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

Murray-Watters, Alexander; Zins, Stefan; Silber, Henning; Gummer, Tobias; Lechner, Clemens

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

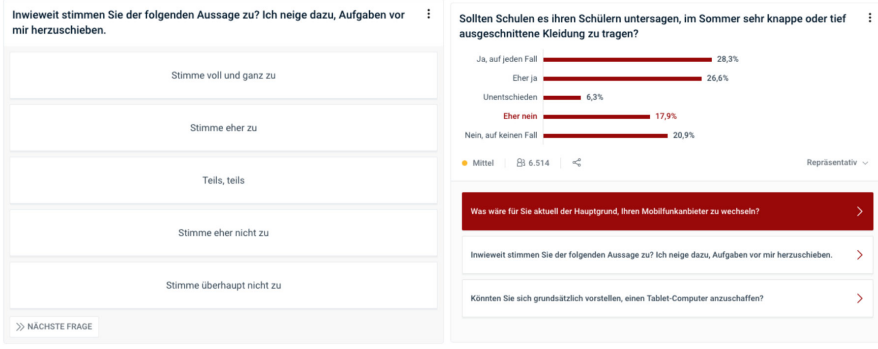


Figure 1 Examples of a single item questionnaire with follow-up questions (River Sample)

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

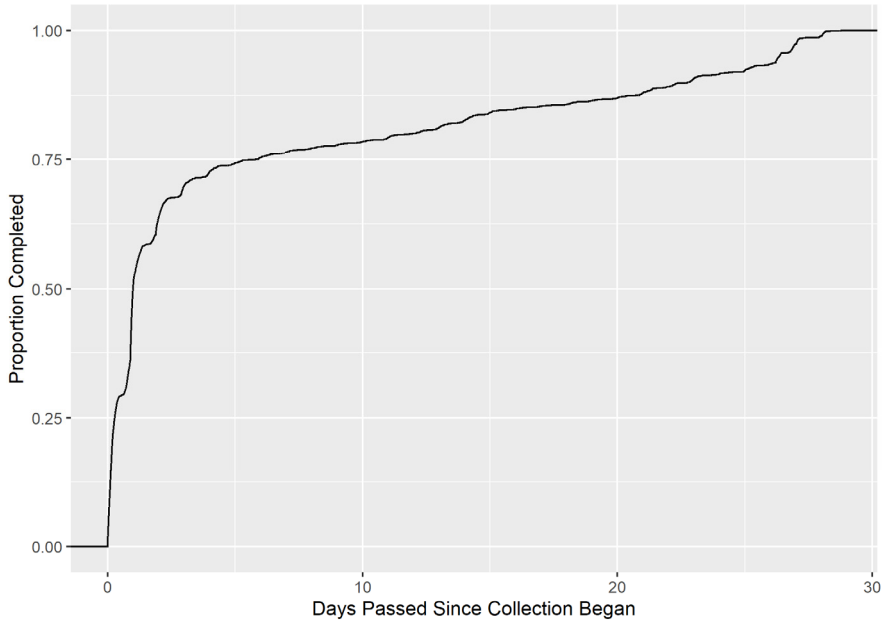
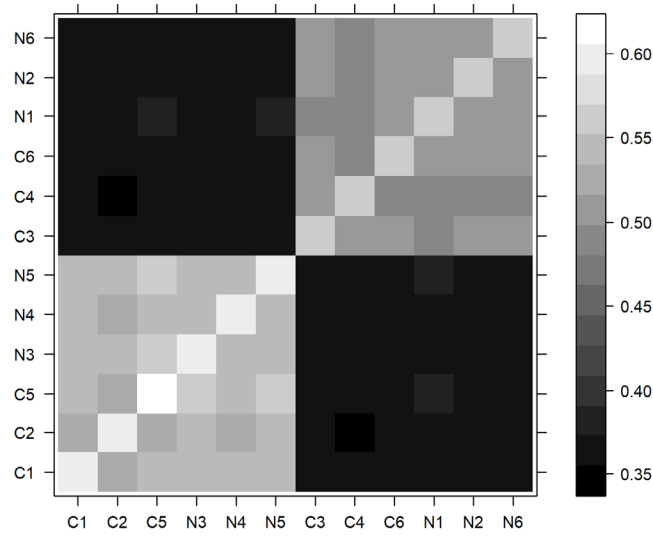


Figure 2 Empirical cumulative distribution of complete cases over time.

