mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 17, 2023 | 1

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession‘s knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.
All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)
Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

methods, data, analyses | Vol. 17(1), 2023, pp. 3-28 DOI: 10.12758/mda.2022.05

River Sampling – a Fishing Expedition: A Non-Probability Case Study

Alexander Murray-Watters[1], *Stefan Zins*[2], *Henning Silber*[3], *Tobias Gummer*[3] & *Clemens M. Lechner*[3]

[1] *Department of Sociology, University of California-Irvine*

[2] *Institute for Employment Research*

[3] *GESIS – Leibniz Institute for the Social Sciences*

Abstract

The ease with which large amounts of data can be collected via the Internet has led to a renewed interest in the use of non-probability samples. To that end, this paper performs a case study, comparing two non-probability datasets – one based on a river-sampling approach, one drawn from an online-access panel – to a *reference* probability sample. Of particular interest is the single-question river-sampling approach, as the data collected for this study presents an attempt to field a multi-item scale with such a sampling method. Each dataset consists of the same psychometric measures for two of the *Big-5* personality traits, which are expected to perform independently of sample composition. To assess the similarity of the three datasets we compare their correlation matrices, apply linear and non-linear dimension reduction techniques, and analyze the distance between the datasets. Our results show that there are important limitations when implementing a multi-item scale via a single-question river sample. We find that, while the correlation between our data sets is similar, the samples are composed of persons with different personality traits.

Keywords: River Sample, Non-probability Sample, BIG-5, Non-linear Dimension reduction, Web Survey Research

Traditional survey methods are under pressure from emerging techniques for conducting web surveys (Baker et al., 2016; Couper, 2011; Miller, 2017). Declining response rates and the increasing cost of traditional surveys encourage practitioners to pursue alternative tactics – such as online surveys. Regrettably, rigorous methodology for online surveys has lagged behind their use in industry. This paper attempts to address some of this methodological lag, by assessing if a widely used psychological measure produces consistent results when it is collected via a novel non-probability sample – a single question river sample. This is of particular interest to psychologists as collecting psychometric data via traditional surveys (e.g., face-to-face or telephone) can be prohibitively expensive.

In order to remedy the lack of research on the applicability of river sampling surveys for scientific research, we conducted a study where we compared data collected through a river sampling single-question approach to data collected in probability and non-probability based panels. River sampling is a self-selected non-probability survey technique, while a "single question approach" involves the invitation to independent follow-up surveys one at a time, in no particular sequence. For a multi-item construct, we used two domains (Conscientiousness and Emotional Stability) from a Big-5 inventory that were fielded in each of the three surveys. For ease of reference, we now refer to the single-question river sample approach simply as a river sample.

The article is structured as follows: In the Background Section, we briefly summarize the existing literature on web-based surveys and some of the new uses of online river sampling. In the Data Section, we describe the Big-5 inventory that we used and the essential properties of the three different samples that we study. In the Methods Section, we describe the different analytical tools that were used to compare the different data sets with each other. The Results Section contains descriptive statistics on the river sample and the results from our comparisons. The descriptive statistics provide insights into the field work and data collection process of the river sample. As is common practice with data from a Big-5 inventory, we calculate correlation matrices and conduct an exploratory factor analysis (EFA) for each sample to compare them. In case there are non-linear relationships in our data (which correlation based methods wouldn't uncover), we also apply a non-linear dimension reduction method, UMAP - Uniform Manifold Approximation and Projection (McInnes et al., 2018). Finally, we analyze the distance between the two non-probability samples and the probability sample and evaluate whether we could weight the non-probability samples to arrive at the same data distribution as seen

Direct correspondence to
Stefan Zins, Institute for Employment Research, Regensburger Strasse 104, Nuremberg, 90478, Germany
E-mail: Stefan.Zins@iab.de

in the probability sample. The Discussion Section closes with a summation of the research findings to give recommendations for researchers and directions for future research.

Background

There are considerable differences in how web surveys are conducted. Couper (2000, p. 477) lists eight types of web surveys, which include three non-probability (polls as entertainment, unrestricted self-selected surveys, and volunteer opt-in panels) and five probability-based methods (intercept surveys, list-based samples, web option in mixed-mode surveys, pre-recruited panel of Internet users, and pre-recruited panels of full population).

One web survey method popular in market research is river sampling (Baker et al. 2010; Baker et al., 2013; Baker et al., 2016; Couper, 2013; DiSogra, 2008; Smith, 2012; Terhanian & Bremer, 2012; Olivier, 2011), often implemented as a collection method in which a pop-up invitation appears on the computer screen of website visitors who can then participate in the survey. Couper (2000) classifies river sampling as an unrestricted self-selected survey based on a non-probability method.

The American Association for Public Opinion Research (AAPOR) task force report on *Opt In Online Panel* stated that:

> There are some indications that river sampling may be on the rise as researchers seek larger and more diverse sample pools and less-frequently surveyed respondents than those provided by online panels. (Baker et al., 2010, p. 725)

Variants of river sampling include website evaluations (Baker et al., 2010) and website customer surveys based on services such as Google Surveys[1] (McDonald, Mohebbi, & Slatkin, 2012; Sostek & Slatkin, 2018). These surveys rely on common collection methods employed in river sampling (i.e., using ads and pop-ups on websites to recruit participants). One advantage of river sampling is that it allows fast, short surveys, possibly consisting of a single question only.

Election and exit polls are two examples of single question surveys (Hillygus, 2011; Kennedy et al., 2018). Their sponsors are usually interested in information on which political candidate or party a respondent intends to vote. Election polls often include a few additional demographic questions if a respondent did not provide this information earlier, for example, during the registration for an online panel. Demographic information is frequently used to adjust survey estimates to a target population and to provide estimates by specific subgroups (e.g., voting intentions

1 Earlier Google Customer Surveys

by gender). River sampling enables rapid studies featuring single questions (or very short questionnaires). Such studies are attractive as relying on a very short questionnaire lowers response burden (Bradburn, 1978; Galesic & Bosnjak, 2009), and can be assumed to foster a more enjoyable survey experience (Silber et al., 2018) than longer surveys. Short surveys collected through river sampling can also provide a novel incentive for participation – instant feedback on how other respondents have answered the same questions (Richter, Wolfram, & Weber n. d.).

Short river sample surveys ask a very limited number of questions – with single-question surveys being the logical extreme (but widely used) – a serious drawback in the social sciences, where general population surveys last 60 minutes or more (e.g., American National Election Study, European Social Survey, World Values Survey). Even in shorter, specialized surveys, scientists are usually interested in multivariate relationships, not estimating a single parameter (e.g., voting intention). They are interested in multivariate relationships, with many variables of interest, such as personality traits (John, Donahue, & Kentle, 1991) or values (Schwartz, Lehmann, & Roccas, 1999). These psychological measures are typically estimated using multi-item scales in order to arrive at reliable estimates of the (latent) traits. This leads us to one of the major research questions of our study: Can a psychometric instrument be successfully fielded with recruitment via a river sample and a sequence of independent single question surveys?

To the best of our knowledge, no published study has explored whether single-question river sampling surveys are feasible for substantive research, whether applying such a survey method will obtain accurate data, and whether weighting can correct biased river sample-based estimates. This dearth of information is concerning, given the rise in popularity of river sampling. In Germany, some of the largest media outlets such as Der Spiegel, Süddeutsche Zeitung, Welt, and Tagesspiegel regularly use this methodology (Höfele, 2018). Results obtained from these surveys (e.g., election polls) attract considerable media attention and are socially and politically important. Scientists, citizens, and policy makers are left without empirical evidence on which they can interpret these results or whether to purchase such data.

Data

Samples

This study is based on three different sample surveys conducted on adults in Germany. The three surveys were similar with regard to the target population but differed with regard to sampling approach (probability sample, online-access panel sample, and river sample), and the measurement approach (single-question vs. mul-

tiple questions).[2] In all three surveys, the same set of items was administered (see Section Measurement Instrument), allowing us to compare the distribution of the data arising from each sample.

The Probability Sample

Our probability sample is the GESIS Panel, a self-administered mixed-mode general population panel in Germany (Bosnjak et al., 2018). There have been two recruitments for the panel. The first GESIS Panel recruitment was done offline in 2013 based on a probability sample, where the target population was defined as persons between 17 and 71 years old that permanently reside in Germany (GESIS Panel, 2018, sec. 1). The sampling design in 2013 had two stages. At the first stage, German municipalities were selected and at the second stage, persons were sampled from the population registers of the selected municipalities. The sampling design for the first wave was planned to give equal inclusion probabilities to all persons in the sampling frame. The second recruitment, in which a refreshment sample was added to the panel, was in 2016. For the refreshment sample the 2016 German General Social Survey was used as a vehicle for the recruitment (see Schaurer & Weyandt, 2018). The register sample was again based on a probability sample and had two stages (persons in municipalities) selecting persons from 148 municipalities. It encompassed the German-speaking population aged 18 years and older.

The GESIS Panel went fully operational in 2014. Since then respondents were interviewed six times per year via web or mail. Each panel wave features a questionnaire duration of about 20 minutes. The measures we use were fielded in the first wave of 2017 (wave *ea*). These data were collected between February 14 and April 18, 2017. 3447 panel members were invited, 1121 in the mail and 2327 in the online mode. The online participants received two reminders, whereas the mail participants did not receive a reminder. Overall, 3125 respondents completed the questionnaire, yielding a completion rate of 90.6% (AAPOR, 2016). Considering the two modes, 2124 respondents completed the survey online (91.3%) and 1001 respondents completed the offline questionnaire (89.3%). The cumulative response rate (CUMR1) of wave *ea* was 20.9% (Pötzschke, Bretschi, & Weyandt, 2017).

The Online-access Panel Sample

Data were collected with an online access panel (OAP) survey conducted by a commercial online survey institute in Germany. A non-probability sampling method was used to select the respondents. The target population were persons between

2 As the analysis is interested in seeing if an online survey produces similar results to more traditional methods, we treat all three samples as if their frame were the same. That is, we will assess whether the sampled populations differ later.

the ages of 18 and 65 years, with access to the internet, who live in Germany. Quota sampling was used to select persons from the OAP, with quotas set for age categories ([18 - 29], [39 - 49], [50 - 65]), gender (male, female), and educational attainment (without/basic degree, secondary degree [10 grade - 13 grade], tertiary degree [university]) based on the German Census 2011. That is, the recruitment of respondents from the OAP continued until the set quotas for the before mentioned variable were fulfilled. A small monetary incentive of approximately 2.50 EUR was paid to respondents upon completion of the survey. Participants who failed an attention check question were excluded from the survey. 419 respondents who completed the survey were screened out and excluded from subsequent analyses because they (1) were not part of the target population because they were still at school or were non-native German speakers; or (2) did not pass an attention check. This attention check consisted in a single item asking respondents to choose one out of 10 response options in order to test the proper functioning of the survey tool. In total, interviews were completed and the completion rate was 84% (AAPOR, 2016).

The River Sample

Our river sample survey was conducted by a commercial vendor from Germany that specializes in gathering data via river samples. The target population consisted of persons aged 18 years or older that resided within Germany at the time of the survey. To conduct its river samples the vendor cooperates with numerous media outlets that embed the vendor's survey tool, a so-called widget, into their websites. The surveys were all single item questionnaires (see left panel Figure 1). The left panel of Figure 1 shows one of our Big-5 items and the right panel shows results to a respondent after completing a single question survey. Although it is not one of our questions, a Big-5 item is shown as the second option for a follow-up single question survey.

Potential respondents who clicked on the widget, if they traversed one of the cooperating websites, had the option to answer a one item survey. With that first survey, the user was asked to register. As part of the registration, the following information was requested: year of birth, gender, and postcode of the place of residence. If the respondent agreed that her or his data can be used and stored by the vendor, a browser cookie was set which was used to recognize a respondent if she or he participated in another survey of the vendor. After a respondent answered its first survey, additional single item surveys were presented to him or her (see right panel in Figure 1). A proprietary algorithm made this suggestion, which could be surveys from other customers of the vendor or from the vendor itself, to gather additional data on the respondents, like education, marital status, and employment. Through the prioritization of certain surveys that were presented to a respondent at a particular time, the algorithm directed the speed with which data for a survey was

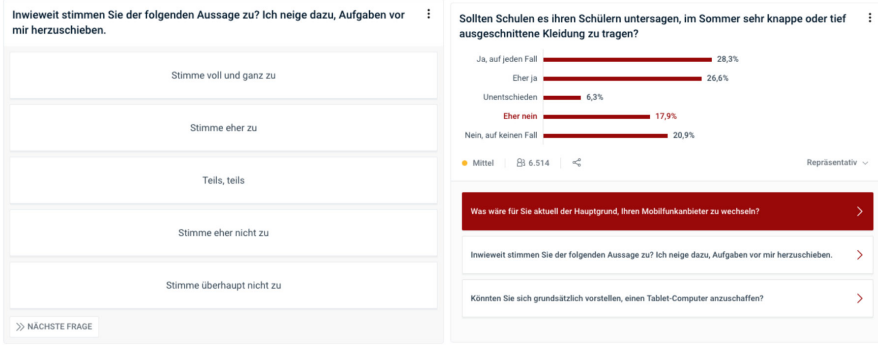


Figure 1 Examples of a single item questionnaire with follow-up questions (River Sample)

gathered. If a high priority was given to a survey, many respondents saw it and were asked to answer it and vice versa. At any time, a respondent could stop answering the suggested surveys. If he or she decided to participate later in another survey of the vendor, the browser cookie was the only tool to recognize the respondent. That is the respondents or users didn't have to actively login, it was sufficient if they accessed the survey tool from the same browser or account, if the browser was synchronized over a cloud service that stores the cookies too.

The reliance on cookies of course means that if a user cleared his or her browser data (including cookies) the survey tool of the vendor (e.g., the widget that is embedded on the website of a media partner) did not recognize the user and treated him or her as a new user and thus asked again to register. Users could also actively create an account with the vendor to log in and to respond to questions that were presented to them. However, most of the users were assumed to be casual users, i.e., the only way to link data to a respondent ID was via a cookie, which could easily be deleted by any user.

Measurement Instrument

As mentioned, the feasibility of administering multi-item inventories through river sampling has not yet been empirically established. Additionally, there were survey methodological and technical limits to the number of items that we could administer through river sampling. These limitations implied that we could not administer a full-length Big Five inventory but had to select a subset of dimensions and items. We chose *Conscientiousness and Emotional Stability* as measured by the short version of the well-validated Big Five Inventory 2 (BFI-2) (Soto & John, 2017; Danner

et al., 2019). Our rationale for choosing Conscientiousness and Emotional Stability was twofold. First, the BFI-2 measures of these two dimensions have very good psychometric properties, including high internal consistencies and good factor-analytic separation (Soto & John, 2017; Danner et al., 2019). These dimensions lent themselves ideally for comparisons of the data of the three surveys under study. Second, Conscientiousness and Emotional Stability show robust links to important life outcomes such as income or health; in other words, they are of high substantive interest to researchers and practitioners alike (Roberts et al., 2007; Rammstedt, Danner, & Lechner, 2017). Each of the two personality domains was measured with 6 items (i.e., 12 items in total), of which three were positively worded, and three were negatively worded, in order to control for acquiescent responding. All BFI-2 items are phrased as short self-descriptions (e.g., 'I am helpful and unselfish with others'). Respondents rated each of these items on the same fully labeled 5-point rating scale (1 = 'Disagree strongly' to 5 = 'Agree strongly').

Although wording and response scales were identical across the three surveys, the way in which these items were presented differed between the river sample and the other two samples. In the OAP sample and the GESIS Panel sample, the item battery was preceded by an introduction that was close to the original introductory statement from the BFI-2, which reads as follows: *Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement*. In both samples, the items were then presented as grid questions. In the river sample, by contrast, the single-question approach necessitated that each item was preceded by the sentence, *to what extent do you agree with the following statement*, followed by the item itself. Tables A.1 and A.2 in the Online Appendix show the BFI-2 scoring information, questions, and item labels we used, respectively, both in English and German.

The twelve items of our river sample surveys were split into two groups of equal size. Within each group, there were 3 items from each of the two Big-5 domains (Conscientiousness and Emotional Stability).[3] The first group of 6 single item surveys was fielded on 09.07.2018 and the second on 11.07.2018. The decision to not field all 12 items on the same day was taken by the vendor to avoid presenting respondents with too many Big-5 items on the same day.[4] The grouping was not

3 Note that the grouping of items into two groups of six items with three items from each of the two Big-5 domains did not change the setup of the river sample as single item questionnaires. The groups only determine when the items were fielded. The reason for the grouping was that the river sample vendor had concerns fielding all 12 items at once and thus suggested fielding half of the items first and the rest 3 days later.

4 Due to the lack of research on the applicability of river sampling to the needs of social science research, we discuss the fieldwork outcomes in more detail as part of our results in Section Fieldwork Outcomes of the River Sample.

random, given the ordering of the questions in the instrument (see Table A.2 in the Online Appendix) every second question was allocated to the second group and the rest to the first group. However, we did not control for the order of questions within the groups, i.e. when the first group was fielded on 09.07.2018, there was not a fixed order in which the questions were presented to the respondents. The algorithm of the vendor, which determines which question is shown to which respondent, prioritized our questions for a certain number of days (around 2 -3 days). However, our questions were not the only ones that vendors fielded during our time of fieldwork. Hence, we do not know what questions, from other customers of the vendor, were also shown to our respondents, between answering our questions.

Methods

Based on our research goals, we focused on addressing three broad research questions: How similar are the datasets? Does an EFA produce similar results when run on each dataset? Is it possible to transform non-probability datasets into equivalents of a probability dataset? Our probability sample serves as a point of reference, to which we compare the non-probability samples. This approach is based on the assumption that the probability sample enables better statistical inference than the non-probability samples (Meng, 2018). All analyses were conducted using only complete cases, as imputation techniques usually require data to be missing completely at random (MCAR), which we certainly violate, or missing at random (MAR), where we have observed the variable(s) determining missingness (which is also unlikely). Planned future research should examine imputation when dealing with non-probability samples.

We address the first and second of the three research questions by examining the multivariate distributions of the datasets, using both linear and non-linear dimension reduction, as well as a simple comparison of each dataset's correlation matrix. The non-linear dimension reduction is particularly important, as a linear method of comparison can mistakenly claim that data are similar when the underlying relationship is non-linear. We did not see a way to build a consistent estimator for the sampling variance of our two non-probability samples, as compared to probability samples (Särndal, Swensson, & Wretman 1992, sec. 2.8). Therefore, the comparisons reported in this study do not include significance tests. While there are publications that discuss measures of variance for non-probability samples (Salganik, 2006) we do not regard these methods as applicable here. We have no information on the sampling design for our non-probability samples in the form that would be needed to conduct design-based variance estimation, i.e. we have no way of knowing how the distribution of any estimators looks like under the non-probability sampling designs.

Our third research question is addressed by evaluating if a frequently referenced method for correcting bias in non-probability samples – weighting – is actually capable of doing so. This is accomplished by displaying the distribution of Euclidean distances of data points between the non-probability samples and the probability sample, as well as examining the results of the non-linear dimension reduction for "holes", that is, parts of the probability sample distribution that have been completely missed by the non-probability sample.

To see how the three samples differ with respect to their gender and age distributions we refer to Tables A.7.1, A.7.2, and A.7.3 in the Online Appendix. The samples display a large dissimilarity regarding age and gender. The river sample, as shown in Table A.7.3, has a very high concentration (45.8%) of male respondents in the age range of 50 to 69 years. The age variable of the river sample also contains some rather implausible values (i.e. a number of values over 90 up to 115), indicating that some respondent might deliberately provide false demographic information.

Linear Dimension Reduction

Factor analysis is a method for linear dimension reduction with the objective of creating a lower dimensional representation of an observed correlation matrix (Spirtes et al., 2000, 76). It was developed during the 1930's (Thurstone, 1935), with roots in Spearman's earlier attempt to justify the existence of a single unobserved variable *g*, which he thought measured *general intelligence* (Spearman, 1904). A large literature has since developed on how to use factor analysis in an *exploratory* way, where the number of common *factors* used to summarize an observed correlation matrix is not initially known. One common method is to determine the point at which adding an additional factor fails to account for a significant improvement in the amount of variance accounted for, often using either a scree plot (Cattell, 1966) (with the inflection point in the plot being the suggested number of factors to reduce to) or various numerical approximations of the scree plot's inflection point. We calculated various numerical approximations of the inflection point using some of the more common methods – the Kaiser rule (Kaiser, 1960), parallel analysis (Horn, 1965), acceleration factor (Raîche et al., 2013), and optimal coordinates (Raîche et al., 2013), and plotted the results. Both the plots and the numeric calculations were performed using the method of Raîche and Magis (2010).

We opted not to test the hypothesized Big-5 psychometric measurement model that underlies our measurement instrument using confirmatory factor analysis (CFA), as the expected bias in our datasets would result in a CFA (or Structural Equation Model) with incorrectly estimated goodness-of-fit statistics. Just as a non-probability sample can result in biased estimates of means (and linear regression coefficients), a CFA would estimate biased goodness-of-fit statistics, making tra-

ditional tests, such as a chi-square test, unreliable. We instead did an Exploratory Factor Analysis (EFA), simply to compare what conclusions (if any) would differ between the EFA when performed on the different datasets[5]. As an EFA does not straightforwardly map onto a discussion of biased fit estimates the way a CFA would, this analysis should be of interest to researchers, despite its deviation from a more traditional approach (i.e. CFA) when analyzing a pre-existing collection of psychometric instruments.

Non-linear Dimension Reduction

As there may also be differences between our samples that are missed by an analysis focused on linear relationships in our data, we also employed another dimension reduction method, UMAP. Uniform Manifold Approximation and Projection (UMAP) is, informally, a non-linear, non-parametric dimension reduction procedure which attempts to perform its reduction on the high dimensional space the observed data occupies, rather than the individual observations. This results in a lower dimensional space, constructed to minimize the amount of information lost about the higher dimensional space, that the observations are then projected onto. As this reduction is non-linear, it can work to preserve relationships that would be excluded when using a linear method (such as factor analysis, which performs its reduction on a correlation matrix), ensuring that we get a more complete picture about the high dimensional distribution of the datasets. We used this method to project the data from each sample onto a two-dimensional plane with continuous measures. Then we applied a two-dimensional kernel density estimation on the reduced datasets to visualize the continuous two-dimensional representation of the three data sets. We used the implementation of UMAP described in (McInnes et al., 2018). The UMAP implementation that we used is relatively new and there are not many publications that feature its use. Nevertheless, Becht et al. (2018) showed an application of the UMAP method to biological data. UMAP, being a non-linear dimension reduction procedure, creates lower dimensional representations that, while useful for prediction, are not interpretable in the normal sense. While the lower dimensional representations have meaningful distances between observations, reifying (i.e., naming and treating the dimensions as if they were something directly measurable) is not typically possible. For example, we cannot justify saying that dimension 1 is “happiness,” but we could say that two observations are separated by 5 units on dimension 1.

5 Performed using varimax rotation as the Big-5 factors are theoretically independent.

Distance Analysis and Weighting

As a supplement to our two dimension reduction methods, we investigated two kinds of distances between our datasets. What and how many (if any) combinations of observations of the 12 items of our measurement instrument we have are never observed (i.e., a *hole* in the distribution), and the distribution of Euclidean distance between observations in the GESIS Panel and the two non-probability samples. Holes are important when considering weighting approaches for making a non-probability sample more similar to a probability sample, as their presence prevents reducing the distance between the two samples to 0, which can result in bias. Because there is no weighting procedure possible that would transform the data distributions of the non-probability samples to that of the GESIS Panel, or any weighted data distribution of the GESIS Panel. For example, if a sample has no observations from some demographic category or group, then one cannot adjust that category's influence (or lack thereof) on a global estimate – say, voting intentions – as there is no data whose influence can be changed.

We conducted our distance analysis as follows: We first looked at the overlap (or lack thereof) in the distribution of each variable, followed by examining the distribution of Euclidean distances from all observations of the non-probability samples to all observations in the GESIS Panel.

Results

The following section includes the results from our analysis of the three samples as described in the Methods Section and the analysis of the fieldwork outcomes of the river sample study. All results shown in Section Comparing the Non-probability Samples to the Probability Samples have been conducted with complete cases only, i.e. only respondents were considered for which data from all 12 items of our measurement instrument was available. For Section Fieldwork Outcomes of the River Sample all cases of the river sample have been considered.

