respond to email invitations. As such, the telephone mode works as a part of the strategy to control for coverage and nonresponse error.

In this thesis, I studied the mixing-modes total survey error phenomena from two perspectives, undercoverage and measurement mode effects. In Chapter 6, we identified small undercoverage bias if excluding the offline population, which showed in almost negligible effect on the quality of most survey estimates; similar results were previously reported for the LISS panel (Eckman 2016). In Chapter 7, we identified measurement mode effects in mixed-method and online panel research, namely a lack of measurement equivalence in the telephone mode and satisficing in the mail mode, and concluded that it was difficult to study mode effects in studies lacking random assignment to modes; similar findings were previously reported by Dennis et al. (2005). In Chapter 8, we extended the measurement mode effect analysis to longitudinal online panel data, and showed how the same respondents switching modes can lead to an increased measurement error in a panel setting.

Finding the balance between these sources of survey error is a common problem in mixed-mode survey research – while using both interviewer-administered and self-administered survey modes can improve representation and response, it can also introduce measurement error due to questions being presented to respondents differently. Combining the key findings from Chapters 6, 7 and 8, I can conclude that we found as much evidence on negative contribution of mixing modes on total survey error as evidence on the improved quality of probability-based online panel data if covering the offline

population. Taking into account higher data collection costs, mostly due to additional telephone interviewing expenses (or printing and postal expenses), all things considered, we did not find convincing evidence that covering the offline population is necessary in countries with similar internet penetration and 'web-demographic' population structure to Australia. It could be argued that money spent on covering the offline population should be spent on panel management to mitigate other sources of survey error. However, there are exceptions to this general recommendation, as online-only panel surveys cannot accurately estimate concepts strongly associated with offline population membership.

## 12.4 The state-of-the-art in online panel research

As online panel research is younger than survey research using more traditional modes, literature on online panels and survey errors is to some extent limited. In 2010, an Opt-In Online Panel Task Force was established by the AAPOR Executive Council to review the empirical findings on nonprobability based online panels. The report written by Baker and his colleagues (2010) represented the most in-depth review of online panel research to that date, although it was predominantly focused on opt-in panels and somewhat limited to the US online research market. In 2014, Callegaro and his colleagues (2014) published a book on online panel research and data quality. In contrast to the report from Baker et al. (2010), it focused on both probability and nonprobability-based panels in chapters covering different aspects of and survey errors in online panel research. It can be argued that more research on survey errors is required in online panels, including an investigation of several concurrent panel-specific errors, to further develop the TSE framework (which would be even more suitable for studying the total survey error in online panels). While more methodological research has been published and presented in recent years, it was much more focused on individual aspects, and has mostly come in the form of journal articles and conference presentations.

Thus, this thesis and its chapters can be potentially considered as the most comprehensive recent overview of the state-of-the-art of online panel research, and Chapter 3 arguably contributes the most to this overview. The meta-analytical study primarily focuses on studying factors affecting recruitment outcomes in probability-based online panels, as well as identifying 28 panels and documenting various recruitment and data collection strategies which have evolved over the years. While we could not collect information on establishment year for every single probability-based online panel, the research presented here estimates that more than one-half of all active panels were established in the 'golden years' for online panels, that is, between 2012 and 2016. There have also been some notable differences in recruitment approaches and methodological changes over time. For example, the mail/postal recruitment mode became more popular, and that was at the expense of the telephone

recruitment mode. The research also identified more end-of-survey/piggybacking recruitment after 2015. In the future and due to cost efficiency, I would not be surprised if there were more RDD-mobile-SMS and IVR recruitment approaches, which were already tested by the Social Research Centre in 2020 (Phillips et al. 2021). Recently, there are more examples of online panels either not collecting data from the offline population (for example, Norwegian Citizen Panel), not recruiting offliners in the most recent recruitment waves (for example, Life in Australia™), or naturally decreasing the proportion of the offline population in the panel (for example, the American Trends Panel). This has been possible with an increasing internet penetration and, consequently, less concern for undercoverage bias. From this perspective alone, many probability-based online panels have become more similar to volunteer online panels.

Those nonprobability-based online panels are still considered less accurate due to self-selection driven representation bias, high nonresponse, professional respondents and bot participants. However, there seem to be notable within-group differences in quality between nonprobability online panels as well, which are conditional on the accreditation and the country of online panel. For example, International Organization for Standardization (ISO) or European Society for Opinion and Marketing Research (ESOMAR) accredited panels should be considered as more accurate than those without accreditation, such as Qualtrics or the extremely cost-effective Mechanical Turk. Also, based on the results from the benchmarking studies from Pennay et al. (2018), Kaczmirek et al. (2019), MacInnis et al. (2018) and other US studies, the gap in quality between volunteer online panels and probability (online) surveys seems to be much smaller in Australia than in the US. This might be a result of the differences in population structures (e.g., greater differences between online and offline populations, or between volunteer panellists and the general population in the US) or the relative quality of the volunteer panels compared to probability samples (volunteer panel selection effect). Also, a new type of volunteer panels has appeared in recent years, that is, volunteer academic panels. With respected survey research sponsors/authority (i.e., universities), those panels should have a better potential for successful recruitment, would not need to incentivize panellists for completing questionnaires, and would attract intrinsically motivated respondents instead of professional respondents. As new methodologies for improving the quality/accuracy of nonprobability online samples are being developed, including in our study which advances the field of post-survey adjustments, there is a potential for more accredited volunteer online panels to be used for data collection outside the market research space, and academic panels to be used in academic and government survey research. On the other hand, we showed how probability-based online panels are still more accurate and they should continue to be used in survey research leaning on a more rigorous methodology. Taking into account various opportunities for mitigation of sources of survey errors and cost-saving solutions

identified in this thesis, there is a strong argument that both types of online panels will play an even more important role in survey research in the future.

## 12.5 References


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

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

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology, 43*, 121-145.

Dennis, J. M., 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. *60th Annual Conference of the American Association for Public Opinion Research*, 1-26.

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

Garrow, L. A., Chen, Z., Ilbeigi, M., & Lurkin, V. (2020). A new twist on the gig economy: conducting surveys on Amazon Mechanical Turk. *Transportation, 47*(1), 23-42.

Groves, R. M. (2004). *Survey errors and survey costs* (Vol. 536). John Wiley & Sons.

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

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

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

Phillips, B., Dove, C., Myers, P., & Neiger, D. (2021, March 4-5). *Expansion of an Australian probability-based online panel using ABS, IVR and SMS push-to-web* [Conference presentation]. CIPHER 2021 Conference, Virtual.