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

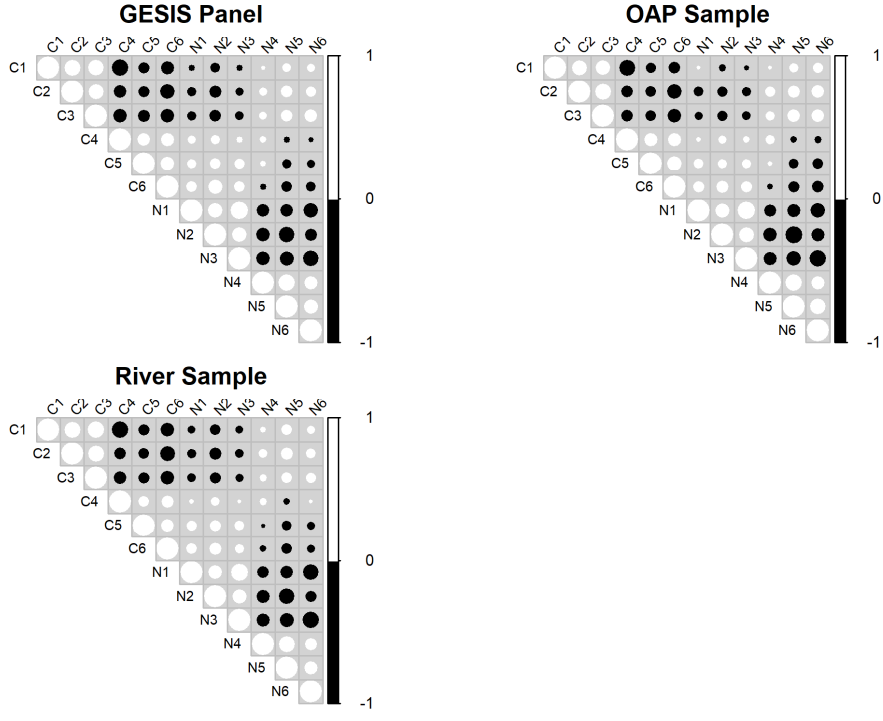


Figure 4 A graphical depiction of correlation matrices for the 12 items of our measurement instrument

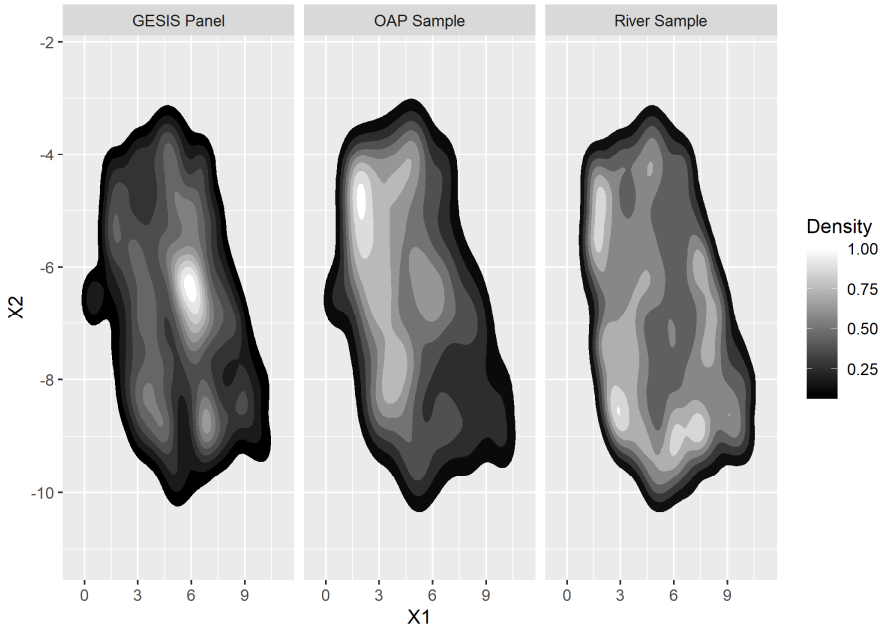
samples is more similar to the probability sample than the other. The complete correlation matrices can be found in the Online Appendix (see Tables A.5.1 - A.5.3).

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.

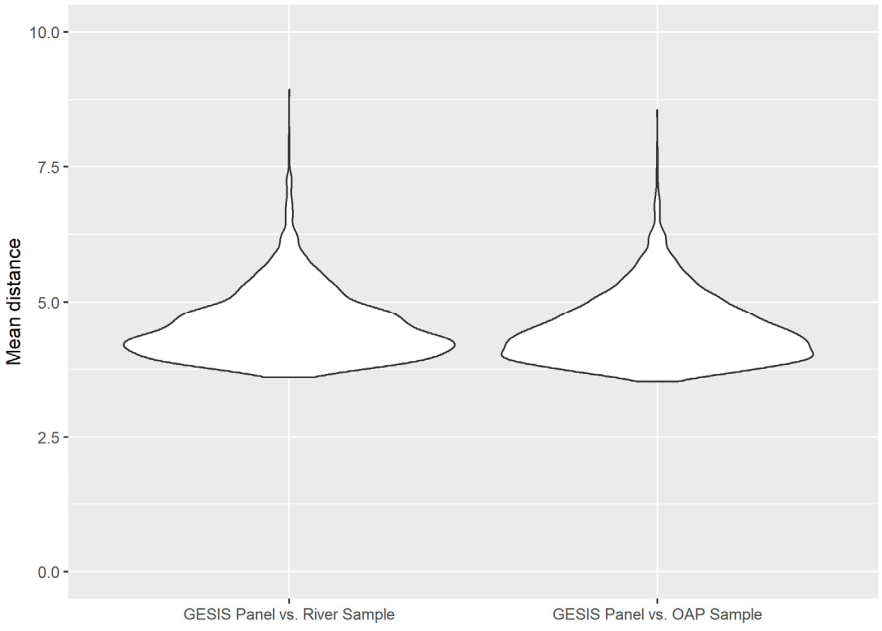


Figure 6 Violin plot of distance between GESIS Panel and non-probability samples

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.