Fieldwork Outcomes of the River Sample

In the river sample, the multi-item scale could only be implemented under the restriction of using a very large sample, since only 29.9% of respondents answered all 12 items. The river sample was gathered over the course of 31 days. 15915 respondents answered at least one of the 12 items, with 4771 complete observations (i.e., respondents answering all 12 items). By the 5th day, we obtained roughly 75% of our total 4771 complete cases. This is observable in the empirical cumulative distribution plot shown in Figure 2. This rapid rate of data collection is likely due

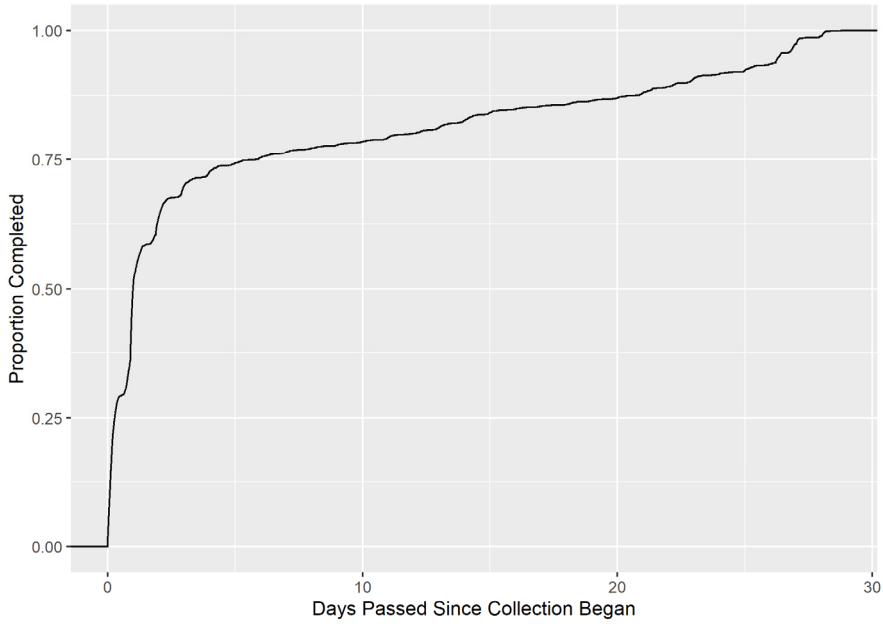
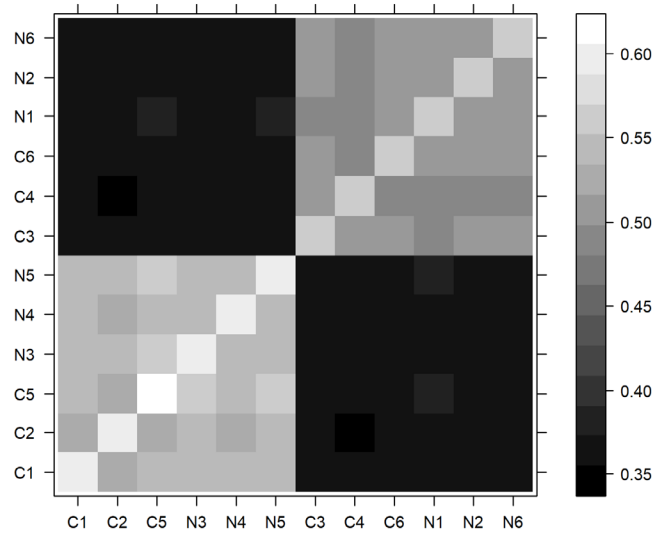


Figure 2 Empirical cumulative distribution of complete cases over time.

to the high priority the vendor gave our questions for the first four days, with half of the 12 questions introduced and prioritized for the first 2 days, then the second half introduced and prioritized for the second two days. No participant could complete all 12 questions until the third day, as only half of the questions were available prior to then. Once the prioritization ended, our questions were answered less frequently, with a brief spike in the number of answered questions in the final days. Respondents usually answered our questions between 10:00AM and 3:00PM, which might coincide with either their work or lunch break. Detailed information on the time of responses is displayed in Figure A.1 within the Online Appendix.

The median gap between a participant answering one question and answering another was 187 seconds, with a mean gap of 28.38 hours, a minimum gap of 2 seconds, and a maximum of 1.18 days.[6] The median gap between a respondent answering their first and last question was approximately 22 minutes. A tabular

6 The large difference of 2 seconds and more than 1 day between answering the questions illustrates that some participants answered the single item questionnaires of the river sample in a similar way as a standard survey, while others took long breaks between answering the questions. Also, we have no information on whether a participant answered other questions before or in between our questions, making the survey context arbitrary.



Note: The shading of the squares represents the proportion of the sample where answers to both the question on the x-axis and y-axis are available.

Figure 3 Level plot of question response overlap.

summary of the time gaps between consecutively answered questions and the time between respondents answering their first and last question can be found in the Online Appendix in Table A.3 and A.4 respectively.

Figure 3 shows the overlap between respondents that answered the same two questions, as a percentage of the total number of respondents (15915). The diagonal of the plot shows the percentage of respondents that answered each individual question. There is a clear pattern to be observed. The initial batch of questions (C1, C2, C5, N3, N4, and N5) were primarily answered by the same people, while the second batch (C3, C4, C6, N1, N2, and N6) were mainly answered by a second different group. Also, a higher percentage of respondents answered the questions of the first batch, which might be explained by the fact that those questions were two days longer in the field than the other questions.

Comparing the Non-probability Samples to the Probability Sample

Tables A.8.1 and A.8.2 in the Online Appendix show the measured means and the coefficients of variation of the 12 survey items for each of the three samples. There appears not to be any large variation between the item means across the samples. The coefficients of variation also do not display any large variation across the sam-

ples. However, the GESIS Panel has for all but one item (C1) the lowest coefficients of variation. Thus, a univariate comparison between the samples does not reveal any notable difference between the measurements obtained from the three samples. The remainder of the section will focus on the multivariate comparison.

Correlation

Figure 4 displays the correlation matrices of the three data sets. The size of the circle are proportional to the correlation coefficients. White circles indicate a positive and balck circles a negativ correlation. For all data sets, we can observe a stronger correlation between variables that should measure the same Big-5 domain, e.g., Conscientiousness for the *C* variables and Emotional Stability for the *N* variables. For all three samples, almost all correlations are in the same direction. In addition, the magnitude of the correlation is similar across the samples, although not as consistent as the direction. However, it cannot be said that one of the non-probability

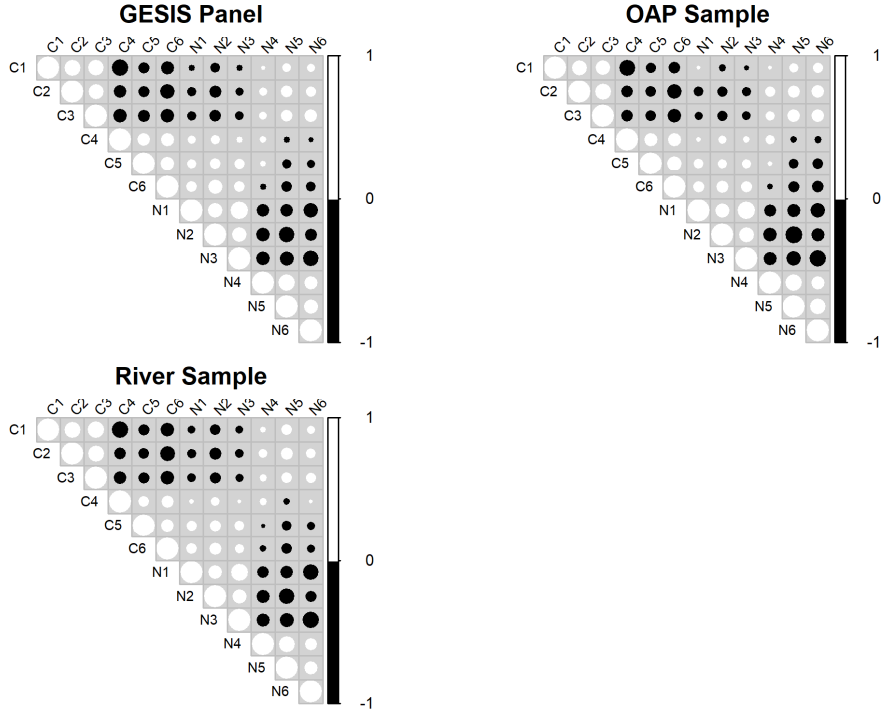


Figure 4 A graphical depiction of correlation matrices for the 12 items of our measurement instrument

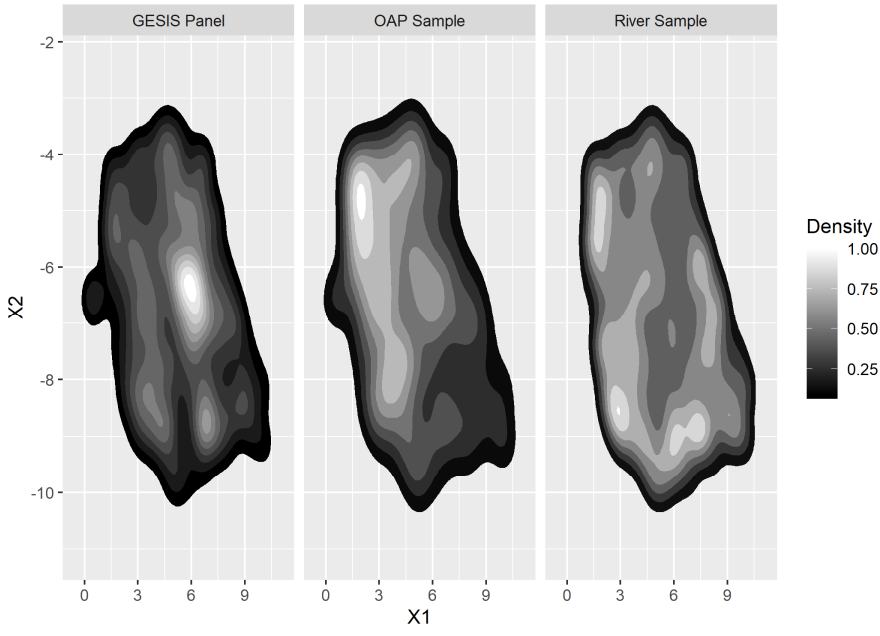
samples is more similar to the probability sample than the other. The complete correlation matrices can be found in the Online Appendix (see Tables A.5.1 - A.5.3).

Exploratory Factor Analysis

All traditional methods for deciding on the number of factors to extract produce similar results (3 factors), with only one selection criteria (the acceleration factor) opting for a different number of factors (1 or 2). The acceleration factor recommends 1 factor when run on the non-probability panel sample, while it recommends 2 in the case of the river sample. The graphical scree plots can be found in the Online Appendix in Figures A.2, A.3, and A.4, for the GESIS Panel, the OAP sample, and the river sample, respectively. All three plots look very similar. Factor loadings (i.e., the correlation between observed measures or items and the hypothesized latent variables) are also moderately similar across all three datasets (signs and magnitudes are fairly similar), though the river sample produces somewhat more different results than the other two samples. The factor loadings, using varimax rotation, for each dataset are available in the Online Appendix in Table A.6. If the factor loadings for each sample strongly differ when varimax rotation is used (i.e., different groups of measures were associated with different factors), then we would be able to conclude there are serious differences between the samples, as such a difference would be unusual. However, as Table A.6 shows this is not the case, as signs and magnitude of the factor loading show the same patterns across the three samples.

Non-linear Dimension Reduction

Applying UMAP to the combined dataset from all three samples, allows us to extract and compare two continuous variables. As these variables are non-linear representations of higher dimensions, their interpretation is unclear, i.e. they have no obvious substantive meaning. Figure 5 shows the contour plots for each of the three samples that visualize the kernel density estimates for their two-dimensional data. As the color of a given level lightens, the density estimate increases, meaning more data is observed in that area (this can be thought of as an increase in elevation in a topographic map used when hiking). When comparing the plots, the probability sample looks very different from the other two datasets. The non-probability samples and the GESIS panel differ in where their peak densities are located, with the peak density of the GESIS panel ($X1 \approx 6$, $X2 \approx -7$) occupying a low density region of both the non-probability and (especially) the river sample. This suggests that, based on the observed dimension reduced data, that there are fundamental differences in the sample composition of people's *personalities* in the sample. Also, in the region where the GESIS Panel and the OAP sample have a number of observa-



Note: Dimensions are non-linear and not straightforwardly interpretable.

Figure 5 Contour plots of two-dimensional data.

tions ($X1 \leq 0$), the river sample has essentially no observations. This missing region in the river sample suggests a possible gap in its distribution, where some kinds of respondents are being systematically missed.

Distance Analysis and Weighting

As we have seen from the non-linear dimension reduction, there seems to be a difference between the multivariate data distribution of the three samples. There are 5^{12} possible permutations of our psychological measures. If we check for how many of these permutations (sometimes referred to as *cells*) exist in our three datasets, we observe the following: The ratio between unique cells and sample size in the river sample is 0.98, for the OAP sample it is 0.99, and 0.93 cells for the GESIS panel. A ratio of 1 would imply that all respondents have different measurements and a ratio of $\frac{1}{n}$ implies the opposite, with n being the number of respondents in the sample. This shows that regardless of sample size almost all respondents in all three samples produce unique measurement combinations. The GESIS Panel has marginally more homogeneous responses, which is also visible in its less dispersed dimension reduced data, seen in Figure 5.

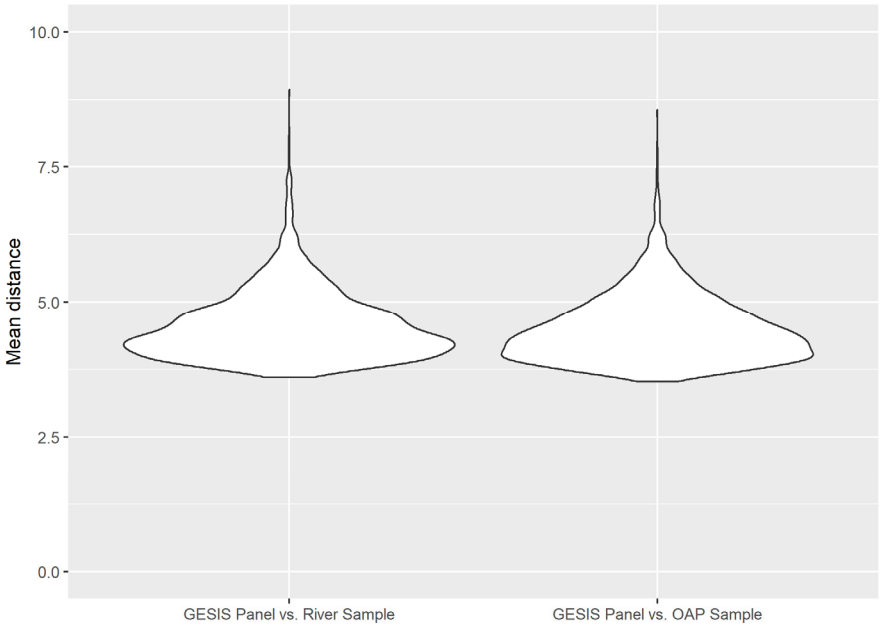


Figure 6 Violin plot of distance between GESIS Panel and non-probability samples

As a measure of overlap between the samples, we examined how many of the cells from the probability sample exist in the two non-probability samples. There are very few elements in common between the two non-probability samples and the GESIS Panel, with the river sample having a ratio of equal responses to its sample size of 0.03, and the OAP sample ratio of 0.05. A ratio of 1 would imply that all observations are the same for the probability and non-probability samples and a ratio of $\frac{1}{n}$ implies the exact opposite, with n being the number of respondents in the non-probability sample.

Given the number of possible permutations of our variables, the lack of overlap does not necessarily imply our datasets are extremely far apart. After all, a single variable differing by 1 would be enough to cause an observation to belong to a different cell. To assess how distant the datasets are from one another, we calculated the distribution of Euclidean distances between all observations of the non-probability samples to all observations in the probability sample. The results are displayed in the violin plot[7] shown in Figure 6. The violin plot displays the distribution of the mean distances that every GESIS Panel respondent has to all respon-

7 Violin plots are similar to box plots, except the box is created by mirroring a density plot around the y-axis (Hintze & Nelson, 1998).

dents in the OAP sample or the river sample. Figure 6 shows that the OAP sample is as distant from the probability sample as the river sample is. This is consistent with the information we received from the UMAP dimension reduction procedure, with the majority of the contour regions overlapping. As it is common practice to weight using demographic variables but some of our datasets use incommensurate definitions for demographic categories (e.g., education), we opted to explore what the best linear transformation of the data would produce. That is, how similar could we make the non-probability sample data set to the probability sample data set, using a linear transformation? The details of the procedure we used, which is essentially a multivariate regression, are discussed in detail within the Online Appendix, along with the results. But the method amounts to having a separate weighting vector for every item. As can be seen in Figure A.5 (Online Appendix), transforming our non-probability datasets using the estimated transformation matrices, as described in the Online Appendix (Weighting section), greatly reduces their average distance from the probability sample. The transformation does not shrink the distance to zero, and the two datasets end up with very similar distance distributions. This suggests that with survey weights, although it is not clear what auxiliary data would be needed to construct them, the two non-probability data sets could produce similar estimates. Note that any single weighting vector for all survey items, as it is usually the case in survey data analysis, could not reduce the dissimilarity between the non-probability and probability data set any more than the method we present in the Online Appendix.

Discussion

In our study, we investigated the possibility of using single-question river sampling surveys for substantive research. We found that many respondents had to be surveyed to achieve a sufficiently large number of complete cases (i.e., respondents who chose to answer all 12 questions), and we show that data can be gathered for projects that only require a very limited number of variables. Yet, from the perspective of survey operations, a variety of questions remain unanswered with respect to river sampling approaches; for instance, how the process of the respondent based selection of questions influences survey-outcomes or whether more complex question formats that exceed the standard closed-ended response formats can be employed. For scientific purposes, non-probability samples have often been used in connection with survey experiments (Mullinix et al., 2015) under the assumption that experiments help to mitigate biases of these samples – some of which have been discussed in this paper. When using non-probability online-access panels, the implementation of experiments seems straightforward, whereas the reliance upon proprietary question allocation algorithms and respondent self-selection into ques-

tions in the river sampling approach might impair the application of similar methods in this setting. More research on design restrictions when using river sampling approaches is warranted in order to shed more light on its applicability for social science research.

If we restrict our discussion to just the measurement instrument, the co-variance structure looks similar across the three samples. This is also what we observe in the EFA results. While the river sample does not use a multi-item questionnaire, the reduced correlation matrix for the respondents appears to be reasonably consistent with the other two samples. At the same time, the UMAP representation for each of the three samples is very dissimilar, which could be an indication that the sample composition of personality types is very different. If we take the Big-5 personality model seriously, this is perfectly consistent with the EFA results, as no sample selection bias should result in a different structure underlying personality. No matter how we sample, we are still sampling people, and the underlying personality structure should not change. One reason we may not observe such a difference in the correlation matrix is that the differences involve non-linear relationships between the variables, which traditional measures of correlation cannot detect. As UMAP allows for non-linear relationships between variables, it would still be capable of detecting such differences. As Thurstone (1935, 206) observed, latent structures are often unlikely to be adequately represented by linear relationships, but rather by non-linear and discontinuous associations.

The evidence of missing kinds of respondents in the two non-probability samples is concerning because, as discussed in Section Distance Analysis and Weighting, weighting cannot be used to reduce the distance between the different classes of sampling procedures, which is a possible sign of data missing not at random (see Särndal, Swensson, & Wretman, 1992, cap. 1). If the river sample and OAP sample selection methods generally behave the same as we have observed in our case study, then there are serious objections to their use in answering substantive science questions. These include the risk of biased parameter estimates *of unknown magnitude* in addition to an inability to determine if the results are significant or not. As we cannot state with any degree of certainty if the observations we made in our case study hold in general, or if the observed differences merely result from sample variation, our conclusions must be somewhat circumspect. The lack of data for our Big-5 scale based on a second probability sample, that has the same target population as the GESIS Panel and a similar sampling design, prevented us from assessing if the data distributions between different probability samples would have been more similar to each other than we observed for either of the three studied samples. Despite these limitations, researchers should exercise caution when using data collected with non-probability - especially river sampling - methods, if their goal is generalizable research. Until more is known, we recommend the use of a probability sample if at all possible. As our analysis showed, the high dimensional

distribution of the BIG-5 items in the two non-probability samples differed quite markedly from the probability sample. These differences might lead to different substantive findings, especially in analyses that involve mean-level comparisons. Further research is needed to assess the sampling variance of non-probability methods, such as river sampling, and reliable methods for assessing, bounding, and reducing their bias need to be developed.

Data Availability

Data from the GESIS Panel used in our study are archived in the German Data Archive for the Social Sciences at the GESIS - Leibniz Institute for the Social Sciences (http://www.gesis.org/dbk). The study number of the data used is: ZA5665 (doi:10.4232/1.12973).

Data from the OAP and the River Sample used in this study are available at the SowiDataNet | datorium, a research data repository, hosted by the GESIS Data Archive for the Social Sciences and can be accessed here: https://doi.org/10.7802/2290

Software Information

For all analytical tasks, including figures, author-originated code was written entirely in R. For the software implementing the UMAP method, an R interface to Python was used. All author-originated code and data are available at the SowiDataNet | datorium (see above).

References

AAPOR (2016). *Standard definitions final dispositions of case codes and outcome rates for surveys.* Retrieved April 19, 2012, from the American Association for Public Opinion Research website: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., Sanford, R. N., et al. (1950). *The authoritarian personality.* New York: Harper & Rowe, Inc.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., et al. (2010). Research synthesis: Aapor report on online panels. *Public Opinion Quarterly*, 74(4), 711-781.

Baker, R., Brick, J., Keeter, S., Biemer, P., Kennedy, C., Kreuter, F., & Terhanian, G. (2016). *Evaluating survey quality in today's complex environment.* American Association for Public Opinion Research, Oakbrook Terrace, IL.

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.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). *Dimensionality reduction for visualizing single-cell data using umap*. Nature biotechnology.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in germany: The gesis panel. *Social Science Computer Review*, 36(1), 103-115.

Bradburn, N. (1978). *Respondent burden*. In Proceedings of the Survey Research Methods Section of the American Statistical Association, 35, p. 40. American Statistical Association Alexandria, VA.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2), 245-276.

Cornesse, C. & Bosnjak, M. (2018). Is there an association between survey characteristics and representativeness? a meta-analysis. *Survey Research Methods*, 12, 1-13.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, 64(4), 464-494.

Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quarterly*, 75(5), 889-908.

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3), 145-156.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). xtable: Export Tables to LaTeX or HTML. R package version 1.8-4.

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C. J., & John, O. P. (2019). Das big five inventar 2: Validierung eines persönlichkeitsinventars zur erfassung von 5 persönlichkeitsdomänen und 15 facetten. *Diagnostica: Zeitschrift für psychologische Diagnostik und differentielle Psychologie*, 65(3), 121-132.

Daróczi, G. & Tsegelskyi, R. (2018). pander: An R ‚Pandoc' Writer. R package version 0.6.3.

DiSogra, C. (2008). *River samples: A good catch for researchers*. GfK Knowledge Networks. http://www.knowledgenetworks.com/accuracy/fall-winter2008/disogra.html

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly*, 73(2), 349-360.

Gandrud, C. (2016). repmis: Miscellaneous Tools for Reproducible Research. R package version 0.5.

GESIS Panel (2018). *Gesis panel study descriptions*. Technical Report 26.0.0, GESIS Leibniz Institute for the Social Sciences. http://dx.doi.org/10.4232/1.12743

Hillygus, D. S. (2011). *The evolution of election polling in the United States*. Public opinion quarterly, 75(5), 962-981.

Hintze, J. L. & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2), 181-184.

Höfele, M. (2018). Meinungsforschungsinstitut Civey: Repräsentativ daneben? *Die Tageszeitung*: taz. https://www.taz.de/!5534782/

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185.

James, D. & Hornik, K. (2019). chron: Chronological Objects which can Handle Dates and Times. R package version 2.3-54.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—versions 4a and 54*.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1), 141-151.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., et al. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1), 1-33.

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

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the joint statistical meetings, American Statistical Association, (pp. 282-296).* Alexandria, VA: American Statistical Association.

Kummerfeld, E. & Ramsey, J. (2016). Causal clustering for 1-factor measurement models. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 1655-1664*). ACM.

Little, R. J. & Rubin, D. B. (2014). *Statistical analysis with missing data*. 2nd Edition. John Wiley & Sons.

McDonald, P., Mohebbi, M., & Slatkin, B. (2012). *Comparing google consumer surveys to existing probability and non-probability based internet surveys*. Google White Paper.

McInnes, L. & Healy, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. ArXiv e-prints. http://arxiv.org/abs/1802.03426

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29), 861.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2), 685-726.

Miller, P. V. (2017). Is there a future for surveys? *Public Opinion Quarterly*, 81(S1), 205-212.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2), 109-138.

Murdoch, D. (2019). tables: Formula-Driven Table Generation. R package version 0.8.8.

Nychka, D., Furrer, R., Paige, J., Sain, S., Gerber, F., & Iverson, M. (2019). fields: Tools for Spatial Data. R package version 10.0.

Oliver, L. (2011). *River Sampling: Non-probability sampling in an online environment*. Web log, November 13 (2011): 2011.

Pötzschke, S., Bretschi, W., & Weyandt, K. (2017). *Gesis panel wave report. Technical Report Wave ea*, GESIS Leibniz Institute for the Social Sciences.

Raîche, G. & Magis, D. (2010). *nfactors: An R package for parallel analysis and non-graphical solutions to the cattell scree test*. R package version, 2(3).

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for cattell's scree test. *Methodology*, 9(1), 23-29.

Ram, K. & Wickham, H. (2018). wesanderson: A Wes Anderson Palette Generator. R package version 0.3.6.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1), 203-212.

Rammstedt, B., Danner, D., & Lechner, C. (2017). Personality, competencies, and life outcomes: Results from the German PIAAC longitudinal study. *Large-scale Assessments in Education*, 5(1), 2.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Version 3.6.1.

Richter, G., Wolfram, T., & Weber, C. (n. d.). Die Statistische Methodik von Civey. Eine Einordnung im Kontext gegenwärtiger Debatten über das Für und Wider internetbasierter nicht-probabilistischer Stichprobenziehung. [in German] derived from: https://assets.ctfassets.net/ublc0iceiwck/3tBBzurQaKhIpNuR7BQJZc/e10b1712b8c73bc8058fd411f8184020/Die_statistische_Methode_von_Civey_Richter_Wolfram_Weber.pdf (accessed 11/15/2021)

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4), 313-345.

Salganik, M. J. (2006). Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 83(7), 98-112.

Särndal, C.-E. & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Schaurer, I. & Weyandt, K. (2018). *GESIS Panel Technical Report. Recruitment 2016 (Wave d11 and d12)*. Leibniz Institute for the Social Sciences, Mannheim, Germany.

Schwartz, S. H., Lehmann, A., & Roccas, S. (1999). *Multimethod probes of basic human values*. Social psychology and culture context: Essays in honor of Harry C. Triandis, p. 107-123.

Silber, H., Daikeler, J., Weidner, L., & Bosnjak, M. (2018). *Web survey*. Wiley StatsRef: Statistics Reference Online, p. 1-6.

Smith, T. W. (2012). Survey-research paradigms old and new. *International Journal of Public Opinion Research*, 25(2), 218-229.

Sostek, K. & Slatkin, B. (2018). *How google surveys works*. White paper, Google Inc.

Soto, C. J. & John, O. P. (2017). The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117.

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201-292.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT press.

Terhanian, G. & Bremer, J. (2012). A smarter way to select respondents for surveys? *International Journal of Market Research*, 54(6), 751-780.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press.

Ushey, K., Allaire, J., & Tang, Y. (2019). reticulate: Interface to 'Python'. R package version 1.13.0-9000.

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). *Non-probability sampling. The Sage handbook of survey methods*, p. 329-345.

Wei, T. & Simko, V. (2017). corrplot: Visualization of a Correlation Matrix. R package version 0.84.

Wickham, H. (2019). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.3.0.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., & Yutani, H. (2019). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.2.1.

Xie, Y. (2019a). bookdown: Authoring Books and Technical Documents with R Markdown. R package version 0.15.

Xie, Y. (2019b). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.26.

methods, data, analyses | Vol. 17(1), 2023, pp. 29-46 DOI: 10.12758/mda.2022.06

Evaluating an Alternative Frame for Address-Based Sampling in Germany: The Address Database From Deutsche Post Direkt

Sven Stadtmüller[1], *Henning Silber*[1], *Tobias Gummer*[1], *Matthias Sand*[1], *Stefan Zins*[2], *Christoph Beuthner*[1] & *Pablo Christmann*[1]

[1] *GESIS – Leibniz Institute for the Social Sciences*
[2] *Institute for Employment Research*

Abstract

In Germany, the population registers with addresses of individuals can be used for address-based sampling. However, unlike countries with a centralized register, municipalities in Germany administer their registers themselves. This not only makes sampling for a nationwide survey more costly and cumbersome but may also result in gaps in the gross sample, as selected municipalities may refuse to allow their registers to be used for sampling purposes. If substitute municipalities are not available, other sampling methods are required. The present study tested the feasibility of using the address database from Deutsche Post Direkt (ADB-DPD) as an alternative frame for address-based sampling in Germany. We simultaneously conducted two almost identical surveys in the German city of Mannheim with gross samples of equal size ($N = 3{,}000$). One sample was drawn from the city's population register, the other from the commercial ADB-DPD. Our findings suggest that the ADB-DPD performs well both in terms of survey response and up-to-dateness. Due to relatively low costs and the fast provision of addresses, the ADB-DPD could be particularly attractive for survey projects with limited budgets and tight schedules. However, these benefits come at considerable cost. First, the use of the ADB-DPD is limited to self-administered surveys. More importantly, in the net sample of the DPD survey, women and young persons were considerably underrepresented. This indicates coverage issues about which DPD provided no further information. Based on our analyses, we offer practical insights into the feasibility of using the ADB-DPD for sampling purposes and suggest avenues for future research.

Keywords: address-based sampling, alternative sampling frame, population register, sample evaluation, sample composition

In many countries, researchers rely on official population registers for address-based sampling. Registers used for personal surveys should ideally include addresses of individuals (as opposed to households). This not only avoids the necessity of selecting target persons within households or dwellings but also allows researchers to personalize their contacts from the beginning, which is known to be beneficial in terms of survey response (Dillman, Smyth, & Christian, 2014).

However, some countries, for example, the United States, the United Kingdom, and France, either lack official population registers completely, or their registers include only addresses of households (Link, Battaglia, Frankel, Osborn, & Mokdad, 2008; Poulain & Herm, 2013). In these instances, survey researchers have to settle for suboptimal registers or, when no register is available at all, use alternatives for address-based sampling. In the United Kingdom, for instance, Royal Mail's Postcode Address File (PAF) has been used as a sampling frame for several national and cross-cultural surveys, for example, the European Social Survey (European Social Survey, 2017). In a similar vein, survey managers in the United States often rely for address-based sampling on address lists updated via the U.S. Postal Service's Computerized Delivery Sequence File (CDS; Harter et al., 2016), even though this sampling frame is known to suffer from systematic undercoverage (e.g., rural household units are more likely to be excluded; Amaya, Zimmer, Morton, & Harter, 2021).

Although register-based sampling is generally regarded as the gold standard for drawing representative samples of the residential population (Lohr, 2009), registers also have their own challenges. In a survey among sampling experts in countries participating in four cross-European surveys, respondents mentioned undercoverage and inaccuracies as the main problems they encountered when using their countries' population registers for sampling purposes (Maineri et al., 2017). Another obstacle mentioned by the sampling experts was access to population registers, which varies considerably across countries: More than half of the sampling experts reporting the use of a register-based sample stated that commercial survey organizations do not have access to population registers for sampling purposes (Scherpenzeel et al., 2017).

In Germany, researchers and survey organizations can access population registers with addresses of individuals for the purpose of address-based sampling for academic surveys. At first glance, this seems to be a comfortable situation. However, unlike most countries with an official population register, Germany does not have one centralized register, but rather local population registers administered by the over 5,000 municipal registration authorities (Federal Ministry of the Interior

Direct correspondence to
Sven Stadtmüller, Survey Design and Methodology Department,
GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
E-mail: sven.stadtmueller@gesis.org

and Community, n.d.). Sampling for nationwide personal surveys in Germany is therefore usually carried out in two stages. In the first stage, municipalities are selected based on a stratified random sample. In the 2014 German General Social Survey (ALLBUS), for instance, 148 municipalities were selected in the first stage (Wasmer et al., 2017). To build the gross sample in the second stage, it is necessary to contact all selected municipalities and request them to provide a random sample of addresses. It is easy to imagine how time-consuming this process can be. Moreover, most municipalities charge a fee for drawing and transmitting a random sample from their registers. In sum, the lack of centralization of the German population registers makes sampling more demanding and a rather costly process in terms of time and funds. Using the German population registers to draw a sample may thus not be feasible for research projects with small budgets or tight schedules.

Most importantly, access to the respective population registers depends on autonomous decisions on the part of each municipality. According to Section 46, Paragraph 1 of the Federal Act on Registration (BMG),[1] information from the population register may be released only if it is in the public interest to do so. As this gives municipalities a certain leeway, some of them refuse to provide the desired addresses if the research purpose is not deemed to be in the public interest for one reason or another. Although the proportion of municipalities that refuse to release information for research purposes appears to have been rather low in the past,[2] we expect municipalities to behave differently in light of the European Union General Data Protection Regulation (GDPR), which came into effect on May 25, 2018. We assume that municipalities might now be more reluctant to supply personal data because of (a) uncertainties related to the GDPR and (b) higher public awareness of data privacy. If a municipality refuses to provide addresses, a substitute is used. However, substitution is methodologically problematic, especially if no structurally equivalent surrogate municipality exists (e.g., for the capital of a federal state). Nonresponse at municipality level may result in selection error that can have a stronger effect than nonresponse at individual level. To avoid this, alternative methods are required, which often include using other sampling frames (for that particular municipality) or other sampling procedures, such as random-route sampling. However, these alternatives often suffer from coverage errors and/or unknown inclusion probabilities that may lead to biased estimates.

With the present study, we tested an alternative sampling frame available in Germany —namely, the address database from Deutsche Post Direkt (DPD), which is referred to in what follows as the ADB-DPD. DPD is a subsidiary of Deutsche Post AG, the leading postal service provider in Germany. It specializes in address marketing and administers the largest commercial address database in Germany—

1 http://www.gesetze-im-internet.de/englisch_bmg/

2 In the 2014 German General Social Survey (ALLBUS), for instance, only six of the 148 selected municipalities did not provide addresses (Wasmer et al., 2017).

the ADB-DPD—which comprises over 77 million active and 143 million inactive private addresses throughout the country, with its population of roughly 83 million people (Deutsche Post Direkt, 2020). However, as DPD does not provide information on the coverage of its frame and on how addresses are obtained, knowledge on the feasibility of using it for survey sampling is lacking. Thus, we carried out the present study to explore the following two research questions:

- Is the ADB-DPD a viable *alternative* to register-based sampling, in particular for projects that cannot afford to draw a random sample from the population registers?
- Is the ADB-DPD a viable *complement* to register-based sampling in the case of those municipalities that refuse to provide addresses from their population registers?

In the next section, we describe the methods employed in our study and introduce the criteria used to assess the performance of the ADB-DPD. After presenting our results, we reflect on the feasibility of using the ADB-DPD for address-based sampling in Germany.

Data and Methods

The Surveys

We fielded two parallel surveys between November 2019 and February 2020 in the city of Mannheim. Mannheim is located in the federal state of Baden-Wuerttemberg in southwestern Germany and belongs to the prosperous Rhine-Neckar Metropolitan Region, which has a high proportion of specialized jobs, especially in technology and pharmaceuticals. The city is characterized by a large university student population (accounting for approximately 10% of the city's inhabitants, compared with a share of 3% of the population of Germany as a whole) as well as by ethnic diversity, with roughly 45% of the residents having a migration background (City of Mannheim, n.d.).

For the first survey (referred to in what follows as the "register survey"), we drew 3,000 individual addresses from the city's population register; for the second survey (the "DPD survey"), the same number of addresses was drawn from the ADB-DPD. For both surveys, simple random samples of all persons aged 18 years and older were drawn from each sampling frame.[3] Before we contacted our target persons for the first time, our field service provider checked the addresses for

3 Like the population registers, the ADB-DPD also allows for stratifying the sample according to age and sex. However, we have no information on the validity of the sociodemographic information in the ADB-DPD and, more importantly, on how this information is collected.

Table 1 Experimental design and groups

			Incentive		
Group #	Sampling frame	Mode sequence	1st contact	2nd contact	Group size
1	Population Register	Concurrent	0 €	0 €	375
2	Population Register	Concurrent	1 €	0 €	375
3	Population Register	Concurrent	2 €	0 €	375
4	Population Register	Concurrent	0 €	2 €	375
5	Population Register	Sequential	0 €	0 €	375
6	Population Register	Sequential	1 €	0 €	375
7	Population Register	Sequential	2 €	0 €	375
8	Population Register	Sequential	0 €	2 €	375
9	ADB-DPD	Concurrent	0 €	0 €	375
10	ADB-DPD	Concurrent	1 €	0 €	375
11	ADB-DPD	Concurrent	2 €	0 €	375
12	ADB-DPD	Concurrent	0 €	2 €	375
13	ADB-DPD	Sequential	0 €	0 €	375
14	ADB-DPD	Sequential	1 €	0 €	375
15	ADB-DPD	Sequential	2 €	0 €	375
16	ADB-DPD	Sequential	0 €	2 €	375

duplicates, and omitted 20 cases that were included in the address files from both sources.

Apart from differences in the cover letter (explained in more detail in the next section), both surveys were identical in terms of recruitment, field time, and questionnaire. We carried out a self-administered mixed-mode survey (web and postal mail), and randomly allocated target persons from each sampling frame to one of eight experimental groups, representing combinations of mode-choice sequence (sequential vs. concurrent) and small prepaid monetary incentives offered on the first or second contact. In both surveys, all groups had the same sample size (for an overview, see Table 1). However, as these experiments go beyond the scope of the present paper, we shall not report their findings here.

The survey was introduced to the target persons as a community survey dealing with the quality of life in Mannheim. However, it also included other topics such as political attitudes, personality traits, and the perception of surveys. The questionnaire took roughly 30 minutes to complete. Target persons received a

stamped addressed envelope in which to return the paper questionnaire. To account for the rising share of respondents answering surveys via mobile devices, the web-based questionnaire was optimized for smartphones.

Operationalization

To assess the performance of the ADB-DPD, we drew on the following criteria:

Feasibility: This criterion included measures that are relevant for survey planning and budgeting, as well as requirements for and restrictions on survey implementation and fieldwork as a consequence of relying on the ADB-DPD for sampling purposes. More specifically, these measures were: (a) the speed of address provision, (b) the costs of drawing a sample, (c) the feasibility of survey modes, and (d) restrictions on fieldwork.

Up-to-dateness: Here, we added up the proportions of target persons for whom (a) the invitation letter could not be delivered by the postal service, due to an incorrect address, death, or relocation; (b) the invitation letter could be delivered, but we subsequently found out (e.g., from relatives) that the target person had died or moved away (and the postal service was not aware of this).

Survey response: Participation in the two surveys was measured using the American Association of Public Opinion Research (AAPOR) Response Rate 2 (RR2; AAPOR, 2016).

Sample composition: Comparing the composition of the net samples for both surveys indicates whether the ADB-DPD suffers from coverage problems by systematically excluding or underrepresenting certain parts of the target population. To this end, we first analyzed the composition of the net samples in terms of sex and age. For these two variables, we also had information on their distribution in the target population (City of Mannheim, 2020). For the purpose of comparison with the official data, we recoded age into five groups (18–24, 25–29, 30–64, 65–79, 80 and older). Sex was coded dichotomously (1 = female, 0 = male).

Second, we compared the two samples in terms of migration background, formal education, marital status, employment status, place of birth, and second residence. To measure migration background, each respondent was asked whether they, their mother, or their father had immigrated to the current territory of the Federal Republic of Germany after 1955. We created a dummy variable indicating whether this was the case for any of the three persons. Formal education was also coded as a dummy variable, with the value 1 indicating that the respondent had a higher education entrance qualification (*Abitur* or *Fachabitur*). We created similar dummies indicating whether the respondent reported that they were married, employed, had lived in Mannheim since birth, and/or had a second residence in Germany. For these variables, we compared the two net samples (with the register survey as a reference), as no official statistics are available for Mannheim.

In addition, we compared the net samples with respect to a set of substantive variables. These variables, which covered a wide range of topics commonly asked in general social science surveys, were:

- *self-reported political interest* measured on a 7-point scale ranging from 1 (*not at all*) to 7 (*very strong*)
- *external political efficacy* operationalized as agreement with the statement "Politicians care about what people like me think," measured on a 7-point scale ranging from 1 (*do not agree at all*) to 7 (*completely agree*)
- *Abstention from voting* measured with the question whether the respondent would vote if there were a federal election on the following Sunday (1 = would abstain, 0 = would vote, missing value = not entitled to vote).
- *Intention to vote Conservative* measured with the voting intention question; a dummy variable indicates whether the respondent stated that they would vote for the Christian Democratic Union (CDU) party (1 = vote for CDU; 0 = vote for another party; missing value = would abstain or is not entitled to vote).
- *Interpersonal trust* measured with the question: "Generally speaking, do you think that most people can be trusted or that you cannot be too careful in dealing with other people?" (1 = most people can be trusted, 0 = cannot be too careful).
- *Institutional trust* measured with three items—trust in the federal government, trust in the media, and trust in political parties—with each item measured on a 7-point scale ranging from 1 (*do not trust at all*) to (*trust completely*).
- *Big Five personality traits* measured with the BFI-10 short scale (Rammstedt & John, 2007). Respondents answered the 10 items on a 7-point scale ranging from 1 (*disagree strongly*) to 7 (*agree strongly*). For each Big Five dimension (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism), we computed the mean of the two items that measured it.

Results

Feasibility

Speed of Address Provision

DPD was able to provide the requested addresses within five working days. Due to the centralized setup of the ADB-DPD, addresses from a larger set of municipalities (throughout Germany) would probably have been provided in a similar timely fashion. To draw a nationwide sample from the population registers, German survey managers usually estimate that it will take up to three months to receive addresses from all selected municipalities.

Costs of Drawing a Sample

For up to 10,000 individual addresses, DPD charged 84 euros per 1,000 addresses (as of October 2019), irrespective of whether these addresses shared the same place of residence or were spread all over Germany. However, the minimum contract value was 1,000 euros per delivery. As our survey used the same addresses twice (i.e., for the initial contact and for a reminder), DPD regarded this as two deliveries and charged an additional fee (of 250 euros, as of October 2019) for data storage and for checking the up-to-dateness of the addresses prior to the second delivery. Thus, for a nationwide survey that fully relied on the ADB-DPD for address-based sampling with a gross sample of 10,000 target persons and three scheduled contact attempts, the cost of sampling would have amounted to 3,500 euros (3 x the minimum contract value, plus the additional fee for the second and third deliveries/contacts) as of October 2019. Based on our own experiences and on an estimate from an experienced survey manager at a German fieldwork agency, the costs for drawing a sample of similar size from the local population registers would have amounted to approximately 30,000 euros for acquiring the addresses alone. Consequently, using the ADB-DPD to draw a sample of residents from all over Germany is considerably cheaper and the data are provided much faster compared with the population registers. However, if researchers were to use the ADB-DPD as a substitute for individual, noncompliant municipalities, the costs for drawing a random sample of the residents of these municipalities would also amount to 3,500 euros (for three contacts), even if just 500 addresses from five municipalities were needed. This is due to the fixed minimum contract value per delivery. Thus, for small substitute samples of a limited number of municipalities, the relative costs per case are considerably higher.

Feasibility of Survey Modes

In terms of feasibility, the ADB-DPD had a major downside: We did not receive the addresses directly but had to engage a print service provider that, in turn, concluded a contract with DPD for receiving and processing the addresses. Although this did not negatively affect our fieldwork management,[4] it considerably limits the feasibility of using the database, as sampling via the ADB-DPD is not feasible for face-to-face surveys. Rather, the use of the ADB-DPD is limited to self-administered surveys with postal mail invitations. Moreover, it was not possible to further reduce survey costs by engaging our own staff (e.g., student assistants) to prepare the invitation and reminder letters, as all operative work with survey materials had to be done by the contracted print service provider.

4 To administer responses, each address was assigned an 8-digit string code by our print service provider. These codes were printed on the paper questionnaire and were also used as passwords to access the online questionnaire.

Restrictions on Fieldwork

When purchasing addresses, researchers are bound by DPD's terms and conditions. This has two important implications. The first relates to the content of the cover letter. DPD required that a rather lengthy text passage (770 characters including spaces) on data protection issues be prominently placed on the first page of our cover letter. This text was pre-written, its content could not be changed, and it was framed for the most common purpose for which the ADB-DPD is used, namely, advertising. This prescribed text could not instead be integrated into accompanying material such as the data privacy sheet. Its inclusion in the cover letter not only reduced the space available for providing relevant information about the survey. It also caused us to fear that target persons would doubt the integrity of the survey, especially because the (standardized) text suggested that the letter served advertising purposes. Finally, we had to submit all materials to DPD (e.g., the cover letter, the data privacy sheet, the questionnaire) in advance to obtain approval to mail them to our target persons. However, this did not prolong the preparation time before fieldwork, as we submitted our materials and received the approval on the same working day.

The second restriction on fieldwork was that we were limited in terms of when we could recontact our target persons (by sending reminders). More precisely, it was not possible to contact the same addresses again until four weeks after the initial contact. In the meantime, DPD checked whether any recipients had requested a deletion (using the procedure outlined in the aforementioned prescribed text passage). For survey managers, this means that a subsequent contact attempt cannot be carried out until four to five weeks after the previous one.[5] This increases the duration of fieldwork and likely diminishes the effect of incentives and other motivational material provided on the first contact attempt.

Up-To-Dateness

The proportion of target persons for whom the invitation letter could either (a) not be delivered due to an incorrect address, death, or relocation, or (b) be delivered, but we subsequently found out (e.g., from relatives) that the target person had died or moved away was 10.1% for the register survey and slightly lower (9.3%) for the DPD survey. According to a two-sample test for the equality of proportions, this difference did not fall below conventional thresholds for statistical significance ($t = 1.01$, $p > .05$).

5 As we used only one reminder in our study, we did not investigate whether a third delivery of the same addresses would have been possible. Thus, we suggest that researchers who plan to purchase addresses from DPD and to contact their target persons more than twice should clarify this matter with the company in advance.

Survey Response

The overall response rate across the two surveys was 24.3% (RR2; AAPOR, 2016). The response rate in the DPD survey (26.1%) was significantly higher than in the register survey (22.3%; $t = 3.26$, $p < .01$).

Sample Composition

As shown in Table 2, the share of females in the target population (official data) was 49.7% (as of 2019). In the register survey, this share was slightly lower (47.9%), but still considerably higher than in the DPD survey, where only 39.0% of all respondents were female. Based on a one-sample test for the equality of proportions, women were significantly underrepresented in the DPD survey compared with the official data ($z = -5.65$, $p < .001$).

With regard to age, the distribution of respondents in the register survey was not significantly different from that in the target population. In the DPD survey, however, the age distribution was heavily skewed toward older people: The share of people aged 65–79 years (27.7%) was almost twice as high as in the target population. By contrast, the share of respondents aged 18–24 and 25–29 in the net sample of the DPD survey was only 3.3%, whereas in the target population 21.4% belonged to these age groups. This indicates that older people are considerably overrepresented in the ADB-DPD.

Regarding further sociodemographic variables (see Table 3, Panel A), the share of respondents with a migration background was rather low in both surveys. This finding is a common phenomenon, especially in self-administered surveys (Salentin, 2014). In the DPD survey, the share of respondents with a migration background was 11.7%, which was significantly lower than that in the register survey (21.7%; $t = 4.58$, $p < .001$).

To account for differences in sample composition between the two surveys with regard to age and sex, we estimated additional multivariate regression models based on the pooled dataset with age and sex as covariates (see Table 3, Panel B). We then calculated the predicted proportions of all sociodemographic variables for the two values of our sample variable (i.e., for our two surveys with the different sampling frames), holding age and sex at the grand mean of the pooled dataset. For the share of respondents with a migration background, the multivariate model revealed a reduced but still significant difference between the two surveys. With regard to formal education, the register survey showed a significantly higher share of respondents with a higher education entrance qualification (61.5% in the register survey vs. 51.7% in the DPD survey; $t = 3.44$, $p < .001$). In the multivariate model, however, these differences disappeared (and were even reversed), suggesting that differences in the sample composition with respect to age and sex were responsible

Table 2 Sex and age distribution in the target population (official data) and in the register and DPD surveys

Demographic variables	Official data (%)	Register survey (%)		DPD survey (%)		*p* (Register vs. DPD)
Female	49.7	47.9	ns	39.0	***	**
Age						
18–24	11.6	13.6	ns	1.3	***	***
25–29	9.8	10.9	ns	2.0	***	***
30–64	56.9	54.5	ns	62.4	**	**
65–79	14.8	15.9	ns	27.7	***	***
80 and over	6.9	5.2	ns	6.6	ns	ns

Note: Differences in demographic variables between the register survey/DPD survey and official data were tested using (two-sided) one-sample tests for the equality of proportions. Differences in demographic variables between the two surveys were tested using a (two-sided) two-sample test for the equality of proportions. ns = not significant.

* $p < .05$, ** $p < .01$, *** $p < .001$.

Table 3 Predicted proportions of other demographic variables in the register survey and the DPD survey with and without age and sex as covariates

	Panel A: Predicted proportions excluding age and sex as covariates			Panel B: Predicted proportions including age and sex as covariates		
Demographic variables	Register survey (%)	DPD survey (%)	*p*	Register survey (%)	DPD survey (%)	*p*
Migration background	21.7	11.7	***	19.4	13.1	**
High level of formal education	61.5	51.7	***	54.4	57.5	ns
Living in Mannheim since birth	32.4	51.0	***	32.6	50.8	***
Married	46.5	66.2	***	51.8	61.8	***
Second residence in Germany	6.6	8.1	ns	5.9	9.0	ns
Employed	58.9	57.1	ns	50.2	63.0	ns

Note: All estimates are predicted proportions based on logistic regression models with the respective demographic variable as dependent variable. The models in Panel A included the sample as the only independent variable and the models in Panel B included the sample, age, and sex as independent variables. The results in the *p*-columns refer to the *p*-values of the regression coefficients of the sample variable in the bivariate regression models (Panel A) and the multivariate regression models (Panel B), respectively. ns = not significant.

* $p < .05$, ** $p < .01$, *** $p < .001$.

for the higher share of highly educated respondents in the register survey. Yet, even after including sex and age as covariates, respondents in the DPD survey had a significantly higher probability of being married, of being employed, and of having lived in Mannheim since birth. For these variables, the net sample of the DPD survey significantly differed from that of the register survey, which we used as our reference distribution.

Finally, Table 4 shows differences between the register survey and the DPD survey with respect to substantive variables. Similarly to above, we report the predicted means and proportions for these variables and for the two surveys based on a bivariate regression model (Panel A) and on a multivariate model including age and sex as covariates (Panel B). Without accounting for the different sample composition, we found that responses in the DPD survey showed higher levels of political interest and lower levels of external efficacy. When age and sex were included as covariates, differences in self-reported political interest disappeared. However, in this scenario, the gap in external political efficacy remained and differences in institutional trust were even more pronounced. Although we found significant differences for some variables either in the bivariate or in the multivariate model, we would like to note that we consider the magnitude of these differences to be small.

Table 4 Predicted means/proportions of substantive items in the register survey and the DPD survey with and without age and sex as covariates

	Panel A: Predicted means/ proportions excluding age and sex as covariates			Panel B: Predicted means/ proportions including age and sex as covariates		
Substantive items	Register survey	DPD survey	*p*	Register survey	DPD survey	*p*
Self-reported political interest 1 = not at all; 7 = very strong	4.98	5.22	**	5.08	5.13	ns
Politicians care about what people like me think (external political efficacy) 1 = do not agree at all; 7 = completely agree	3.01	2.69	***	2.98	2.72	**
Abstention from voting	7.6%	5.6%	ns	8.0%	5.3%	ns
Intention to vote Conservative	19.3%	24.5%	*	22.2%	22.2%	ns
Interpersonal trust 0 = one cannot be too careful; 1 = most people can be trusted	56.2%	56.4%	ns	54.3%	58.0%	ns

Table 4 continued

Substantive items	Panel A: Predicted means/ proportions excluding age and sex as covariates			Panel B: Predicted means/ proportions including age and sex as covariates		
	Register survey	DPD survey	*p*	Register survey	DPD survey	*p*
Institutional trust						
Trust in the federal government 1= do not trust at all; 7= trust completely	4.07	3.96	ns	4.12	3.92	*
Trust in the media 1= do not trust at all; 7= trust completely	3.59	3.65	ns	3.66	3.58	ns
Trust in political parties 1= do not trust at all; 7= trust completely	3.27	3.14	ns	3.29	3.12	*
Personality Traits (Big Five)						
Openness to Experience 1= very low; 7= very high	4.75	4.59	ns	4.72	4.62	ns
Conscientiousness 1 = very low; 7 = very high	5.38	5.48	ns	5.44	5.43	ns
Extraversion 1 = very low; 7= very high	4.49	4.37	ns	4.43	4.42	ns
Agreeableness 1 = very low; 7 = very high	4.31	4.35	ns	4.32	4.34	ns
Neuroticism 1 = very low; 7 = very high	3.39	3.35	ns	3.33	3.41	ns

Note. All estimates are predicted means/proportions based on linear/logistic regression models (logistic regression models were estimated for "Abstention from voting," "Intention to vote Conservative," and "Interpersonal trust") with the respective substantive item as the dependent variable. The models in Panel A included the sample as the only independent variable; the models in Panel B included the sample, age, and sex as independent variables. The results in the *p*-columns refer to the *p*-values of the regression coefficients of the sample variable in the bivariate regression models (Panel A) and the multivariate regression models (Panel B), respectively. ns = not significant.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Discussion

With the present study, we tested the feasibility of using the ADB-DPD as a sampling frame for scientific surveys. In many countries, commercial enterprises—in particular postal service providers—have specialized in collecting and selling addresses of their residential populations. In Germany, DPD administers the largest database of this kind, which is used mainly for marketing purposes. How-

ever, this database, referred to in the present study as the ADB-DPD, can also serve as a frame for address-based sampling for nationwide surveys, as addresses can be randomly drawn from all current entries throughout the country.

The starting point for our study was the assumption that sampling from the official population registers is the gold standard for drawing representative samples of the German residential population. Nevertheless, we assumed that the ADB-DPD might be attractive as a complement to register-based sampling or even as an alternative sampling frame in two survey scenarios. In the first scenario, survey projects have limited resources in terms of time and funds but need to include target persons with a high regional diversification, or even aim to carry out a nationwide survey. In the second scenario, survey projects have sufficient resources to comply with the gold standard for sampling—that is, register-based sampling—but are faced with the problem that some municipalities refuse to provide the requested addresses. In this scenario (and particularly in cases where these municipalities cannot be easily substituted), the ADB-DPD might complement the population registers for address-based sampling. For these potential scopes of application, our study aimed to provide a first assessment of the feasibility of using the ADB-DPD as a frame for address-based sampling, especially in light of the fact that DPD provides no information on the coverage of this database.

A key limitation of the ADB-DPD is that it is meaningfully usable as a sampling frame only for self-administered surveys. This is because researchers do not receive the addresses directly but can communicate with their target persons only via a print service provider. Of course, this considerably limits the scope of application. However, despite this important restriction, our results were rather promising in terms of feasibility. Due to the centralized setup of the ADB-DPD, addresses from all over Germany can be randomly drawn and provided by DPD within five working days and at low costs. The ADB-DPD appears to be especially cost-effective for nationwide survey projects that consider relying fully on this sampling frame. In contrast, when a random sample is needed for only a small number of noncompliant municipalities that cannot be substituted (which we consider the most important case for scientific surveying), sampling via the ADB-DPD is rather expensive, due mainly to the fixed minimum contract value. Moreover, due to DPD's insistence on trade secrecy, the generation of the sampling frame is essentially a black box. As a result, coverage error is not computable.

Having said that, the ADB-DPD performed well in terms of survey response and up-to-dateness. With regard to survey response, the DPD survey even yielded a significantly higher response rate than the register survey. This outcome is surprising, given that DPD required that more data privacy information be included in the cover letter. However, the higher response rate might be related to the fact that the ADB-DPD (over)covers demographic groups with higher response propensities.

Moreover, the rate of undeliverable invitation letters was slightly lower in the DPD survey.

Regarding sample composition, we found a substantial underrepresentation of women and young people in the DPD survey. Given that we drew a simple random sample from both frames and administered an almost identical survey in terms of recruitment, field time, and questionnaire, this may indicate that the ADB-DPD suffers from coverage error for these demographic groups. Similarly, when comparing the net samples, the DPD survey showed a remarkably higher proportion of people who had been living in Mannheim since birth, but a considerably lower proportion of respondents with a migration background. All this suggests that the ADB-DPD covers a larger share of the less mobile segment of the population (older people, non-migrants, people who rarely change their place of residence). This reasoning might also explain the higher response rate in the DPD survey, as the less mobile segment of the population might have been particularly attracted by the local focus of the survey. Turning to substantive variables, we found some differences between the net samples of our two surveys—before and after including age and sex as covariates in our statistical models. However, these appeared to be of low magnitude.

Overall, we would like to caution that using the ADB-DPD comes with its own challenges and uncertainties, especially regarding the way in which addresses are obtained and how well the database covers the residential population. Regarding the feasibility of using the ADB-DPD for sampling purposes, this seems to be an option for smaller research projects that cannot afford to draw a nationwide sample from the population registers and/or that operate on a tight schedule. This holds true especially when the estimation of valid parameters for the residential population is not a high priority, but rather the focus is on experimentation.

Our study can be extended in various ways. First, we encourage future research to compare different alternatives to the gold standard (i.e., register-based sampling) and to detail the pros and cons of each alternative. If a noncompliant municipality cannot be substituted, there will be a trade-off between having no addresses for this regional unit (i.e., risking systematic undercoverage) and drawing on alternative frames with their own challenges and possible errors. As was the case in our study, we suggest investigating this question by fielding similar surveys in parallel in the same regional units using different sampling methods.

Second, our study focused on the municipality of Mannheim. Although our findings might hold true for other cities or regions in Germany, replication studies are required. Such studies should also aim to cover different topics. Our survey was framed as a community survey that covered a diverse set of substantive social science questions. However, it would be interesting to see how well the ADB-DPD performs for specific topics (e.g., election studies) or for questionnaires without a topic of local relevance.

Third, given the country-specific nature of the ADB-DPD, we refrain from generalizing our findings to other countries. However, we provide empirical evidence on how sampling approaches and their impact on survey outcomes differ. In our study, we show these differences for a presumably more cost-efficient alternative to the established gold standard—register-based sampling—in Germany. We would thus welcome further research that conducts similar studies in other countries. Such studies could aim to investigate how other country-specific commercial address services perform, and whether they are an alternative to or can complement register-based sampling.

References

Amaya, A., Zimmer, S., Morton, K., & Harter, R. (2021). Does undercoverage on the U.S. address-based sampling frame translate to coverage bias? *Sociological Methods & Research*, *50*(2), 812–836. doi:10.1177/0049124118782539

American Association of Public Opinion Research (AAPOR). (2016). *Standard definitions. Final dispositions of case codes and outcome rates for surveys.* Retrieved October 19, 2021, from the AAPOR website: www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx

City of Mannheim (n.d.). *Einwohner mit Migrationshintergrund*. Retrieved October 19, 2021, from the website of the city of Mannheim: www.mannheim.de/de/stadt-gestalten/daten-und-fakten/bevoelkerung/einwohner-mit-migrationshintergrund

City of Mannheim (2020). *Einwohnerbestand 2019 in kleinräumiger Gliederung* (Statistische Daten Mannheim, No. 1/2020). Retrieved October 19, 2021, from the website of the city of Mannheim: www.mannheim.de/de/stadt-gestalten/daten-und-fakten/bevoelkerung

Deutsche Post Direkt. (2020). *Zusatzinformation datenschutzkonforme Adresslösungen*. Retrieved October 19, 2021, from the website of Deutsche Post Direkt: www.deutschepost.de/de/d/deutsche-post-direkt/deutsche-post-direkt-datenschutz.html

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed mode surveys: The tailored design method* (4th ed.). New York: John Wiley & Sons.

European Social Survey. (2017). *ESS Round 8 (2016/2017) technical report*. Retrieved October 19, 2021, from the ESS website: www.europeansocialsurvey.org/data/download.html?r=8

Federal Ministry of the Interior and Community (BMI). (n.d.). *Meldewesen*. Retrieved October 19, 2021, from the website of the BMI: www.bmi.bund.de/DE/themen/moderne-verwaltung/verwaltungsrecht/meldewesen/meldewesen-node.html

Harter, R., Battaglia, M. P., Buskirk, T. D., Dillman, D. A., English, N., Fahimi, M., … Zukerberg, A.L. (2016). *AAPOR report: Address-based sampling.* Retrieved October 19, 2021, from the AAPOR website: www.aapor.org/Education-Resources/Reports.aspx

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). Comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for general population surveys, *Public Opinion Quarterly*, 72(1), 6–27, doi:10.1093/poq/nfn003

Lohr, S. L. (2009). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury Press.

Maineri, A. M., Scherpenzeel, A., Bristle, J., Pflüger, S-M., Butt, S., Zins, S., Emery, T., & Luijkx, R. (2017). *Evaluating the quality of sampling frames used in European cross-national surveys.* (Deliverable 2.2 of the SERISS project). Retrieved October 19, 2021, from the SERISS website: www.seriss.eu/resources/deliverables/

Poulain, M., & Herm, A. (2013). Central population registers as a source of demographic statistics in Europe. *Population*, 68(2), 183–212. doi:10.3917/popu.1302.0215

Rammstedt, B., John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1), 203–212. doi:10.1016/j.jrp.2006.02.001

Salentin, K. (2014). Sampling the ethnic minority population in Germany. The background to "migration background." *methods, data, analyses*, *8*(1), 28. doi:10.12758/mda.2014.002

Scherpenzeel, A., Maineri, A. M., Bristle, J., Pflüger, S-M., Mindarova, I., Butt, S., Zins, S., Emery, T., & Luijkx, R. (2017). *Report on the use of sampling frames in European studies.* (Deliverable 2.1 of the SERISS project). Retrieved October 19, 2021, from the SERISS website: www.seriss.eu/resources/deliverables/

Wasmer, M., Blohm, M., Walter, J., Jutz, R., & Scholz, E. (2017). *Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2014* (GESIS Papers No. 2017/20). Mannheim: GESIS – Leibniz Institute for the Social Sciences. Retrieved October 19, 2021, from the GESIS website: www.gesis.org/allbus/inhalte-suche/methodenberichte_

methods, data, analyses | Vol. 17(1), 2023, pp. 47-70 DOI: 10.12758/mda.2022.08

How to Optimize Online Mixed-Device Surveys: The Effects of a Messenger Survey, Answer Scales, Devices and Personal Characteristics

Caroline Marjanne Menken & Vera Toepoel
University of Utrecht

Abstract

The goal of this research was to determine the best way to present mixed-device surveys. We investigate the effect of survey method (messenger versus regular survey), answer scale, device used, and personal characteristics such as gender, age and education on break-off rate, substantive answers, completion time and respondents' evaluation of the survey. Our research does not suggest that a messenger survey affects mixed-device surveys positively. Further research is necessary to investigate how to optimally present mixed-device surveys in order to increase participation and data quality.

Keywords: Mixed-device survey, messenger, answer scale, online survey, mobile friendly

Online surveys are often used by researchers (Schlosser & Mays, 2018; Zhang, Kuchinke, Woud, Velten & Margraf, 2017). A traditional online survey is designed to be completed on a computer (Schlosser & Mays, 2018). However, since the use of mobile devices has grown, online surveys are also being completed on other devices such as mobile phones and tablets (De Bruijne & Wijnant, 2013; Mavletova, 2013; Millar & Dillman, 2012). Surveys that are being completed on different devices are called mixed-device surveys (Toepoel & Lugtig, 2015). What is the best way to present these mixed-device surveys?

Previous research has shown that there are several differences between devices, when it comes to response behavior (Couper & Peterson, 2017; Schlosser & Mays, 2018). For instance, the screen size of mobile phones is smaller than the screen size of computers and tablets (Schlosser & Mays, 2018), making it more difficult to answer questions on a mobile phone (De Bruijne & Wijnant, 2014; Mavletova, 2013). Especially close-ended questions with many answer options are not desirable on a small screen, because some answer options fall off screen (Couper & Peterson, 2017) and respondents need to scroll to see all the answer options. Alternatives are open-ended questions or close-ended questions with few answer options. However, research shows open-ended questions take a lot of effort to answer and result in higher (item) nonresponse (Couper & Peterson, 2017; De Bruijne & Wijnant, 2014; Couper & Peterson, 2017; Mavletova, 2013; Schlosser & Mays, 2018). Often, researchers make online surveys more suitable to be completed on mobile phones, for instance by making the design of the survey responsive. With a responsive design, the layout of the survey adapts to the device being used (Antoun, Katz, Argueta & Wang, 2018; De Bruijne & Wijnant, 2014; Mavletova, 2013). However, although a responsive design makes it easier for respondents to complete online surveys on mobile phones, it might still not be optimal for data quality (Antoun et al., 2018). A solution to increase the data quality of online surveys could be to make the survey more interactive by adding a conversational element (Kim, Lee & Gweon, 2019). Since mobile phones are mostly used for short messaging, a Whatsapp-type messenger survey could be a way to make an online survey more suitable for mobile phones. A research messenger survey is more similar to text messaging and adds a conversational element to the online survey.

In this study, we randomly assigned respondents in the American Amazon MTurk Panel to a regular responsive online survey design and a messenger survey. In addition, we randomly assigned panel members to a closed-ended answer scale with many answer options, a closed-ended answer scale with few answer options (that would fit small screens of mobile phones), and an open-ended answer scale to investigate the effect of survey method and type of response format. We also

Direct correspondence to
Caroline Menken, University of Utrecht
E-mail: caroline-menken@hotmail.com

investigate the effect of the device used to complete the survey and the effect of personal characteristics such as age, gender and education. We compare break-off rate, substantive answers, completion time and respondents' evaluation of the survey to provide suggestions on how to optimally design mixed-device surveys. We conducted exploratory research.

Theoretical Background

Online Surveys

Since Internet has become more and more important in daily life, the use of online surveys has grown (Schlosser & Mays, 2018; Solomon, 2000; Zhang, Kuchinke, Woud, Velten & Margraf, 2017). Online surveys have a number of advantages and disadvantages. One of the advantages is that there is no need for interviewers (Tourangeau, Maitland, Rivero, Sun, Williams & Yan, 2017). Therefore, online surveys are anonymous, which reduces socially desirable responding (Tourangeau et al., 2017). Furthermore, online survey research takes less time and the costs of online survey research are typically low (Couper & Miller, 2008; Solomon, 2000; Wright, 2005; Zhang et al., 2017). For respondents it takes less effort to participate in a survey since respondents can complete online surveys in their own time and space (Solomon, 2000; Wright, 2005; Zhang et al., 2017). However, the absence of an interviewer also has disadvantages (Bowling, 2005; De Leeuw, 2008; Kim, Lee & Gweon, 2019). An interviewer can give the respondent additional instructions and can clarify the questions when needed (Bowling, 2005; De Leeuw, 2008; Harris & Brown, 2010). Besides, the presence of an interviewer leads to a higher response rate and a higher completion rate since an interviewer can convince and motivate reluctant respondents to participate and to finish the survey (De Leeuw, 2008; Heiervang & Goodman, 2011; Kim, Lee & Green, 2019). If an interviewer is present, a survey is more similar to a conversation (Bowling, 2005).

Mixed Devices

A traditional online survey is designed to be completed on a computer (Cunningham, Neighbors, Bertolet & Hendershot, 2013; Schlosser & Mays, 2018). However, nowadays online surveys are also being completed on other devices such as mobile phones and tablets (De Bruijne & Wijnant, 2013; Mavletova, 2013; Millar & Dillman, 2012). There are several differences between the different devices (Couper & Peterson, 2017; Schlosser & Mays, 2018). First of all, the screen of a mobile phone is much smaller than the screen of a computer (Maslovskaya, Smith & Durrant, 2020; Schlosser & Mays, 2018; Stapleton, 2013). So, it takes longer to complete a

survey on a mobile phone than on a computer. The time to complete a survey on a tablet is typically in between (Couper & Peterson, 2017). The Internet connection on mobile phones is often slower than on computers, also leading to a higher completion time (Couper & Peterson, 2017; Schlosser & Mays, 2018). Furthermore, respondents who complete a survey on a mobile phone or tablet are more likely to do this away from home, and could therefore be distracted (Couper & Peterson, 2017; Maslovskaya, Smith & Durrant, 2020; Schlosser & Mays, 2018). The higher completion time on mobile phones often leads to a higher break-off rate (Couper & Peterson, 2017; Cunningham et al., 2013; Mavletova, 2013; Schlosser & Mays, 2018). Respondents generally experience a higher difficulty to complete an online survey on a mobile phone (De Bruijne & Wijnant, 2013). Research shows that the response-rate of traditional, non-optimized online surveys on mobile phones is very low (Mavletova, 2013).

Additionally, the usage of mobile phones is different than the usage of computers and tablets. Mobile phones are mostly being used to communicate through messaging apps, especially Whatsapp (O'Hara, Massimi, Harper, Rubens & Morris, 2014). Whatsapp is being used for sharing information, images and videos and for ongoing conversations by sending short messages (Ahad & Lim, 2014; O'Hara et. al., 2014). Mobile phones are used for more casual conversations, whereas computers are generally being used for more formal communication (O'Hara et. al., 2014).

Mobile Friendly Survey

Antoun and others (2018) created general guidelines to alter the design of a survey to make it suitable for mobile phones. First of all, it should be easy for respondents to read the questions and answer options. The font size should be large enough and answer options should be large enough to be easily selected by respondents with touch screen. In addition, the content of the survey should fit the width of the screen. If not all of the answer options fit on screen, the answer options should be presented vertically not horizontally. Besides, the features of the design should be simple, and respondents should be able to understand how to use them. At last, the design should work on different devices. A way to achieve this is by making the design responsive. With a responsive design the layout of the survey adapts to the device being used (Harb, Kapellari, Luong & Spot, 2011; Hussain & Mkpojiogo, 2015). The layout adapts to suit different screen sizes, larger buttons and texts are provided when using a mobile phone and non-essential elements are being hidden when the screen is small (Harb et al., 2011).

Several researchers have created a mobile friendly design to make a traditional online survey more suitable to be completed on mobile phones. For instance, De Bruijne & Wijnant (2013) made the content of their survey fit the width of the screen and made the font size larger. Moreover, the answer options were made wide

buttons in order to be easily selected with touch screen. The answer options were also presented vertically instead of horizontally. However, respondents who completed the survey with a mobile friendly design still reported a longer completion time than the respondents who completed a traditional online survey on a computer. This implies that completing a survey on a mobile phone still takes more effort and time, even when the survey is adapted to mobile phones. This conclusion is supported by other researchers who investigated the differences between a survey with a mobile friendly design and a traditional online survey (Couper & Peterson, 2017; Mavletova, 2013; Schlosser & Mays, 2018). Antoun and others (2018) also concluded that the guidelines might not be enough to make an online survey optimal for mobile phones. Therefore, more research into how to present an online survey on mobile phones is needed.

Answer Scales

Survey questions can have different types of answer scales. Questions could be open-ended, close-ended with many answer options or close-ended with few answer options. The non-response rate for open-ended questions is higher than for close-ended questions in online surveys (Reja, Manfreda, Hlebec & Vehovar, 2003) because it takes more effort and time for the respondents to answer the open-ended questions (Couper & Peterson, 2017). Close-ended questions have a number of answer options. Respondents tend to choose one of the answer options even if their true answer is not one of the options (Reja et al., 2003). Few answer options can give respondents too little information. The chance that their true answer is not one of the options is higher (Reja et al., 2003). However, too many options make the questions too complicated for respondents. This can lead to not considering all the options (Chung et al., 2010). The order in which the answer options are presented in close-ended questions also affects answers. Respondents are more likely to choose the answer options that are visualized on the screen than the answer options that fall off the screen (De Bruijne & Wijnant, 2013; Mavletova, 2013; Stapleton, 2013). Thus, too many answer options in mixed-device surveys might not be desirable.

In online surveys, the answers on open-ended questions completed on a mobile phone are shorter than the answers on open-ended questions completed on a computer (Mavletova, 2013). According to respondents, it is easier to type an answer on a computer keyboard (Mavletova, 2013). Therefore, it takes more time to answer open-ended questions on mobile phones than on a computer (Couper & Peterson, 2017). In addition, the completion time for close-ended questions with many options is higher than for close-ended questions with fewer answer options on a mobile phone. Research of De Bruijne & Wijnant (2014) has shown that the completion time for an 11-point answer scale was significantly higher than a 5-point or 7-point answer scale on mobile phones. This could also be explained by the fact that

the 5-point answer scale was visible for 99%, the 7-point answer scale for 94% and the 11-point answer scale only for 59%. According to Couper and Peterson (2017), the need to scroll on a mobile phone leads to a higher completion time. Moreover, because more answer options are off screen, the tendency of respondents to choose the visible answer options is especially a problem when the survey is completed on mobile phones (De Bruijne & Wijnant, 2013; Mavletova, 2013; Stapleton, 2013).

Personal Characteristics

Younger people are on average faster in completing a web survey than older people. An explanation could be that the working memory capacity of older people is reduced, which makes the web survey more difficult (Yan & Tourangeau, 2008). However, although younger people complete a web survey faster, the break-off rate is higher for younger people than for older people, possibly due to motivation issues (Peytchev, 2009). Furthermore, younger people use mobile devices, in particular mobile phones, more than older people (De Bruijne & Wijnant, 2013).

Research shows that the response rate of women in online surveys is higher than the response rate of men (Smith, 2008). However, the break-off rate of women is higher than the break-off rate of men (Steinbrecher, Roßmann & Blumenstiel, 2015). Research has shown that men use a smartphone more than women to complete an online survey (De Bruijne & Wijnant, 2013).

In general, the time to complete a web survey is higher for respondents who did not complete high school than for respondents who did. The break-off rate of lower-educated respondents is higher than of higher-educated respondents (Peytchev, 2009; Yan & Tourangeau, 2008). In addition, higher-educated people use mobile phones more often than low-educated people (De Bruijne & Wijnant, 2013).

Innovative Ways to Conduct Surveys

Symon, Cassel and Dickson (2000) argue that there should be more alternative and innovative research methods. Online surveys are often seen by respondents as boring, which leads to reluctance to complete the survey (Dolnicar, Grün & Yanamandram, 2013). For this reason, researchers should look into new ways to present a survey (Dolnicar et al., 2013; Symon et al., 2000). One innovative research method is the gamification of online surveys. Gamification is the use of game design elements in non-game contexts (Harms, Seitz, Wimmer, Kappel & Grechenig, 2015). Gamification leads to more motivation of the respondents, a better user experience and positive feedback of the respondents (Harms et al., 2015). It is a way to make an online survey more interactive and dynamic. This can lead to a higher response rate and lower break-off rate of online surveys (Dolnicar et al., 2013). However, gamification of a survey takes effort (Seaborn & Fels, 2014).

Another way to make an online survey more interactive is a chatbot survey. Kim, Lee and Gweon (2019) used a text-based chatbot in their research to investigate the effect of a conversational element in an online survey. The researchers compared a chatbot survey with a regular web survey. The conversational style of the chatbot survey increased the differentiation in the responses of the respondents, leading to a higher quality of the response data. Furthermore, the respondents who completed the chatbot survey evaluated the survey more positively than respondents who completed the regular web survey (Kim, Lee & Gweon, 2019). The researchers suggest that the conversational style should be casual (Kim, Lee & Gweon, 2019). The research of Kim, Lee and Gweon (2019) did not focus on mixed-device surveys. As mentioned before, mobile phones mostly are used to communicate online by using short messaging apps, like Whatsapp (O'Hara et. al., 2014). A research messenger could be a way to make mixed-device surveys more interactive with a casual conversational style, since a research messenger is similar to short messages that are sent via Whatsapp. No previous research has been conducted to investigate the use of messenger type surveys in mixed-device research.

Methods

Respondents

Respondents could participate by completing an online survey, which was distributed among Amazon Mechanical Turk panel members in the United States of America. Amazon Mechanical Turk is a crowdsourcing marketplace that makes outsourcing of processes and jobs to a distributed workforce, which can perform these tasks virtually, easier. Participation in this research was possible from June to August in 2018. There were 2078 respondents in this research. However, 201 respondents did not complete a single question. In addition, 149 respondents did not complete the survey. The remaining 1728 respondents form the base of our analytic sample.

Survey

The respondents could self-select the device (computer, tablet or mobile phone) to complete the survey. At the beginning of the survey the respondents were randomly assigned to either a regular responsive survey design or a research messenger survey design. Appendix 1 shows images of both survey methods. In addition, respondents were randomly assigned to one of three response option conditions: a condition with open-ended questions; close-ended questions with few answer options; or close-ended questions with many answer options.

The survey consisted of four modules. Module A was about media use, model B about most important issues in the country, module C about politics and module D about sports. The order of the modules was randomized to avoid order effects. The order of the questions within the modules was not randomized. However, in this research only one question of each module is used. Therefore, the question order does not affect results. The respondents were assigned to the same response option condition and the same design in every module. After the four modules, the respondents had to answer questions about their background and their opinion about the survey (evaluation questions).

Analyses

The main goal of this research was to determine the best way to present mixed-device surveys. In order to do this, we investigated if there are differences between three different types of answer scales, two survey methods and the different devices used to complete the survey. Since there were only 104 respondents who completed the survey on a tablet, we decided not to treat tablets as a separate group. Tablets are sometimes grouped with mobile phones, because they are both mobile devices (De Bruijne & Wijnant, 2014). Some researchers group tablets together with computers because tablets are more similar to computers to complete a survey on, for instance both tablets and computers have large screen sizes (Couper & Peterson, 2017). We decided to group tablets with computers, because both devices are in general not used for Whatsapp type of messaging. However, we checked if the results are different when we group tablets with mobile phones. We added the personal characteristics age, gender and education to all analyses as control variables. Despite the low theoretical evidence for the effects of these personal characteristics, we checked if these variables affect the results of this research. To analyze the data we used IBM SPSS Statistics, 26.0.0.

Number of Completes

First, we conducted a simple binary logistic regression analysis to investigate if there is a difference in the proportion of completes between the types of answer scale, the survey methods and the devices used. We added age, gender and education to the analysis as control variables. The analysis has been done to investigate the break-off rate of the respondents. A survey is complete if all the questions of the survey were completed. We used the data of the 1877 respondents who started the survey. However, for fifteen respondents the device used to complete the survey could not be determined, so we did not include these respondents in the analysis. So, we used the data of 1862 respondents. We conjecture that the number of completes of respondents who completed the regular survey on a mobile phone is lower

than the number of completes of respondents who completed the research messenger survey on a mobile phone and the number of completes of respondents who completed the survey on a computer. Furthermore, we conjecture that the number of completes of respondents with open-ended questions is lower.

Substantive Answers

We choose one question per module to investigate if answers differ between the types of answer scale, survey methods, devices, and personal characteristics. We used the first question of the modules media use and sports. From the module about the most important issues in the country we used the only question which had the answer option "other, please write" so that all respondents could give the same answer despite the different types of answer scale. The module about politics did not contain an experiment with answer scales, therefore we did not use a question of this module. The questions we used are in Appendix 1.

The question about media use was "On a typical day, about how much time do you spend watching, reading or listening to news about politics and current affairs?". Respondents in the short scale got five answer options, respondents in the long scale eight. Respondents in the open format had to give their answer in hours and minutes. We dichotomized answer options; values lower than 75 minutes were coded as 0 and values of 75 minutes and higher were recoded as 1 (75 minutes was about the median time). Don't know answers were treated as non-substantive answer options. We dichotomized the answer options to make the answers comparable between the different conditions. The closed-ended questions have nominal time categories as answer options, not single numerical values. These time categories differ between the short scale and long scale, because the long scale has more answer options. The difference in mean could be caused by these different answer scales, therefore we dichotomized the answer options. The question about the most important issues in the country was "Which people or organizations you think have the most influence on the actions of the American government?". The close-ended questions had nine or twelve answer options for the short and long scale, respectively. The question about sports was "What sport or physical activity do you take part in most frequently?". Respondents in the short scale received thirteen answer options and respondents in the long scale twenty. In the analyses, the answers on the questions about the most important issues and sports were adjusted to the closed answer scale with many options. The answers of respondents with the short scale that answered "other" and the answers of respondents with the open answer scale were recoded manually to the long list of the closed answer scale with many options. After that, the questions were dichotomized; answer options that were only options in the long list received the value 1.

We perform simple binary logistic regressions with answer scales, survey methods, devices used, and personal characteristics in order to investigate if respondents

in different conditions give different answers. Interaction effects between answer scale, survey method, device, gender, age and education were also investigated, by adding the interaction terms to the regressions. We used the data of the 1728 respondents who completed the survey. However, of fifteen respondents it could not be determined which device was used to complete the survey. Furthermore, seven respondents choose the answer option "Other" or "Would rather not say" on the question "What is your gender?" and nineteen respondents did not fill in their age or answered with an invalid number. These respondents are not included in the analyses, so we used the data of the remaining 1687 respondents. We conjecture that there is a difference in answers on the questions between the different answer scale conditions. We also conjecture that respondents who completed the survey on a computer choose more often an answer option that is only in the long list than respondents who completed the survey on a mobile phone.

Completion Time

We use simple multiple regression to investigate if there is a difference in completion time between the types of answer scale, the survey methods, the devices used and age, gender and education. We also investigated interaction effects between answer scale, survey method, device, gender, age and education, by adding the interaction terms to the regression. The completion time is the time it took the respondents to complete the survey, so the time between the start and the end of the survey, and it is measured in seconds. We used the log of the completion time, because the distribution of the completion time is right-skewed. The data of the 1687 respondents of whom we had all the data was used. We conjecture that respondents with the research messenger have a higher completion time than respondents with the regular survey (due to the conversational element). Additionally, we conjecture that respondents with open-ended questions have a higher completion time than respondents with close-ended questions. Furthermore, we conjecture that respondents who completed the regular survey on a mobile phone have a higher completion time than respondents who completed the regular survey on a computer.

Evaluation Questions

Finally, we use simple multiple regression analyses to investigate if the answers on three evaluation questions differ between the types of answer sale, survey methods, the devices used and age, gender and education. Interaction effects between answer scale, survey method, device, gender, age and education were also investigated, by adding the interaction terms to the regressions. We used the data of the 1687 respondents of whom all the variables are known. The questions were: "Was it difficult to answer the questions?", "Did you enjoy answering the questions?" and "Was the subject interesting?". Answers on the evaluation questions were investigated in

order to determine the preferences of respondents with regard to presentation of mixed-device surveys. We conjecture that respondents who completed the research messenger evaluated the survey more positively than respondents who completed the regular survey, in particular respondents who completed the survey on a mobile phone. In addition, we conjecture that respondents in the open answer scale condition evaluated the survey more negatively than respondents in a closed answer scale condition.

Results

Descriptives

Table 1 shows the number of respondents per device, survey method, type of answer scale, gender and education. For fifteen respondents the device used to complete the survey could not be determined. Seven respondents choose the answer option "Other" or "Would rather not say" on the question "What is your gender?". Moreover, table 1 shows the minimum and maximum age of the respondents, the mean age, the mean completion time and the recoded binary variables from the different modules.

Table 1 Descriptive statistics

	n	%
Mobile phone (0)	538	31.4
PC (1)	1071	62.5
Tablet (1)	104	6.1
Total	1713	
Research messenger (RM) (0)	871	50.4
Regular survey (1)	857	49.6
Total	1728	
Open	580	33.6
Closed with few options	574	33.2
Closed with many options	574	33.2
Total	1728	
Female (0)	1157	67.2
Male (1)	564	32.8
Total	1721	

	n	%
Less than high school/ high school graduate (0)	1015	58.7
Some college/ college graduate (1)	713	41.3
Total	1728	
	min	max
Age in Years	18	84
	M	SD
Age in Years	34.87	10.119
Completion time in seconds	764.39	451.853
%	0	1
Question Media use *1:75 minutes or more*	51.6	48.4
Question Important issue *1: answers only in the long list*	82.8	17.2
Question Sports *1: answers only in the long list*	91.1	8.9

Completes

A binary logistic regression analysis is conducted to predict the proportion of completes. Table 2 shows that device, survey method, type of answer scale, age, gender and education do not significantly predict if the respondent completed the survey. The regression model is also not significant. However, as expected a lower proportion of respondents in the open answer condition completed the survey compared to respondents in a closed-ended condition. Furthermore, a higher proportion of respondents in the regular survey condition than in the research messenger condition completed the survey.

Table 2 Results binary logistic regression analysis predicting completes

	B	Exp(B)
Device (0: mobile phone, 1: tablet/PC)	.038	1.038
Survey method (0: RM, 1: regular survey)	-.495	.609
Open	-.820	.440
Closed few	-.350	.705
Closed many (ref.)		
Age	-.004	.996
Gender (0: female, 1: male)	.285	1.330
Education (0: ≤ high school graduate, 1: ≥ college)	-.437	.646
Nagelkerke R^2	.026	
n	1862	

Substantive Answers

We conducted binary logistic regression analyses to predict the answers on the questions about media use, important issues and sports from device used, survey method, answer scale, age, gender and education. Table 3 shows that respondents that completed the survey on a mobile phone significantly chose more often the response options in the long list on the question about most important issues. However, on the other questions respondents that completed the survey on a computer or tablet chose more often an answer option of the long list. The survey method did not have a significant effect on responses. Respondents with the short answer scale gave significantly different answers compared to the respondents with the long answer scale. In addition, in two out of three questions (media use and sports), responses from respondents in the open condition were significantly different than respondents in the long answer scale condition. There were two significant interaction effects, suggesting that respondents with high education that completed the survey on their computer/tablet reported less time in the media question; while men in the open format also reported to spend less time on media use. The model predicts 5.1% of the answers on the question about media use, 24.1% of the answers on the question about important issues and 11.1% of the answers on the question about sports.

Table 3 Results binary logistic regression analyses predicting answers on survey questions

	Media use		Important Issues		Sports	
	B	Exp(B)	B	Exp(B)	B	Exp(B)
Device (0: mobile phone, 1: tablet/PC)	.173	1.189	-.429	.651*	.209	1.233
Survey method (0: RM, 1: regular survey)	-.005	.995	.093	1.097	.224	1.252
Open	.380	1.462*	-.049	.952	-.513	.598*
Closed few	.646	1.908**	-4.595[1]	.010**	-2.537	.079**
Closed many (ref.)						
Age	.020	1.020**	.001	1.001	.001	1.001
Gender (0: female, 1: male)	.400	1.492*	-.530	.588**	.162	1.176
Education (0: ≤ high school graduate, 1: ≥ college)	.203	1.225	-.075	.928	-.044	.957
Education*device	-.498	.608*				
Gender*open	-.513	.599*				
Nagelkerke R^2	.051**		.241**		.111**	
n	1646		1687		1687	

*p< .05, **p< .001
1: the sample size is small.

Note: all the other interaction effects are not significant and were therefore not included in this model. The first question has fewer cases because of the DK option that is omitted from the analysis.

Completion Time

A multiple regression analysis is conducted to predict completion time from device used, survey method, type of answer scale, and personal characteristics. Table 4 shows that the time to complete the survey on a mobile phone was shorter than on a computer or tablet. The research messenger took significantly longer to complete than the regular survey method. Furthermore, the time to complete the survey with a closed answer scale with few options was significant higher than the time to complete a survey with another answer scale condition. There was no significant difference in completion time between the open answer scale and the closed answer scale with many options. Older respondents had a significantly higher completion

time than younger respondents. Women had a significant higher completion time than men. There was no difference in completion time between different levels of education. Although older respondents and respondents on a computer/tablet took longer to complete the survey, the interaction effect shows that older respondents on a computer/tablet took less time to complete the survey. Note that 6% of the variance in completion time can be predicted by the regression model.

Table 4 Results multiple regression analysis predicting the log completion time

	Beta
Device (0: mobile phone, 1: tablet/PC)	.330**
Survey method (0: RM, 1: regular survey)	-.114**
Open	.022
Closed few	.137**
Closed many (ref.)	
Age	.546**
Gender (0: female, 1: male)	-.073*
Education (0: ≤ high school graduate, 1: ≥ college)	.042
Age*device	-.472**
R^2	.086
F	19.666**
n	1686

* p< .05, ** p< .001

Note: One outlier is removed. Other interaction effects are not significant and were therefore not included in this model.

Evaluation Questions

To predict the three evaluation questions from device used, survey method, type of answer scale and personal characteristics, we conducted multiple regression analyses. Table 5 shows that respondents who completed the survey on a computer or tablet answered the evaluation questions significantly more negatively than the

respondents who completed the survey on a mobile phone. There is no significant difference in answers on the evaluation question between the survey methods nor types of answer scale. Older respondents enjoyed the survey significantly more and evaluated the survey as significantly more interesting than younger respondents. Men found the survey significantly more difficult than women. There is a significant interaction effect of gender and education on the difficulty of the survey (men with a high education found the survey more difficult) and a significant interaction effect of gender and survey method on how interesting the respondents evaluated the survey (men that received the regular survey design found the survey less interesting).

Table 5 Results multiple regression analyses predicting answers on evaluation questions

	Difficulty	Enjoyment	Interesting
	Beta	Beta	Beta
Device (0: mobile phone, 1: tablet/PC)	.043	-.066*	-.050*
Survey method (0: RM, 1: regular survey)	-.019	-.009	.024
Open	-.001	-.024	-.042
Closed few	-.023	-.025	-.000
Closed many (ref.)			
Age	-.012	.107**	.102**
Gender (0: female, 1: male)	.064*	-.001	.029
Education (0: ≤ high school graduate, 1: ≥ college)	-.055	.062*	.044
Gender*education	.084*		
Gender*survey			-.084*
R^2	.015	.015	.014
F	4.228**	4.663**	4.080**
n	1687	1687	1687

* p< .05, ** p<.001

Note: Other interaction effects are not significant and were therefore not included in this model.

Discussion and Conclusion

The goal of this research was to determine the best way to present mixed-device surveys. In order to do this, we investigated the differences between three different types of answer scales, two survey methods and the devices used.

Respondents with the close-ended answer scale with few options choose less often an answer option in the long list (e.g. by choosing other please specify) than respondents in other answer scale conditions. The completion time of the respondents who completed the survey with the close-ended answer scale with few options is also longer. An explanation for this could be that the true answer of the respondents is not in the list with few options. Previous research has shown that respondents tend to choose one of the options of the list even if it is not their true answer (Chung, Boyer & Han, 2010; Reja et al., 2003). Therefore, the respondents have to think longer about their answers and eventually choose an answer that is one of the options. Especially on the question about the most important issue, a very small number of respondents with the close-ended answer scale with few options have chosen an answer option that was only in the list of the close-ended answer scale with many options. This question does not have one possible answer, because respondents could think more issues are important. Therefore, it is likely that the respondents tend to choose an answer that is available rather than "other, please specify". For the question about sports the answers of respondents with the open-ended answer scale are also different than the answers of respondents with the close-ended answer scale with many options. Respondents that choose an answer option that was only in the longer list is in general low. This could be explained by the fact that the sports that were only in the longer list are in general not popular sports. For example, bowling was popular in America in 1960 but that popularity has declined (McIntosh, 2011). The difference in answers on the media question between the respondents in different answer scale conditions might have been caused by the difference in the scales instead of only by the difference in the presentation of the scales. As shown in appendix 1, the answer options are different, for example the first category of the closed-question with few options is "less than an hour" while the first category of the closed-question with many options is "less than half an hour". The answer options can be suggestive and serve as anchors for the respondents (Desai & Reimers, 2019). Therefore, the different scales with different categories could explain the difference in answers between respondents with few answer options and respondents with many answer options.

The time to complete the research messenger survey is longer than the time to complete the regular survey. The conversational element in the research messenger survey takes more time, because the respondents have to wait for the next question. However, the respondents who completed the research messenger survey did not answer the evaluation questions of the survey more negative. The break-off rate of

respondents who completed the research messenger survey is also not higher. So, the respondents with the research messenger survey did not seem to mind that it took longer to complete. However, the respondents did not evaluate the research messenger more positive compared with the regular survey. Which indicates that a research messenger survey might not be a better way to present mixed-device surveys.

In contrast to other studies and our assumptions, the respondents who completed the survey on a computer or tablet evaluated the survey more negative than the respondents who completed the survey on a mobile phone. However, the effects are small and did not depend on survey method.

In general, respondents can self-select the device to complete an online survey on. Therefore, in this research the respondents also could self-select the device. However, we expect that the effect of mobile phones would have been greater if respondents were assigned to a device, because then there were also respondents in the mobile phone condition who are less experienced with using a mobile phone. Furthermore, only single items were used in this research. We expect that the effects would be greater if we analyzed rating scales that encompass multiple items, since the effect on multiple items would be measured instead of the effect on a single item.

The completion time of older respondents was higher than of younger respondents. Despite the higher completion time, older respondents found the survey more enjoyable and interesting. Also, women had a significant higher completion time than men. However, men evaluated the survey as more difficult than women.

The survey was distributed among Amazon Mechanical Turk panel members in the United States of America. These panel members received a compensation for completing the survey. This could be a reason for the low break-off rate of the survey. Furthermore, the panel members might evaluate the research messenger survey more negatively than respondents who are not in a panel, because they are used to completing regular online surveys. Moreover, since the panel members are trained in completing surveys, they might have less problems with open-ended questions or closed-ended questions with many answer options, such as a longer completion time. Although the respondents are not representative of the population, the sample was heterogeneous. The sample is heterogenous, though not representative of the population. Especially female respondents are overrepresented. Future research using a probability-based sample should be used to replicate our results and test robustness.

In conclusion, based on our results we recommend to use a close-ended answer scale with many options or an open-ended answer scale since a closed-ended answer scale with few options results in other frequencies hence outcomes. The research messenger survey did not seem to be a better method to present mixed-device surveys than a regular survey. Further research is necessary to investigate

how to present mixed-device surveys in order to increase participation and data quality in mixed-device surveys.

References

Ahad, A. D., & Lim, S. M. A. (2014). Convenience or Nuisance?: The 'WhatsApp' dilemma. *Procedia–Social Behavioral Sciences, 155*, 189-196. DOI:10.1016/j.sbspro.2014.10.278

Antoun, C., Katz, J., Argueta, J., & Wang, L. (2018). Design heuristics for effective smartphone questionnaires. *Social Science Computer Review, 36*(5), 557-574. https://doi.org/10.1177/0894439317727072

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*(3), 281-291. https://doi.org/10.1093/pubmed/fdi031

Brant, J. M., Haas-Haseman, M. L., Wei, S. H., Wickham, R., & Ponto, J. (2015). Understanding and evaluating survey research. *Journal of the Advanced Practitioner in Oncology, 6*(2), 168–71. DOI:-

Chung, C., Boyer, T., & Han, S. (2010). How many choice sets and alternatives are optimal? Consistency in choice experiments. *Agribusiness, 27*(1), 114-125. https://doi-org.proxy.library.uu.nl/10.1002/agr.20252_

Couper, M. P., & Miller, P. V. (2008). Web survey methods. *Public Opinion Quarterly, 72*(5), 831-835. https://doi.org/10.1093/poq/nfn066

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? Social Science Computer Review, 35(3), 357-377. https://doi.org/10.1177/0894439316629932

Cunningham, J. A., Neighbors, C., Bertholet, N., & Hendershot, C. S. (2013). Use of mobile devices to answer online surveys: Implications for research. *BMC Research Notes, 6*, 258. https://doi.org/10.1186/1756-0500-6-258

De Bruijne, M., & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review, 31*(4), 482-504. https://doi.org/10.1177/0894439313483976

De Bruijne, M., & Wijnant, A. (2014). Improving response rates and questionnaire design for mobile web surveys. Public Opinion Quarterly, 78(4), 951-962. https://doi.org/10.1093/poq/nfu046

De Leeuw, E. D. (2008). Self-administered questionnaires and standardized interviews. *Handbook of social research methods*, 313-327. DOI: -

Desai, S. C., & Reimers, S. (2019). Comparing the use of open and closed questions for Web-based measures of the continued-influence effect. *Behavior research methods*, *51*(3), 1426-1440. https://doi.org/10.3758/s13428-018-1066-z

Dolnicar, S., Grün, B., & Yanamandram, V. (2013). Dynamic, interactive survey questions can increase survey data quality. *Journal of Travel & Tourism Marketing, 30*(7), 690-699. https://doi.org/10.1080/10548408.2013.827546

Harb, E., Kapellari, P., Luong, S., & Spot, N. (2011). Responsive web design. *Version of*, *6*. Retrieved from http://courses.iicm.tugraz.at/iaweb/surveys/ws2011/g3-survey-resp-web-design.pdf

Harms, J., Seitz, D., Wimmer, C., Kappel, K., & Grechenig, T. (2015). Low-cost gamification of online surveys: Improving the user experience through achievement badges. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in* Play*, 15*, 109-113. http://dx.doi.org/10.1145/2793107.2793146

Harris, L. R., & Brown, G. T. L. (2010). Mixing interview and questionnaire methods: Practical problems in aligning data. *Practical Assesment, Research, and Evaluation, 15*(1), 1-14. https://doi.org/10.7275/959j-ky83

Heiervang, E., & Goodman, R. (2011). Advantages and limitations of web-based surveys: Evidence from a child mental health survey. *Social Psychiat Epidemiol, 46*, 69-76. https://doi.org/10.1007/s00127-009-0171-9

Hussain, A., & Mkpojiogu, E. O. (2015). The effect of responsive web design on the user experience with laptop and smartphone devices. *Jurnal Teknologi*, *77*(4), 41-47. DOI:-

IBM SPSS Statistics (26.0.0) [Computer software]. United States: IBM Company.

Kim, S., Lee, J., & Gweon, G. (2019). Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. *Proceedings of the 2019 CHI conference on human factors in computing systems, 86,* 1-12. DOI:-

Maslovskaya, O., Smith, P., & Durrant, G. (2020). Do respondents using smartphones produce lower quality data? Evidence from the UK Understanding Society mixed-device survey. *National Centre of for Research Methods Working Paper, 1* (20), 2-32. DOI:-

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review, 31*(6), 725-743. https://doi.org/10.1177/0894439313485201

McIntosh, P. (2011). Bowling: Entertainment for All Ages. *English Teaching Forum, 49*(4), 37-45. DOI:-

Millar, M., & Dillman, D. A. (2012). Encouraging survey response via smartphones. *Survey Practice*, *5*(3), 3095. https://doi.org/10.29115/SP-2012-0018

O'Hara, K., Massimi, M., Harper, R., Rubens, S., & Morris, J. (2014). Everyday dwelling with WhatsApp. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 14*, 1131-1143. https://doi.org/10.1145/2531602.2531679

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74-97.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, *104*(1), 1-15. https://doi.org/10.1016/S0001-6918(99)00050-5

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. Close-ended Questions in Web Questionnaires. *Developments in applied statistics*, *19*(1), 159-177. DOI: -

Schlosser, S., & Mays, A. (2018). Mobile and dirty: Does using mobile devices affect the data quality and the response process of online surveys? *Social Science Computer Review, 36*(2), 212-230. https://doi.org/10.1177/0894439317698437

Seaborn, K., & Fels, D. I. (2014). Gamification in theory and action: A survey. *International Journal of human-computer studies*, *74*, 14-31. https://doi.org/10.1016/j.ijhcs.2014.09.006

Smith, G. (2008). Does gender influence online survey participation?: A record-linkage analysis of university faculty online survey response behavior. *ERIC Document Reproduction Service No. ED 501717,* 2-21.

Solomon, D. J. (2000). Conducting web-based surveys. *Practical Assessment, Research, and Evaluation, 7(*19), 1-4. https://doi.org/10.7275/404h-z428

Stapleton, C. E. (2013). The smartphone way to collect survey data. *Survey Practice, 6*(2), 1-7. DOI:-

Steinbrecher, M., Roßmann, J., & Blumenstiel, J. E. (2015). Why do respondents break off web surveys and does it matter? Results from four follow-up surveys. *International Journal of Public Opinion Research*, *27*(2), 289-302.

Symon, G., Cassell, C., & Dickson, R. (2000). Expanding our research and practice through innovative research methods. *European Journal of Work and Organizational Psychology, 9*(4), 457-462. https://doi.org/10.1080/13594320050203076

Toepoel, V. & Lugtig, P. J. (2015). Online surveys are mixed-device surveys. Issues associated with the use of different (mobile) devices in web surveys. methods, data, analysis, 9(2), 155-162. https://doi.org/10.12758/mda.2015.009

Tourangeau, R., Maitland, A., Rivero, G., Sun, H., Williams, D., & Yan, T. (2017). Web surveys by smartphone and tablets: Effects on survey responses. *Public Opinion Quarterly*, *81*(4), 896-929. https://doi.org/10.1093/poq/nfx035

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.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(1), 51-68.

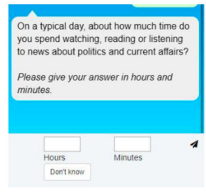
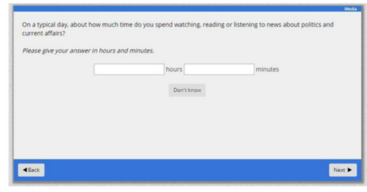
Zhang, X., Kuchinke, L., Woud, M. L., Velten, J., & Margraf, J. (2017). Survey method matters: Online/offline questionnaires and face-to-face or telephone interviews differ. *Computers in Human Behavior*, *71*, 172-180. https://doi.org/10.1016/j.chb.2017.02.006

Appendix 1

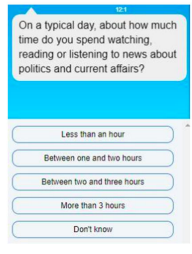
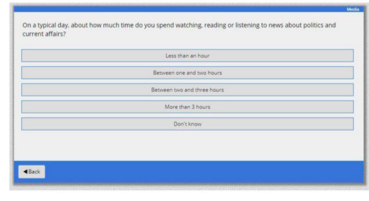
Screenshot of the answer scales and survey methods of the questions used. Video's are available upon request

Question about media use.

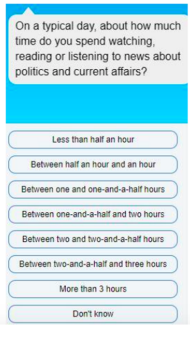
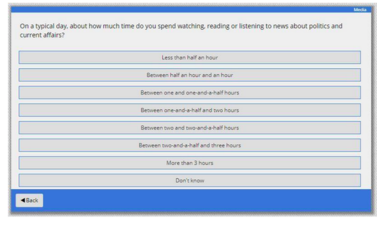
Section A: Media use: A1A

Questionnaire	RM version	Traditional version
[1A]. On a typical day, about how much time do you spend watching, reading or listening to news about politics and current affairs? Please give your answer in hours and minutes. **(source: European Social Survey (round 8))** ___ (open answer) Don't know		

Section A: Media use: A1B

Questionnaire	RM version	Traditional version
[1b]. On a typical day, about how much time do you spend watching, reading or listening to news about politics and current affairs2? **(source: European Social Survey (round 8))** Less than an hour Between one and two hours Between two and three hours More than 3 hours Don't know		

Section A: Media use: A1C

Questionnaire	RM version	Traditional version
[1c]. On a typical day, about how much time do you spend watching, reading or listening to news about politics and current affairs2? **(source: European Social Survey (round 8))** Less than half an hour Between half an hour and an hours Between one and one-and-a-half hours Between one-and-a half and two hours Between two and two and a half hours Between two and-a half and three hours More than 3 hours Don't know		

Question about most important issue.

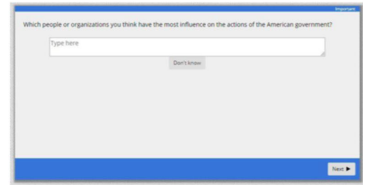
Section B: Most important problem:B_A4

Questionnaire

4. **(GSS)** Which people or organizations you think have the most influence on the actions of the American government?
___ **(open answer)**

RM version

Traditional version



Section B: Most important problem:B_B7

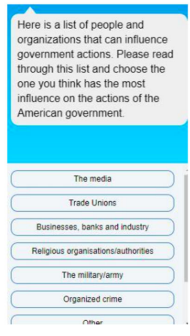
Questionnaire

7. **(GSS**)Here is a list of people and organizations that can influence government actions.
Please read through this list and choose the one you think has the most influence on the actions of the American government

The media
Trade Unions
Businesses, banks and industry
Religious organisations/authorities
The military/army
Organized crime
Other, ___
I can't choose
Don't know

[if "can't choose or "don't know" -> go to module c]

RM version



Traditional version

Section B: Most important problem:B_C3

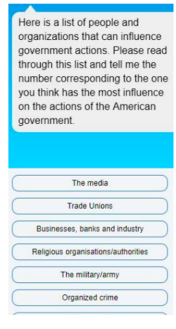
Questionnaire

3. (GSS) Here is a list of people and organizations that can influence government actions.
Please read through this list and tell me the number corresponding to the one you think has the most influence on the actions of the American government

The media
Trade Unions
Businesses, banks and industry
Religious organisations/authorities
The military/army
Organized crime
Schools/education
Ordinary People
Rich people
Other, ___
I can't choose
Don't know

[if "can't choose or "i don't know" -> go to module c]

RM version



Traditional version

Question about sports.

Section D: Leisure Time and Sports: D_1A

Questionnaire

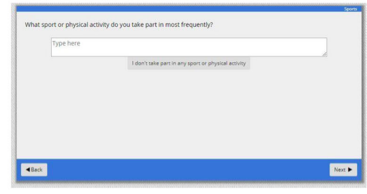
What sport or physical activity do you take part in most frequently? ((If you do not take part in any sport or physical activity, please tick the box provided below.)) *<OPEN-ENDED>*

Most frequent sport or physical activity
_____ (open answer)

I do not take part in any sport or physical activity

RM version

Traditional version



Section D: Leisure Time and Sports: D_1B

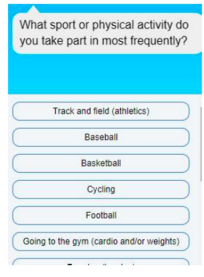
Questionnaire

[1b] What sport or physical activity do you take part in most frequently?

Track and field (athletics)
Baseball
Basketball
Cycling
Football
Going to the gym (cardio and/or weights)
Running (jogging)
Swimming for fitness
Treadmill or Gym equipment at home
Walking for fitness
Other, please write ___ **(open answer)**
Not sure

Do not know

RM version



Traditional version

Section D: Leisure Time and Sports: D_1C

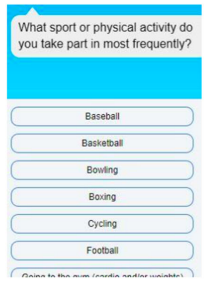
Questionnaire

[1c] What sport or physical activity do you take part in most frequently?

Baseball
Basketball
Bowling
Boxing
Cycling
Football
Going to the gym (cardio and/or weights)
Golf
Horse riding
Ice Hockey
Running (jogging)
Soccer
Swimming for fitness
Tennis
Track and field (athletics)
Treadmill or Gym equipment at home
Walking for fitness
Yoga/Pilates
Other, please write
Not sure

Do not know

RM version



Traditional version

methods, data, analyses | Vol. 17(1), 2023, pp. 71-92 DOI: 10.12758/mda.2022.12

Measuring Congruence Between Voters and Parties in Online Surveys: Does Question Wording Matter?

Bastiaan Bruinsma

Chalmers University of Technology

Abstract

Congruence on policies between political parties and voters is a frequently assumed requirement for democracy. To be able to study this, we should be able to calculate accurate and precise measures of policy congruence in political systems. This could then tell us more about the political system we study, and the "distances" that exist between parties and voters on either issues or broader ideological dimensions. Here, I draw on experimental data from a Voting Advice Application to show that the wording of the issues can influence the degree of congruence one measures. Yet, this comes with the complication that this influence depends on the type of issue, the characteristics of the voters themselves, and the party the congruence is calculated with. These findings should serve as a warning for those who aim to measure congruence that even minor changes in question-wording can (but do not have to) cause relatively large changes in congruence, especially when many parties are involved and the differences between the congruences are small.

Keywords: congruence, wording effects, voting advice applications

Congruence between voters and parties is one of the cornerstones of democratic representation (e.g., Huber & Powell, 1994; Katz, 1997; Powell, 2004). The idea is that the higher the level of congruence, the better-represented we can say citizens are. As such, we can use the degree of congruence to say something about the degree of democracy in a country (Diamond & Morlino, 2005). While congruence is usually high on broad ideological dimensions - such as economic left-right - it is often lower for individual policy issues (Dolný & Baboš, 2015; Dalton, 2017). This is problematic, as while dimensions are still important, there is a significant number of voters who vote on issues, instead of broad ideological dimensions and partisanship (Dalton, 2017). So, it is no surprise that the number of studies that focus on issue congruence is increasing (e.g., Belchior & Freire, 2013; Romeijn, 2020; Rosset & Stecker, 2019; Costello et al., 2020).

To measure issue congruence, one needs valid and unbiased positions of both voters and parties on these issues, as well as a mechanism to compare them. To find the position of a party, there are various methods, each of which has its advantages and drawbacks (Mair, 2001; Benoit & Laver, 2006). Often, they are dependent on experts, though they may also use the opinion of the party itself. For the position of the voter, the only option seems to be to ask the voter themselves, most often in the form of a survey. Here, one presents them with a series of issues and ask them to state their agreement or disagreement with them. While straightforward, this requires a certain degree of response quality on part of the voter to be useful. In other words, we need the voter to provide a response that is as close to their "true" opinion as possible.

Yet, such a high degree of response quality is by no means guaranteed. At its best, the response quality of the average voter is often only sub-optimal (Blasius & Thiessen, 2012, p.3). This is especially so if the voter perceives the formulation of the issue to be difficult, which can happen when the wording of the issue is long and complicated, or when the issue uses negative wording. Given the increased relevance of issue congruence, the question is then how relevant response quality is when measuring congruence. To put it in other words: to what extent does the "difficulty" of the formulation of the issue influence its measure?

Of the various ways in which an item can be difficult, one of the simplest forms is whether an issue is positive or negative. This is also known as the *polarity* of an issue. Negative issues, as well as their positive counterparts, occur in almost all surveys. Survey designers include them to: a) minimize response styles such as acquiescence bias, and b) allow for the inclusion of negatively worded issues from previous surveys. Yet, it is well known that negative issues come with certain prob-

Direct correspondence to
Bastiaan Bruinsma, Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden
E-mail: sebastianus.bruinsma@chalmers.se

lems. For example, respondents might miss the negation in the sentence, or become confused about how to map their response to the response options. As a result, they spend longer on reading the item (Kamoen et al., 2011) and mapping their response to the correct option (Chessa & Holleman, 2007). This, in turn, leads to an increase in non-response and respondents giving the opposite response to the one they intended. Holleman et al. (2016) show the latter in an experiment where they changed some issues from positive to negative, such as changing *forbid* to *allow*. They found that the responses were different than they should be (i.e., their direct opposite), though the effect they found was rather small (on a scale of 1 to 5, only 1 out of 20 respondents showed a scale point difference) (Holleman et al., 2016, p. 9).

Yet, many small differences in response might have large effects when it comes to the congruence between the voter and the party. Besides, it is unclear if the effect of changing the polarity is the same for all issues, parties, and voters. For example, certain issues might be more susceptible to such effects while others might not show any difference at all. Also, if a party associates itself with a certain formulation of an issue it might suffer if the formulation of the issue is the opposite. Finally, voters with different levels of political sophistication might show differences in the way polarity influences them. For example, voters with low political sophistication could find it more difficult to handle negatively worded issues than voters with a high level of political sophistication.

In this paper, I will thus assess the following: *does changing the polarity of an issue affect the degree of congruence between a voter and a party?* For this, I will run an experiment in which I present voters with both the positive and negative versions of the same issues. In addition, I will also take the effects of issue salience and respondent characteristics into account. The experiment itself is carried out in the context of a Voting Advice Application (VAA) - an online tool voters can use before the elections to calculate their congruence with certain parties. Not only is congruence central to the idea of the VAA, but the VAA community itself has also been actively involved in studying how and when variations in congruence occur (Louwerse & Rosema, 2014).

From here on, the structure of this paper is as follows. First, I will introduce the increased relevance of issue congruence, which leads to a discussion on how we should measure it. Based on the congruence literature, I will identify several factors that might influence the effect of the question wording such as issue salience and the political sophistication of the respondent. Then, I will describe the research design and the set-up of the experiment as well as the measures used. Finally, I will attempt to answer the main research question as well as discuss several implications of the findings.

Relevance of Issue Congruence

We can define congruence as the distance between a citizen and their representative (Dalton, 1985). This representative could be either a single candidate or a party. The idea of congruence itself was for a long time equated with ideological congruence, which referred to the congruence between voters and parties on an ideological dimension. If on this dimension the distance between the voter and those that represent them neared or equaled zero, they were said to be congruent.

While simple, there are several problems with the ideological congruence approach. The first is that the positions of voters and parties often have different meanings. Voters are, by definition, a larger group than parties, resulting in their positions varying more than those of the parties. Second, issues not included in the dimension might be relevant for either the voter or the party. It could thus be that a voter and a party appear to be congruent, even though they disagree on issues that are fundamental to both. Besides, there is no reason to suppose that congruence on a dimension also implies congruence on a certain issue (Thomassen, 2012). And while this might not be a problem when the issue is not salient to either the party or the voter, it becomes a problem when it is (Giger & Lefkofridi, 2014). As a result of this, Dalton (1985, 2017) shows that congruence not only differs between issues but between parties as well. Finally, for ideological congruence to work, one needs to position both the party and the voter on the same metric. Yet, due to data limitations, this is often not possible, and parties and voters are often positioned on different metrics that scholars assume to be similar.

Partly because of the first two issues, the focus on issue congruence has increased over the last years. To measure it, one can use either of three approaches, dependent on how one measures the positions of the parties and the voters (Powell, 2009). These are a) a voter identification and voter perception approach, b) a party vote and party manifesto approach, and c) a voter survey and expert survey approach. In the first, voters' position both themselves and any parties on a certain issue. While this tackles the problem of not using the same metric, it assumes that voters have a good knowledge of both their own and of the parties' position. This goes against the experience that voters are rarely that well informed on these matters. Also, a voter's perception of the position of a party is dependent on their own position as well. In the second approach, one takes voters in part out of the equation by looking at their votes during the elections to estimate their position. For the position of the party one then takes the electoral manifesto the party released for the same elections. While this circumvents the problems with the voters, one's vote is rarely indicative of one's true position - especially if choices are limited. Also, the use of party manifestos to position parties is not uncontroversial as well (Dinas & Gemenis, 2010). The third option - combining a voter survey and expert survey approach seems to tackle most problems and is, therefore, the approach used

by most issue-congruence scholars (e.g., Giger & Lefkofridi, 2014; Costello et al., 2020). The reason it works is that voters are often better aware of their positions on single issues than on a whole dimension and expert surveys are often viewed as the "gold standard" for party positioning and are flexible to implement.

Effects of Question Wording

The remaining problem with measuring issue congruence are then the issues themselves. Not only does the actual content of an issue matter, but also its formulation (e.g., Hippler & Schwarz, 1986). For example, Schuldt et al. (2011) found Republicans less likely to endorse the existence of "global warming" than the existence of "climate change". This is because different terms draw the readers' focus to different aspects of the phenomenon. One might accept the "warming" while agreeing on "change" is something very different. The same occurs when we alter the polarity of the question. That is, agreeing with something does not mean that one disagrees with the opposite. For example, agreeing that soft drugs should *be forbidden* does not mean that one disagrees that they should *be legalized*.

A negation such as *forbidden* is known as an implicit negative. That is, the negation is in the word itself. Its alternative is the explicit negative, where one places the word "not" in front of the positive verb (Clark, 1976). Thus, it would be *legalize* versus *not legalize*. Of the two, the implicit version is often the most popular option. This is because one can only use an implicit negative in a context where it makes sense. For example, one does not talk of *keeping someone from* suicide unless one supposes that someone intended to commit suicide in the first place (Horn, 1989, p.523). Also, the alternative, the explicit negatives, take longer to understand. This is as one first must reconstruct the positive version, and then make it negative. As an example, Kaup et al. (2006) asked respondents to imagine a non-open door. Most respondents took twice as long in imaging this as they would in imaging an open door, as they first had to open the door and then "non-open" it.

The effect of a change in the polarity of an item can then differ on three levels: the respondent, the party, and the issue. As such, we should focus on each of them to see how they behave when we change the polarity of an item. To begin with, we set out to see whether question polarity has any influence on the response behavior of the respondents at all. Given that it seems reasonable to assume that respondents will respond differently when an item has a different formulation, I hold that:

H1 Question polarity affects the responses respondents give to items.

Turning first to the respondents, it is likely that the characteristics of the respondent condition the effect of changing the polarity of the question. On considering public policy issues, political sophistication is one of the main relevant characteristics.

The concept itself is multi-faceted. While it relates to political knowledge, it also includes the facility to acquire new information, the ability to translate personal values into opinions and behavior, and one's motivation to do so (Luskin, 1990; Highton, 2009). Thus, respondents with low levels of political sophistication are more likely to have problems with negative questions, as the negative version of a question is more complicated (Holleman et al., 2017). Besides, respondents with a low level of political sophistication will pay less attention to these questions (Bassili & Krosnick, 2000), again increasing the degree of misunderstanding. It is thus no surprise that Holleman et al. (2016) found that lower levels of political sophistication led to larger effects of the polarity. It seems thus reasonable that the opposite is also the case. Therefore:

H2 The higher the sophistication of the respondent, the lower is the effect of the polarity.

As for the issues, one characteristic that might influence the responses is the salience of an issue. Most often, one measures this by asking after the "most important issue" during a certain period. The more important the issue to people, the higher its salience. As a result, voters are better informed on those issues, as they tend to be so when the issue is important to them (Giger & Lefkofridi, 2014). This makes the issue less difficult for the voter, leading them to earlier spot a change in the difference of wording. Thus:

H3 The higher the salience of the issue, the lower the effect of question polarity.

Finally, we turn to the parties. As there is a seemingly one-on-one relation between the response of the respondent and the congruence (as the position of the party does not change here), we expect that any change in the response will also lead to a change in the degree of congruence. As an extension of the first hypothesis, therefore:

H4 The polarity of the item influences the congruence between respondents and parties.

Research Design

To measure the influence of the question wording on the degree of congruence, I will turn to an instrument that has congruence at its heart: Voting Advice Applications (VAAs). These are online questionnaires that compare the answers of their respondents with those of political parties on the same issues. This comparison is then shown as the degree of agreement between the respondent and the party - in other words, their congruence (see also Costello et al., 2020). Using VAAs to calcu-

late issue congruence has several advantages. First, the items are the same for both respondents and parties, thus avoiding the complications other methods have when comparing the positions (Krosnick & Presser, 2010; Costello et al., 2020). Also, we can calculate the match without requiring any further scaling analysis. Finally, as VAAs are popular online instruments, reaching the desired number of responses is easy.

As VAAs, in general, have only a single version of the question wording, to test for the effect of question polarity, we have to run two similar VAAs. Even so, as the wording of the question could potentially influence the respondent and thus the outcome of the VAA, it would be unethical to run two versions with either only positive or only negative items. To get around this, I designed a VAA with two different versions (hereafter Version A and Version B). Both versions had 25 questions in total, with 13 questions being similar for both (these questions all had a positive wording). Of the remaining 12 questions, the polarity of the wording differed for each version. Of the negative questions, 4 of them were explicit negatives and 8 were implicit negatives. The topics of the questions were decided upon in cooperation with other members of the design team based on their expected relevance during the elections. For a full overview of the questionnaire, see Appendix A.

The VAA designed for this study – Stem-Consult – launched several weeks before the elections for the House of Representatives in the Netherlands on March 15, 2017[1] (Gemenis et al., 2017). I reached out to potential respondents through word-of-mouth and targeted Facebook advertisements. Upon entering the website, the VAA assigned respondents at random to either Version A or B of the VAA. Besides the main questionnaire, optional questions asked respondents for their age, political interest, education, and gender. The VAA included 14 of the 28 parties taking part in the elections, which were included either because they were already represented in parliament or showed in the polls a consistent chance of gaining at least a single seat. As some parties were favored only by a small percentage of the respondents, I decided to only include here the 8 largest parties. These are the CDA, a Christian-democratic party, the CU, a social Christian party, D66, a social-liberal party, GL, a green party, the PvdA, a social-democratic party, the PVV, a radical right populist party, the SP, a radical left party, and the VVD, the main center-right liberal party. For an overview of these parties and their abbreviations, see Appendix B.

1 Stem-Consult was designed and launched in cooperation with the PreferenceMatcher consortium.

Voter and Party Positions

In a VAA, respondent positions come from the main questionnaire. Here, the VAA presents the respondent with 25 items - one item at the time - and asks them to respond on a 5-point Likert scale. This scale ranges from *Completely Disagree* to *Completely Agree*, with a *No Opinion* option included. For these questions, there is no time limit and respondents have the opportunity of returning to a previous item if they wish to change their response.

For the party positions, I employed a team of coders that coded the positions of the parties. This coding takes place on the same issues and uses the same response options as the respondents are presented with. For the coding, I used the iterative Delphi-process (Gemenis, 2015). This process presents the coders with the issues in a first round and asks what they think would be the party's position on it. Together with their answer, they are then asked by the moderator to back up their idea of the party's position using any source of information they please. Also, they have to tell how confident they are of their opinion. This information is then collected, anonymized, and fed back to the coders. This allows them to see how the other coders positioned a party on an issue, as well as how confident they are of this position. Then, the moderator asks them if they want to reconsider their original position. The idea then is that those respondents who were not very confident of the position will alter theirs to be more in line with those that are confident. This process then repeats until the coders reach a certain degree of agreement between them. Here, I calculate this agreement following Gemenis (2015) and use Van der Eijk's coefficient α (van der Eijk, 2001). I do so as the ordinal rating scales used to position parties in VAAs do not lend themselves well to other common methods of agreement, such as the standard deviation, as these often reflect the skewness of distribution in addition to the dispersion (van der Eijk, 2001, p.328). Van der Eijk's coefficient α circumvents this by taking a weighted average of the degree of agreement that is there for the individual categories. Though Van der Eijk (2001) does not offer any cut-off point for α, I follow Gemenis and Van der Ham (2014), who carried out a similar analysis during the Dutch elections of 2014 and require the agreement to be higher than $\alpha > 0.7$ for the process to finish. When at this point the positions of the coders still differ, I take the average between them as the position of a party. For this VAA, the coders positioned all parties on the questions as they appear in version A of the VAA. The positions for Version B were then generated by reversing the positions where necessary. While this is in no way ideal, constraints of both time and resources led us to settle on this approach.

Congruence Measures

To calculate the agreement between a respondent and a party, VAAs can draw on two schools of thinking. The first draws on Downs (1957) and holds that agreement is the distance between a party and respondent. The second finds its origin in Rabinowitz and MacDonald (1989) and supposes it is not the distance, but the intensity that counts - relevant is thus if the respondent and the party are on the same side of the argument, regardless of the distance between them. As I deem both to be relevant, I opt to use a hybrid model that splits the difference between both methods (Mendez, 2017). With this algorithm, the degree of congruence can range between -100 to 100. Here, -100 means that when the respondent completely disagreed the party completely agreed, or the respondent completely agreed the party completely disagreed. On the other hand, +100 means that both the respondent and party completely disagreed or completely agreed. Also, 0 means that either the party or the respondent was neutral while the other completely agreed or disagreed (cf. Mendez, 2014). For a complete overview of how this algorithm works, see Appendix C.

Political Sophistication

For political sophistication, I create an additive scale using the education and political interest variables, both of which are measured on a five-point scale. The result of this is an additive scale running from 2 to 10, which lower values indicating a low level of political sophistication and higher values indicating a higher level of political sophistication.

Issue Salience

To measure the degree of issue salience, I will make use of the data supplied by the Dutch Parliamentary Election Study 2017 (van der Meer et al., 2017), which was fielded around the elections during which the VAA was implemented. Here, I only used the responses collected by CAPI, which provided 927 respondents who completed the questionnaire. Then, for issue salience, I used the question "What do you think are the most important problems in our country?"[2]. This question resulted in an open answer where the respondent could name more than one problem, they considered important. These answers were then re-coded into nine categorical variables ranging from the most important problem to the ninth, sorted into twenty-two different issue categories. Of the 927 respondents, 857 mentioned at least a single problem. I used the average frequency of mentions of an issue over the total number

2 The original wording in during the survey was: "*Wat zijn volgens u de belangrijkste problemen in ons land?*"

of issues mentioned as the average importance of that issue (Hosch-Dayican et al., 2013). I then related the twenty-two categories from the DPES to the topics of the items in Stem-Consult. The results of this are in Table 1. Here, the first column refers to the item in the VAA, the second to which of the twenty-two different issue categories it belongs, and the third shows the percentage of respondents that mentioned that issue.

Table 1 Overview of the issues and their percentage of issue salience

#	Item	Type of Issue	Salience (%)
1	Art subsidies	Media	0.2
2	Public broadcasting	Media	0.2
3	Development cooperation	Inequality/Poverty	5.4
4	Pension Age	Social security	3.1
5	Mortgage relief	Housing	1.5
6	Anonymous application	Inequality/Poverty	5.4
7	Insurance	Income/Price levels/Taxes	1.5
8	End of life	Norms and values	5.0
9	Threat of Islam	Norms and values	5.0
10	Soft drugs	Norms and values	5.0
11	Remain in EU	European integration	1.4
12	EU Expansion	European integration	1.4
13	Immigration	Norms and values	5.0
14	Acceptance of refugees	Minorities	18.3
15	Environmental measures	Environment	7.9
16	Energy saving measures	Regulation/Big government	0.9
17	Coal Plants	Environment	7.9
18	Loan system	Education	7.3
19	Binding referendum	Politics	3.8
20	Defense	Defense	0.9
21	Spending on social work	Social security	3.1
22	Own risk in healthcare	Healthcare	19.4
23	Healthcare and market	Healthcare	19.4
24	Mileage charge	Traffic/Mobility	1.3
25	Multicultural society	Minorities	18.3

Data

Given that VAA data are online data, Mendez et al. (2014) and Andreadis (2014) advise cleaning up the data before using it. Following their advice, I removed respondents when: the time taken to complete the total of 25 issue statements was less than 75 seconds; they answered at least one issue in less than 2 seconds or when they answered 12 or more consecutive statements in the same way. Besides, I removed returning respondents – identified by similar entries from the same computer – as well as all respondents taking the VAA after March 15, which was the date of the election, and those of the VAA between 10-12 March, when the VAA was taken offline for a security update. Finally, I removed the first 50 entries as these most likely were filled out during initial testing. After doing so and selecting those respondents who filled out all the items and for which data on political interest and education was available, 2674 respondents remained.

Of these, 1328 were in Version A and 1346 in version B. Both groups were not different with regard to Sex ($X^2(1)=0.07$, $p=0.79$), Age ($t(2670)= -0.61$, $p=0.54$), or Education ($X^2(1)=0.39$, $p=0.53$), and though there was a significant difference in Political Interest ($X^2(1)=4.75$, $p<0.05$), the actual differences of 2.40 for Version A and 2.47 for Version B are too small on a 5-point scale to be expected to make any conceivable difference.

Within the sample, 45.8% was male and 53.9% was female. Compared to the general population with 49.6% male and 50.4% female, this indicates that Stem-Consult attracted more females than males. As for age, the mean age for Stem-Consult was 38.2 for men and 43.0 for women, while in the general population these are 40.7 and 42.5. For education, I find that 51.7% of the respondents had a graduate or post-graduate education, compared to 23.5% in the population. This means that Stem-consult reached younger and higher educated respondents, as is common for most VAAs (van de Pol et al., 2014).

Results

Before we look at the hypotheses, it might be instructive to look at those items not included in the analysis - that is, those items that were positive in both versions. For these items, in contrast to those that had different versions, there should be no difference in the response between the two versions of the VAA.

Table 2 Independent samples t-test for the unaltered items of the VAA. A and B refer to either of the two versions of the VAA. For significance, * (p<0.05), ** (p<0.01), *** (p<0.001)

	Mean					CI	
Item	A	B	t	p	Sig.	Low	High
q3	3.51	3.57	-1.32	0.19	-	-0.14	0.03
q4	2.57	2.53	0.79	0.43	-	-0.06	0.14
q6	3.39	3.52	-3.06	0.00	**	-0.22	-0.05
q7	2.66	2.73	-1.76	0.09	-	-0.15	0.01
q8	2.07	2.08	-0.34	0.73	-	-0.09	0.07
q9	2.81	2.83	-0.30	0.76	-	-0.11	0.08
q12	3.68	3.71	-0.70	0.48	-	-0.10	0.05
q13	1.58	1.59	-0.43	0.66	-	-0.07	0.04
q14	3.65	3.67	-0.48	0.63	-	-0.10	0.06
q19	2.97	2.97	-0.05	0.96	-	-0.10	0.09
q20	2.60	2.56	0.80	0.42	-	-0.05	0.11
q23	3.61	3.73	-2.71	0.01	**	-0.20	-0.03
q24	2.71	2.78	-1.67	0.10	-	-0.16	0.01

To see if degree this is the case, Table 2 shows the independent samples t-test for each of these thirteen items. As expected, in all but two cases, there is no statistically significant difference in the respondents' responses. The two exceptions to this are item 6 ("Anonymous application must become the norm for government jobs") and item 23 ("Through free-market operation, healthcare functions better"). In these two cases, the differences between both versions are 0.13 and 0.12 respectively. While there is no clear evidence on what causes these differences, given their small size it seems quite unproblematic to ignore them.

Table 3 Table of Coefficients for the first linear model, as well as the percentage of salience for each of the items

Item	Intercept		Pos/Neg		Salience (%)
q1	2.91***	(0.03)	-0.03	(0.05)	0.2
q2	3.38***	(0.03)	-0.1**	(0.04)	0.2
q5	3.65***	(0.03)	0.01	(0.04)	1.5
q10	2.5***	(0.03)	0.2***	(0.04)	5.0
q11	3.64***	(0.03)	-0.06	(0.05)	1.4
q15	3.26***	(0.03)	-0.18***	(0.04)	7.9
q16	2.38***	(0.03)	0.17***	(0.04)	0.9
q17	3.34***	(0.03)	0.01	(0.04)	7.9
q18	3.4***	(0.03)	-0.29***	(0.04)	7.3
q21	3.88***	(0.03)	-0.05	(0.03)	3.1
q22	3.52***	(0.03)	-0.02	(0.05)	19.4
q25	3.4***	(0.03)	-0.09**	(0.04)	18.3

Effect of Question Polarity

Turning then to the hypotheses, Table 3 shows the results for the first hypothesis. For this, I used a linear model with the response of the respondent as the dependent variable and the polarity as the independent (binary) variable. Note that the response scale runs from 1 (negative) to 5 (positive) and the polarity is either negative (0) or positive (1). The results show that the polarity only had a significant effect on the response for six of the twelve items. The effect is the largest for item 18 ("The loan system for students should be abolished/maintained") with -0.29. This means that if one would alter the wording of the item from "abolished" to "maintained" this would lead to respondents giving a more negative response. There are lower (but significant) values for items 15, 16 and 10, with coefficients of -0.18, 0.17 and 0.2. As with item 18, these values are small considering the scale. Besides, there is little consistency in the direction of the effect. In two cases (items 10 and 16) a positive wording instead of a negative one led to a more positive response. In the other four (items 2, 15, 18 and 25) a similar change would lead to a more negative response. Thus, not only do only some of the items show an effect of polarity, of those that do so the effects are small and of inconsistent direction. Taken together, this means that we can only partly confirm Hypothesis 1.

Table 4 Table of Coefficients from the linear models with the interaction term

Item	Intercept		Pos/Neg		Soph.		Pos/Neg x Soph.	
q1	2.88***	(0.16)	-0.39*	(0.23)	0	(0.03)	0.06	(0.04)
q2	2.99***	(0.15)	0.48**	(0.21)	0.06***	(0.02)	-0.09***	(0.03)
q5	3.55***	(0.15)	0.02	(0.21)	0.01	(0.02)	0	(0.03)
q10	2.86***	(0.16)	-0.05	(0.22)	-0.06**	(0.02)	0.04	(0.03)
q11	2.4***	(0.17)	0.21	(0.24)	0.19***	(0.03)	-0.04	(0.04)
q15	3.71***	(0.15)	-0.3	(0.21)	-0.07***	(0.02)	0.02	(0.03)
q16	2.3***	(0.14)	0.25	(0.19)	0.01	(0.02)	-0.01	(0.03)
q17	3.26***	(0.13)	0.03	(0.19)	0.01	(0.02)	0	(0.03)
q18	3.74***	(0.16)	-0.65***	(0.22)	-0.05**	(0.02)	0.06*	(0.03)
q21	4.09***	(0.11)	-0.19	(0.16)	-0.03*	(0.02)	0.02	(0.03)
q22	4.64***	(0.16)	0.07	(0.23)	-0.18***	(0.02)	-0.01	(0.04)
q25	2.3***	(0.15)	0.26	(0.21)	0.17***	(0.02)	-0.05*	(0.03)

Effect of Sophistication

We now turn to the second hypothesis. Here, we considered whether the effect of the polarity is lower when the sophistication of the respondent is higher. To see if this is the case, we first run a second linear model to see if we can expect an interaction between the two to begin with. Thus, in this model, we include both sophistication and its interaction with polarity. Table 4 shows the results for this. Here, we find a significant interaction for only three of the items: 2, 18 and 25. Thus, only in three cases is the effect of sophistication different for either the positive or negative version of the item.

Yet, to see how different, we have to look at the effect at various levels of sophistication. For this, we must look at the average marginal effects at representative cases (MERs). The representative cases here are all those cases that have one of the nine levels of political sophistication. The marginal effects are the contribution of the polarity to the response. Table 5 shows these marginal effects at each of the nine levels of political sophistication. As higher levels of sophistication should lead to a lower effect of the polarity, the marginal effects there should tend towards zero. In other words, at those levels, the polarity contributes little to nothing to the eventual response of the respondent.

Table 5 Table of Margins

Item	2 (Low)	3	4	5	6	7	8	9	10 (High)
q1	-0.28	-0.22	-0.17	-0.11	-0.05	0.00	0.06	0.11	0.17
q2	0.30	0.21	0.12	0.03	-0.06	-0.15	-0.24	-0.33	-0.42
q5	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.00	-0.00
q10	0.03	0.07	0.11	0.15	0.19	0.23	0.27	0.31	0.35
q11	0.13	0.09	0.05	0.01	-0.03	-0.07	-0.11	-0.15	-0.19
q15	-0.27	-0.25	-0.23	-0.21	-0.19	-0.17	-0.15	-0.13	-0.11
q16	0.22	0.21	0.20	0.19	0.17	0.16	0.15	0.14	0.13
q17	0.02	0.02	0.01	0.01	0.01	0.00	0.00	-0.00	-0.00
q18	-0.54	-0.48	-0.42	-0.37	-0.31	-0.25	-0.20	-0.14	-0.08
q21	-0.15	-0.12	-0.10	-0.08	-0.06	-0.04	-0.01	0.01	0.03
q22	0.05	0.03	0.02	0.01	-0.00	-0.01	-0.03	-0.04	-0.05
q25	0.16	0.10	0.05	-0.00	-0.06	-0.11	-0.17	-0.22	-0.27

From the Table, we see that in the case of items 2 and 25, the marginal effects run from positive to negative. Thus, at low levels of political sophistication, changing the item from negative to positive leads to a more positive response. On the other hand, those with a high level of political sophistication would provide a more negative response. Thus, not only does the effect not disappear at high levels, but it also even flips to behave opposite from expected. Even for item 18, where the negative effect does decrease, it remains negative even at the highest level.

The other items fare not much better. Items 5, 17 and 22 are closest to 0 (no effect) but are stable for all levels of political sophistication. Other items are equally stable and are either positive or negative for each case. Finally, item 10 shows the opposite of what we expect - that is, values that tend towards zero at the lower end of political sophistication. Note though that as before the size of all these effects is small (-0.54 being the largest for item 18). Taken together, this means that we cannot confirm the second hypothesis.

Effect of Issue Salience

As per Hypothesis 3, the effect of the polarity should be smaller for questions with a high salience than for those with a low salience. Looking at Table 3 we find that in the three cases where the effect is significant, the salience is rather low: (0.2% for

questions 1 and 2, and 7.3% for question 18). Also, for the question with the highest salience (question 22, with 19.4%) the effect is both small and insignificant. At the same time, another question with high significance (question 25) does have a high salience, at 18.3%. Moreover, there is no significant correlation between the salience and the effect ($r(10) = 0.02$, $p = 0.94$). As such, we have to reject the third hypothesis.

Effect on Congruence

The fourth hypothesis asked if question polarity can lead to actual differences in congruence. In other words: does changing the polarity influence the degree of how close or far a respondent is from a party? For this, I calculated how often a party appeared as the "best match" for a respondent. This best match is also that party that appears at the top of the list of matches the respondent receives after filling out the VAA. To then see if polarity made a difference, Table 6 shows the differences between both cases.

Here we see that not only are there differences between the various parties, but there are differences between the items as well. Also, when comparing these results with Table 3, we find that the effect of polarity is not always a good predictor for the change in congruence. For example, while item 18 showed the highest effect of polarity (at -0.29), the changes in the best match are average. Yet, item 5, which did not show any effect of polarity, shows a large "swing" for the GreenLeft. That is, when switching from the negative to the positive version of that item ("Mortgage relief has to be abolished/maintained"), the party lost 18.64% of its best matches, with their loss being equally distributed over the other parties. Something similar occurs for the PVV in the case of item 11, where they lose around a third (34.49%) of their best matches. While in both cases this loss is equally shared between the other parties, in other cases, it is clearer which parties' profit. For example, for item 25 (The multicultural society is a not good thing/is a good thing), where the PVV loses 26.86%, these losses benefit both CDA and CU in an equal manner.

On why certain parties gain and lose matches, the results are mixed. For item 25, we can argue that the cause of the losses of the PVV is that the party is often associated with the negative version of the item. The same goes for item 11 ("The Netherlands has to remain in/leave the European Union"), as the party is well known for favoring leaving the EU. Yet, in other cases, such as for the Green Left for item 5, or why the CDA and CU profit from the change for item 25, the results are less clear. Overall though, we can conclude that the polarity of the item does influence the congruence, though not always in a consistent manner. Thus, we can confirm the fourth hypothesis.

Table 6 Differences in the percentage of best match for each party, when switching from the negative to the positive version

	CDA	CU	D66	GL	PvdA	PVV	SP	VVD
q1	-2.21	-2.21	2.24	2.24	-2.32	2.24	2.24	-2.21
q2	5.86	5.86	-7.44	-7.44	5.86	-1.13	-7.44	5.86
q5	2.67	2.67	2.67	-18.64	2.67	2.67	2.67	2.67
q10	16.04	3.78	-6.78	-6.78	-15.85	16.04	-6.78	0.32
q11	4.72	4.72	4.72	4.72	4.72	-34.49	6.21	4.72
q15	1.24	1.86	4.22	1.86	4.22	-7.63	1.86	-7.63
q16	-9.05	11.28	-9.05	2.39	11.28	-9.05	11.28	-9.05
q17	6.63	-2.71	-2.71	-2.71	-5.88	6.63	-5.88	6.63
q18	-5.38	-5.38	3.81	3.81	3.81	1.25	-5.38	3.64
q21	1.16	-3.50	1.16	1.16	1.16	-3.50	1.16	1.16
q22	8.81	8.81	7.20	-8.01	-8.01	-8.01	-8.01	7.20
q25	13.48	13.48	-0.01	-0.01	-0.01	-26.86	-0.01	-0.06

Conclusion

One of the cornerstones of the field of issue voting are accurate measures of the congruence between voters and parties on issues. In this paper, I focused on one threat to the accuracy of these measures by looking at how the formulation of an issue can influence it. In other words, could one get alternative degrees of congruence by doing no more than altering the formulation of an issue? For this, I focused on a common alteration of the polarity of the issue, that is, whether an issue has a negative or positive formulation. After selecting the items, I provided two groups of respondents with either formulation using a Voting Advice Application, which launched during the 2017 elections in the Netherlands. This led to mixed results. While for some items there were significant differences between the responses for the two versions of the item, for others there were not. The same was true for the influence of political sophistication and issue salience: sometimes it was influenced by the effect of polarity, other times not. Yet, in all cases, the actual influence on the mean response was low. That is, even when the polarity led to a different mean response, this difference was very small (especially on a 1-5 scale). Yet, these small differences did often have a significant effect on the congruence between the party and the respondent. Thus, for some items, parties received a higher congruence with a respondent with a different formulation of an item.

These findings are interesting for several reasons. To begin with, that a different item wording can lead to different responses is well-known. Yet, what these differences are and whether they matter is much less studied. With regards to congruence, it is not only important to look at the actual change in the response, but also at the change in congruence. Here, we saw that even small differences in the mean response could lead to large differences in the congruence.

As such, these findings can find application in several fields. An obvious one is that of VAAs, which provide respondents with matches between them and various parties, based on the degree of congruence between them. Designers of VAAs should thus consider that not only can the initial selection of the items matter, but their formulation can also as well. Given the increasing evidence that VAAs can influence party choice, this makes it even more relevant for designers to pay close attention to the wording of the issues. Another complication for designers of VAAs is the conditional effect of political sophistication. In some cases, the effects of changing the polarity of the question were higher when the respondent had a lower level of political sophistication. This is a challenge for designers of VAAs as it is for those respondents that VAAs are most beneficial. As these findings show that they are vulnerable to such design choices, there should be an increased focus on these choices. This then leads to the question if they should or should not include negative questions at all. The problem here is that it is difficult *not* to include any negative questions at all. Some of the questions only exist in their negative form in the debate and including them in their positive form could be confusing for the respondents. Also, there is no reason to assume that negative questions are inherently problematic. They are only different from their positive counterparts. The best designers could do is at least to consider which questions to make negative and why. Besides, designers would do well not to use a certain wording if it would favor a certain party.

Apart from VAAs, we can also extend our conclusions to the measurement of congruence in general. In this case, the main conclusion is that congruence not only depends on the content of the item but also on its formulation, with the effect being influenced by the political sophistication of the respondent and the party one calculates the congruence with. Here we saw that while these differences may be small, their influence can be large. As such, ignoring the effect of the formulation of an item can lead one to draw conclusions based on measurement variation instead of substantive variation. Besides, changing the polarity of an item is a simple change. More rigorous changes - such as including or not including examples in the item - are likely to cause equal, or even larger, differences.

From here, there are several avenues for further research. The first one is to extend the current research to other countries. The country here – The Netherlands – is in many ways a unique case. The country has a representative system of government that is one of the most proportional in the world. As a result, there are a large

number of political parties (Lijphart, 1999). In such systems, parties are more likely to adopt issues as their own and stress them during the campaign (Kim, 2020). A second avenue is to carry out similar research, but also pay attention to the positions of the parties. In this study, the coders coded the party on only one version of the question. The position for the opposite wording was nothing less than the reverse of the score. Yet, given that parties might own the wording of certain issues it might be that the positioning of the party is different under different wording. This could explain at least some of the variations found here. A third avenue might be to change other aspects of the wording of the questions or the questionnaire. One example might be the effect of the number or the order of the response options the respondent can use. Another might be to what degree quantifiers or explanations show any influence. Fourth, one could repeat the experiment on dimensions instead of issues, as the wording of the dimensions is most likely affected in the same way.

That question wording is no neutral exercise is clear. Yet, the precise effects of it are often not clear. Here, I showed that even a simple aspect of question-wording could lead to changes in the size of the congruence between voters and parties. Moreover, parties can benefit when the question uses their favored wording. Thus, scholars working with congruence should take not only the effects of question-wording into account, but also realize that no wording can truly be neutral.

References

Andreadis, I. (2014). Data Quality and Data Cleaning. In D. Garzia and S. Marschall (Eds.), *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective* (pp. 79–92). Colchester: ECPR Press.

Bassili, J. N. & Krosnick, J. A. (2000). Do Strength-Related Attitude Properties Determine Susceptibility to Response Effects? New Evidence From Response Latency, Attitude Extremity, and Aggregate Indices. *Political Psychology 21* (1), 107–132.

Belchior, A. M. & Freire, A. (2013). Is party type relevant to an explanation of policy congruence? catch-all versus ideological parties in the Portuguese case. *International Political Science Review 34* (3), 273–288.

Benoit, K. & Laver, M. (2006). *Party Policy in Modern Democracies*. London: Routledge.

Blasius, J. & Thiessen, V. (2012). *Assessing the Quality of Survey Data*. London: Sage.

Chessa, A. G. & Holleman, V. (2007). Answering attitudinal questions: Modeling the response process underlying contrastive questions. *Applied Cognitive Psychology 21* (2), 203–225.

Clark, H. H. (1976). *Semantics and Comprehension*. Den Haag: Mouton.

Costello, R., Toshkov, D., Bos, B., & Krouwel, A. (2020). Congruence between voters and parties: The role of party-level issue salience. *European Journal of Political Research*.

Dalton, R. J. (1985). Political parties and political representation. *Comparative Political Studies 18* (3), 267–299.

Dalton, R. J. (2017). Party representation across multiple issue dimensions. *Party Politics 23* (6), 609–622.

Diamond, L. & Morlino, L. (2005). *Assessing the Quality of Democracy*. Baltimore, MD: The Johns Hopkins University Press.

Dinas, E. & Gemenis, V. (2010). Measuring Parties´Ideological Positions With Manifesto Data: A Critical Evaluation of the Competing Methods. *Party Politics 16* (4), 427–450.

Dolný, B. & Baboš, P. (2015). Voter–representative congruence in Europe: A loss of institutional influence? *West European Politics 38* (6), 1274–1304.

Downs, A. (1957). *An Economic Theory of Democracy*. New York, NY: Harper.

Gemenis, K. (2015). An iterative expert survey approach for estimating parties' policy positions. *Quality & Quantity 49* (6), 2291–2306.

Gemenis, K. B., Bruinsma, C., Djouvas, V., Manavopoulos, & Mendez, F. (2017). *StemConsult: Voting Advice Application data for the 2017 parliamentary election in the Netherlands*. DANS. https://doi.org/10.17026/dans-24r-225b

Gemenis, K. & van Ham, C. (2014). Comparing Methods for Estimating Parties' Positions in Voting Advice Applications. In D. Garzia and S. Marschall (Eds.), *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective* (33–47). ECPR Press.

Giger, N. & Lefkofridi, Z. (2014). Salience-Based Congruence Between Parties & their Voters: The Swiss Case. *Swiss Political Science Review 20* (2), 287–304.

Highton, B. (2009). Revisiting the Relationship between Educational Attainment and Political Sophistication. *The Journal of Politics 71* (4), 1564–1576.

Hippler, H.-J. & Schwarz, N. (1986). Not Forbidding Isn't Allowing: The Cognitive Basis of the Forbid-Allow Asymmetry. *Public Opinion Quarterly 50* (1), 87–96.

Holleman, B., Kamoen, N., Krouwel, A., van de Pol, J., & de Vreese, C. (2016). Positive vs. Negative: The Impact of Question Polarity in Voting Advice Applications. *PLOS ONE 11* (10), 1–17.

Holleman, B., van den Bergh, H., Mak, P., Sanders, T. & Kamoen, N. (2017). Why Are Negative Questions Difficult to Answer? On the Processing of Linguistic Contrasts in Surveys. *Public Opinion Quarterly 81* (3), 613–635.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.

Hosch-Dayican, B., Aarts, K., Amrit, C., & Dassen, A. (2013). Issue salience and issue ownership online and offline: Comparing Twitter and survey data. Paper presented at the American Political Science Association 2013 Annual Meeting; 29 August - 01 September, Chicago, IL.

Huber, J. D. & Powell, G. B. (1994). Congruence Between Citizens and Policymakers in Two Visions of Liberal Democracy. *World Politics 46* (3), 291–326.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or Disagree? Cognitive Processes in Answering Contrastive Survey Questions. *Discourse Processes 48* (5), 355–385.

Katz, R. S. (1997). *Democracy and Elections*. Oxford: Oxford University Press.

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics 38* (7), 1033–1050.

Kim, Y. J. (2020). Issue Ownership and Electoral Rule: Empirical Evidence from Japan. *Government and Opposition 55* (1), 147–162.

Krosnick, J. A. & Presser, S. (2010). Question and Questionnaire Design. In P. V. Marsden and J. D. Wright (Eds.), *Handbook of Survey Research (Second ed.)* (pp. 263–313). Bingley: Emerald.

Lijphart, A. (1999). *Patterns of Democracy*. New Haven, CT: Yale University Press.

Louwerse, T. & Rosema, M. (2014). The design effects of Voting Advice Applications: Comparing methods of calculating matches. *Acta Politica 49* (3), 286–312.

Luskin, R. C. (1990). Explaining Political Sophistication. *Political Behavior 12* (4), 331–361.

Mair, P. (2001). Searching for the Positions of Political Actors: A Review of Approaches and a Critical Evaluation of Expert Surveys. In M. Laver (Ed.), *Estimating the Policy Positions of Political Actors* (pp. 10–30). London/New York, NY: Routledge.

Mendez, F. (2014). What's behind a matching algorithm? A critical assessment of how VAAs produce voting recommendations. In D. Garzia and S. Marschall (Eds.), *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective* (pp. 49–66). Colchester: ECPR Press.

Mendez, F. (2017). Modeling proximity and directional decisional logic: What can we learn from applying statistical learning techniques to VAA-generated data? *Journal of Elections, Public Opinion and Parties 27* (1), 31–55.

Mendez, F., Gemenis, K., & Djouvas, C. (2014). Methodological challenges in the analysis of voting advice application generated data. In 2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization, pp. 142–148.

Powell, G. B. (2004). Political Representation in Comparative Politics. *Annual Review of Political Science 7* (1), 273–296.

Powell, G. B. (2009). The ideological congruence controversy: The impact of alternative measures, data, and time periods on the effects of election rules. *Comparative Political Studies 42* (12), 1475–1497.

Rabinowitz, G. & MacDonald, S. E. (1989). A Directional Theory of Issue Voting. *The American Political Science Review 83* (1), 93–121.

Romeijn, J. (2020). Do political parties listen to the(ir) public? Public opinion-party linkage on specific policy issues. *Party Politics 26* (4), 426–436.

Rosset, J. & Stecker, C. (2019). How well are citizens represented by their governments? issue congruence and inequality in Europe. *European Political Science Review 11* (2), 145–160.

Schuldt, J. P., Konrath, S. H., & Schwarz, N. (2011). "Global warming" or "climate change"? Whether the planet is warming depends on question wording. *Public Opinion Quarterly 75* (1), 115–124.

Thomassen, J. (2012). The Blind Corner of Political Representation. *Representation 48* (1), 13–27.

van de Pol, J., Holleman, B., Kamoen, N., Krouwel, A., & de Vreese, C. (2014). Beyond Young, Highly Educated Males: A Typology of VAA Users. *Journal of Information Technology & Politics 11* (4), 397–411.

van der Eijk, C. (2001). Measuring Agreement in Ordered Rating Scales. *Quality & Quantity 35* (3), 325–341.

van der Meer, T. W. G., van der Kolk, H., & Rekker, R. (2017). Dutch Parliamentary Election Study 2017 (DPES/NKO 2017). DANS.

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

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

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

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

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

Abstract

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

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

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

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

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

Acknowledgements

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

Direct correspondence to

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

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

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

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

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

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

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

Literature Review

Socio-demographic Undercoverage Bias in Online Panels

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

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

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

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

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

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

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

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

Estimation of Survey Accuracy with Benchmarking

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

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

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

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

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

Methods

Data

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

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

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

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

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

Population, Samples, and Data Collection Modes

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

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

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

Data Processing and Analysis

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

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

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

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

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

particular statement, which further justifies the undercoverage bias calculation with dummies.

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

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon \quad (2)$$

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

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

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

For better statistical power, the Life in Australia™ ordinal variables were included in all models, both the ones for categorical variables (with *Cramer's V value* as the dependent variable) and for models with non-parametric effect sizes as dependent variables (with *Rank-Biserial Correlation coefficient*). Since R-BS coefficient values range from −1 to 1, and we were only interested in the magnitude

of effect sizes and not the direction, an absolute version of *R-BS coefficient* with positive values only was used.

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

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

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

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

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

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

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

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

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

Results

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

Undercoverage Bias – Extent of Univariate Bias

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

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

Table 2 Undercoverage bias in six Life in Australia™ waves

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

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

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

Undercoverage Bias – Survey Item Characteristics

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

(see 'Attitudinal, behavioral, and other factual differences between the online and offline populations' subsection of the Literature review), while also presenting the magnitude of those differences.

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

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

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

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

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

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

*p<0.05, **p<0.01

The R-BS models showed that the differences between onliners and offliners were captured the most prominently in *internet*, but also in *labor, employment, work,* and partially in *health, environment,* and *multiculturalism*. The topics with below-average differences were *finance* (in contrast to the Cramer's V model), *politics and elections,* and *government and policy.* Except for the *internet* and *government and policy* topics (and to some extent *international relations*), there were no observable trends – in some cases, weighting decreased bias in others it had

no effect; effect sizes differed substantially between Cramer's V and R-BS models for the same topics; topics with above and below-average effect sizes could not be grouped further into broader homogenous topics with more or less undercoverage bias.

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

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

Accuracy of Estimation – Benchmarking

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

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

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

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

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

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

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

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

Discussion and Recommendations

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

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

The findings of this research are based on data from one country only (Australia) and country-specific effects cannot be ruled out. The results indicate that inclusion of the offline population in probability-based online panel research seems to be, to some extent, unnecessary from the coverage error and accuracy perspectives. This could potentially be generalized to other developed countries with high internet penetration rates, narrower socio-economic and demographic distributions, and consequently, relatively minor differences between those with and without internet connection. At the very least, offliners could be temporarily excluded for certain topics which the current study identified as lesser predictors of differences between the populations, such as *household and family, government and policy,* or partially,

finance. On the other hand, it might be more prudent to think reversely - what items should never be included in probability-based online panel surveys if data are collected from an online sample only, e.g., *internet* or *international relations* items in Life in Australia™. However, overall, the current study observed differences across the majority of topics with no particular trends. This is in line with the findings of Rookey et al. (2008) and other authors who have reported differences for various topics (e.g., Blom et al., 2015; Bosnjak et al., 2013; Eckman, 2016; Keeter et al., 2015; Schaurer & Weiß, 2020; Zhang et al., 2009).

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

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

The current analyses were limited, to some extent, by the number of studied items and their characteristics. With a larger sample of items and variables with available benchmarks, possibly from questions related to different broad topics and with more continuous variables, future studies would have greater statistical power and better evidence for data-informed decision making. The current findings might have to be slightly adjusted in that case. This study presents a combined approach

to studying undercoverage bias and its effects on data accuracy, and as this was examined in the Australian context only, future research should focus on online-offline population differences in other countries. This is particularly pertinent in regions with both lower internet penetration rates and wider socio-economic and demographic distributions. Such studies could help establish how necessary mixing modes and inclusion of the offline population are in a particular country's context.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Pennay, D. W., Neiger, D., Lavrakas, P. J., & Borg, K. (2018). The Online Panels Benchmarking Study: a Total Survey Error comparison of findings from probability-based

surveys and nonprobability online panel surveys in Australia. *CSRM and SRC Methods Paper, 2018* (2).

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

Selection of Statistical Tests and Effect Size Measures

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

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

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

Data Processing and Effect Size Analysis

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

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

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

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

Outlier Detection Analysis

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

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

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

Table 5 Benchmarking data sources and nationally representative benchmarks

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

methods, data, analyses | Vol. 17(1), 2023, pp. 121-130 DOI: 10.12758/mda.2022.07

Different Approaches to Incorporate the Aspect of Practical Relevance in the Statistical Inferential Process

Andreas Quatember
Johannes Kepler University Linz, Austria

Abstract

In different scientific areas, empirical studies are typically carried out by statistical null hypothesis tests. Despite the long tradition of applications, misinterpretations and misuses of the concept have led to a substantial confidence crisis in its inferential quality. One of the discussed issues is the significance-relevance discrepancy of the results of standardly applied zero-effect null hypothesis tests. This means that statistically significant test results do not automatically also have to be of scientific relevance in the specific research context. Therefore, this article is aimed at practitioners of empirical research who might want to include the aspect of practical relevance in their statistical conclusions. Different approaches to include this aspect in the inferential process are discussed with an example from the field of educational research.

Keywords: Null hypothesis significance test; effect size; significance thresholds; relevance thresholds; statistical literacy

In areas such as the social, behavioral, or educational sciences; in economics; or in medicine, empirical studies are commonly carried out with the application of statistical null hypothesis significance tests. For many of these methods, R. A. Fisher provided the theory in his book, *Statistical Methods for Research Workers*, first published in 1925 (Fisher, [1925] 1990) and he described the general theoretical framework by a famous experimental setup, "The Lady Tasting Tea", which was published in his book, *The Design of Experiments*, in 1935 (Fisher, [1935] 1990). Despite this long history of applications of this technique from the field of inferential statistics, misinterpretations of its results and misuses of the procedure have led to a veritable confidence crisis with regard to its inferential quality (for instance, Greenland et al., 2016: 341; Wasserstein & Lazar, 2016: 129). Under these circumstances, the *American Statistical Association* (ASA) decided to publish a statement on statistical significance and *p*-values containing several broadly agreed upon principles underlying the proper use of this method of inferential statistics (Wasserstein & Lazar, 2016). Furthermore, the editors of *The American Statistician*, a journal published by the ASA, decided to dedicate a special issue of the journal to the topic. The contributions contained many ideas that were published to enable wider consideration and debate (Wasserstein, Schirm & Lazar, 2019).

One of the issues under discussion is the significance-relevance discrepancy (for an example, see Nuzzo, 2014: 151f). By this term, it is meant that so-called statistically significant test results do not automatically also have to be of practical (or scientific) relevance in the specific research context. But, empirical researchers "rarely distinguish between the statistical and the practical significance of their results. Or worse, results that are found to be statistically significant are interpreted as if they were practically meaningful" (Ellis, 2010: 4).

In this article, which is mainly intended to practitioners of empirical research, the approaches that incorporate the aspect of practical importance of survey results in the statistical inferential process are described as a contribution to this debate. For this purpose, a research question from the field of educational sciences will serve as an explanatory example. Section 2 addresses the difficulty of the specification of the thresholds, which have the task to separate the practically important from the nonimportant test results. Section 3 discusses different concepts of the consideration of their practical importance. The concluding fourth section summarizes the aspects of the significance-relevance discrepancy.

Direct correspondence to
Andreas Quatember, Institute of Applied Statistics, Johannes Kepler University JKU Linz, Science Park 2, Altenberger Str. 69, A-4020 Linz, Austria, www.jku.at/ifas
E-mail: andreas.quatember@jku.at

The Aspect of Practical Relevance

Throughout the article, the following research question from the field of educational sciences will serve as the explanatory example, from which similar considerations can be derived for other study questions: Are the obtained test results of the students of country A in an interesting competence area better than the results of the students of country B? Based on this research question, the null hypothesis H_0 and the alternative hypothesis H_1 for the statistical null hypothesis significance test of the difference $\delta = \mu_A - \mu_B$ between the true mean values, μ_A and μ_B, of the countries' students are formulated as follows:

$$H_0: \delta \leq 0 \text{ and } H_1: \delta > 0$$

Only with a full survey of the students in both countries, it would have been possible to make a definitely correct decision between these two hypotheses.

However, is really each difference $\delta > 0$ practically relevant? In other words, is really each "effect" (of different school systems, forms of teaching, etc.) larger than zero practically meaningful? There cannot be given a general answer to this question because the answer completely depends on the research context. In any case, this aspect also occurs with population surveys. But if not all effects $\delta > 0$ are of practical importance, the next question that automatically arises is: How big an effect δ in the specific scientific context has to be in order to be of practical importance?

In the specific scientific context, different approaches can lead to the determination of a certain relevance threshold, which shall separate the nonrelevant from the relevant effects. First, such a threshold may be directly derived from the given research question (*research-driven approach*). In our example, the actual research question under investigation may be that the difference δ of the mean values of the two groups became larger compared to the difference δ_0 in a previous population survey. Accordingly, the derived relevance-threshold δ_R of the difference δ should be set at δ_0.

Second, there may be a consensus about those effect sizes that are of practical importance (*expertise-driven approach*). In our example from the field of student assessment, experts may, for instance, agree on a certain relevance-threshold δ_R with regard to the difference δ.

Third, a convention might be applied with respect to the calculation of a reasonable relevance threshold (*convention-driven approach*). In his milestone book in the field of behavioral sciences, Cohen (1969), for instance, expresses relevant effect sizes in units of the variability of the variable under study. For population differences δ (with the known standard deviation σ of the variable under study assumed to be equal in both populations), he specifies a relevance threshold

- of $\delta_R = 0.2 \cdot \sigma$ for the search for an at least small,
- of $0.5 \cdot \sigma$ for the search of an at least medium, and
- $0.8 \cdot \sigma$ for the investigation of a large effect (Cohen, 1969: Section 2.2).

For our example, assuming that a relevant effect has at least to be a small one, one can use the pooled estimated standard deviation from the last survey to determine the corresponding convention-driven δ_R.

Of course, because such a relevance threshold is a continuous quantity, one can object that there is no content-related reason that test statistics being only a little bit smaller or larger, respectively, than δ_R shall be differently interpreted with respect to its practical meaningfulness. However, one can argue against this that there are countless other examples for the usefulness of such arbitrary limits in everyday life. Just think, for instance, in medicine of the categorizations of the total cholesterol level of adults. Values of less than 200 mg/dL are "considered desirable", values from 200 to 240 mg/dL are called "borderline high", and those more than 240 mg/dL are called "high". Depending on the category in which a person belongs, different therapeutic measures are recommended (MNT, 2021). Other examples include the thresholds of the risk of poverty in official statistics, the legal limit of blood alcohol for driving a car, the permissible fine dust pollution in a city, or also the significance level α of a statistical null hypothesis test (see for its history: Cowles & Davis, 1982). In all of these examples, there is no reasonable justification for the strict categorizations except for one: They are all undeniably pragmatic with regard to the objectivity of the criteria for decisions derived from them.

Clearly, the specification of such relevance thresholds is crucial when the practical meaningfulness of test results shall be included in the inferential process. If it is not at all possible to fix such a threshold before the investigation, then it will also not be possible afterward to assess the practical importance of the test statistic. Assuming that such a threshold can be determined, the next question is naturally: How can the aspect of practical meaningfulness of a result be incorporated in the statistical inferential strategy?

A Marriage Between Statistical Significance and Practical Relevance

In the practice of empirical research, independently of any research context, the null hypothesis postulates the complete absence of an effect as a rule. The impact of the implementation of such a "zero-effect null hypothesis" H_0 is that with increasing sample size even for very tiny, practically irrelevant effects larger than zero, the probability of the, then, correct rejection of H_0, which is the test power, increases.

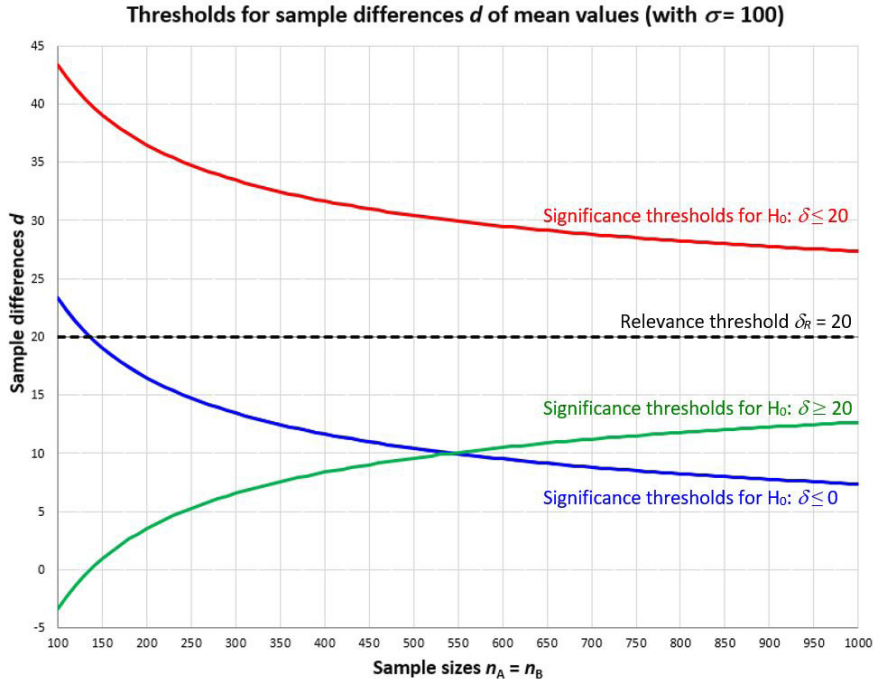


Figure 1 Thresholds for sample differences d of mean values for the different approaches to the incorporation of the aspect of practical relevance.

This is particularly problematic in the big data context of survey statistics (Meng, 2018).

In Figure 1, for the two sample t-test from our example with the test statistic $t = d/\sigma_d$ (with the difference d of the two sample means and the standard deviation σ_d of d) under simple random sampling with replacement, among other things, the (upper) limits for d, which separate the weak from the strong indicators against H_0: $\delta \leq 0$ at the significance level $\alpha = 0.05$, are exemplarily shown for an assumed $\sigma = 100$ and varying equal sample sizes n_A and n_B in the range from 100 to 1,000 (blue curve). For $n_A = n_B = 750$, for example, the limit between the significant and the non-significant test results is approximately $d = 8.5$. But, is, for instance, a sample difference $d = 10$, which in this case does speak statistically significantly ($p < 0.05$) against H_0: $\delta \leq 0$, really of practical importance in the contextual background? Based on the convention-driven approach from the previous section, for example, the relevance-threshold could be specified by $\delta_R = 0{,}2 \cdot \sigma = 20$ (dashed line in Figure 1). In this case, as an estimate of the true effect δ the survey result $d = 10$ would not indicate the presence of practical relevance because it is below the dashed line. For $n_A = n_B = 100$, a result of $d = 22$, which is below the blue curve in

Figure 1, would not be statistically significant, but at the same time, it would indicate a practical relevance because it is above the dashed line.

On the one hand, the standardly applied, context-unrelated formulation of a zero-effect null hypothesis does not at all consider a context-related relevant effect-size threshold. On the other hand, the categorization of a test statistic based solely on such a context-related relevance threshold without testing also for statistical significance would not at all take into account the sample fluctuation of the test statistic.

Goodman, Spruill, and Komaroff (2019) suggested a combination of these two approaches. In the hybrid method of "decision by minimum effect size plus *p*-value" (Goodman, Spruill, & Komaroff, 2019: 171f), the zero-effect null hypothesis is rejected only in cases where the test statistic's *p*-value is not larger than the significance level α, and at the same time, the test statistic itself is larger than a minimum practically meaningful effect. In Figure 1, such results d lie above the blue curve as well as the dashed line. Compared to the standardly applied zero-effect null hypothesis test, this concept incorporates also the practical relevance of the statistically significant results. However, it must be noted that it only takes account of the sampling error with respect to the null hypothesis of the complete absence of an effect and not with respect to the relevance threshold.

If the research aim is not to check whether there is a relevant effect, but rather whether there is no effect at all, a certain type of statistical significance testing, the so-called "equivalence tests", is suggested (see, for instance, Ramert & Westphal, 2020). In the field of pharmacokinetics, for example, researchers sometimes want to show the non-inferiority of a new cheaper drug compared to an established one (Lakens, 2017). In the statistical inferential process, the alternative hypothesis H_1 always serves as the statistical translation of the research hypothesis. Therefore, in this case, it consists of the range of parameter values that support the equivalence-hypothesis, whereas the null hypothesis H_0 consists of the range of values that do not. Consequently, the null hypothesis H_0 comprises, for instance, all differences δ that are equal or larger than a relevance (or non-equivalence) threshold δ_R:

$$H_0\text{: } \delta \geq \delta_R \text{ and } H_1\text{: } \delta < \delta_R$$

However, this approach should not be applied to research questions that are intended to test the opposite, namely the existence of a practically relevant effect. A look at Figure 1 illustrates the problem. The green curve marks the (lower) thresholds of statistically significant sample differences d with respect to the equivalence test with H_0: $\delta \geq 20$. A sample difference d, which is above this green curve but below the dashed line of $\delta_R = 20$ (like, for example, $d = 0$ for $n_A = n_B = 100$), indicates on the one hand that the null hypothesis of the existence of a relevant effect cannot be rejected when the sample fluctuation of the test statistic under the actual null

hypothesis is taken into account, but on the other hand, as an estimator of the effect size δ, it clearly indicates that there is no relevant effect.

In Fisher's framework, it is crucial that the statistical hypotheses of the test are formulated in such a way that it is really tested what is wanted to be tested. In practice, far too often these hypotheses are not the correct translations of the research questions, when zero-effect null hypothesis tests are standardly performed. If in our example from the field of educational sciences it is to be checked whether there is a statistically significant and at the same time practically relevant positive difference δ between the means in two countries, H_0 must contain all effect sizes δ that are considered as not practically important. Hence, the statistical hypotheses would have to be

$$H_0: \delta \leq \delta_R \text{ and } H_1: \delta > \delta_R \, .$$

This approach leads from a standardly applied zero-effect significance test, which completely ignores the research context, to a context-related statistical significance test for a practically relevant effect. Only if δ_R actually equals zero because even the tiniest effects are scientifically meaningful in the specific research context, this strategy corresponds to a zero-effect significance test.

With these hypotheses, a p-value of a relevant test statistic, which is not larger than the significance level α, signifies that the observed data are unlikely under the null hypothesis of no practically relevant parameter value. Consequently, a statistically significant result is automatically interpreted as being also of practical importance. Furthermore, in the case of $\delta_R > 0$, in contrast to the standardly applied zero effect test with $H_0: \delta \leq 0$, by an increase of the test power, the probability of the detection of a tiny but practically meaningless effect converges to zero.

For our example, the appropriate test statistic is given by $t = (d–\delta_R)/\sigma_d$. From this test statistic, the upper limits for d, which separate the weak from the strong indicators against $H_0: \delta \leq \delta_R$ at the significance level $\alpha = 0.05$, can be derived. In Figure 1, these are shown for $\delta_R = 20$ for different sample sizes $n_A = n_B$ by the red curve. Hence, sample differences d from the area above are considered to speak statistically significant against this null hypothesis of no relevant effect.

For the implementation of this conceptual shift from the standardly applied context-ignoring zero-effect null-hypothesis significance test toward a content-driven significance test for a practically relevant effect, for the investigation of a statistical parameter, the appropriate test statistic and its sample distribution under the null hypothesis have to be considered. This may require that users apply a test statistic that is unfamiliar to them.

Summary

Results from null hypothesis significance tests are interpreted as not enough indication or as strong indication against the null hypothesis, whatever this hypothesis was formulated. The significance-relevance discrepancy of test results only exists if the research hypotheses are not correctly translated into the statistical hypotheses. For this purpose, relevance thresholds have to be specified with respect to the parameters under study. This can be done in the given scientific context, based directly on the research question, on the basis of the expertise of an experienced researcher, or on conventions. Taking into account the relevance of test results, besides other approaches to incorporate the aspect of scientific relevance in the inferential process, statistical significance tests for a practically relevant effect can be performed. These have the advantage to be applicable within the traditional framework of statistical null hypothesis significance tests. Such tests consider the scientific importance of the test results and, at the same time, their sample fluctuation under the actual null hypothesis. For their application, possibly unfamiliar, but known appropriate test statistics and their sample distributions are to be used. Consequently, by making the experiment more accurate, for example, by a larger sample size, the increased test power does not lead to practically irrelevant, statistically significant results anymore.

References

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cowles, M., & Davis, C. (1982). On the Origins of the .05 Level of Statistical Significance. *American Psychologist*, 5, 553-558. doi:10.1037/0003-066X.37.5.553

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes*. Cambridge: Cambridge University Press.

Fisher, R. A. (1990). *Statistical Methods, Experimental Design, and Scientific Inference. A Re-Issue of Statistical Methods for Research Workers, The Design of Experiments, and Statistical Methods and Scientific Inference*. Oxford: Oxford University Press.

Goodman, W. M., Spruill, S. E., & Komaroff, E. (2019). A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting its Use. *The American Statistician*, 73(1), 168-185. doi: 10.1080/00031305.2018.1564697

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical Tests, *P* Values, Confidence Intervals, and Power: a Guide to Misinterpretations. *European Journal of Epidemiology*, 31, 337-350. doi: 10.1007/s10654-016-0149-3

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analysis. *Social Psychological and Personality Science*, 8(4), 355-362. doi: 10.1177/1948550617697177

Meng, X. L. (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. *The Annals of Applied Statistics*, 12(2), 685-726.

MNT (2021). What should my Cholesterol Level be at my Age? MedicalNewsToday Newsletter. Retrieved December 02, 2021, from the website: https://www.medicalnewstoday.com/articles/315900#treatment-options

Nuzzo, R. (2014). Statistical Errors. *Nature*, 506, 150-152. doi: 10.1038/506150a

Ramert, A., & Westphal, E. (2020). Equivalence Testing. STAT Center of Excellence, STAT COE-Report-12-2020. Retrieved December 02, 2021, from the website: https://www.afit.edu/stat/statcoe_files/1005AFIT2020ENS09117%201005rame%202-2.pdf

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, 70(2), 129-133. doi: 10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$". The American Statistician, 73(1), 1-19. doi: 10.1080/00031305.2019.1583913

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)