# When are non-probability surveys fit for my purpose?

Björn Rohr
GESIS – Leibniz Institute for the Social Sciences

Barbara Felderer
GESIS – Leibniz Institute for the Social Sciences

Henning Silber
GESIS – Leibniz Institute for the Social Sciences

Jessica Daikeler
GESIS – Leibniz Institute for the Social Sciences

Joss Roßmann
GESIS – Leibniz Institute for the Social Sciences

Jette Schröder
GESIS – Leibniz Institute for the Social Sciences

Publication date: December 1, 2024

*The affordability of non-probability surveys, especially online surveys, has made them a popular alternative to probability surveys in social science research. However, their quality is questionable, and inferences based on non-probability surveys rely on strong, hard-to-test, and often unrealistic assumptions, such as that every population unit has a chance to participate in the survey. Commonly applied tools to improve accuracy, like adjustment weighting or the use of quotas, are very unlikely able to eliminate or sufficiently reduce the (self-) selection bias in non-probability surveys, as it is hard to identify and measure all variables needed. The use of non-probability surveys should thus be carefully considered and limited to specific (non-inferential) purposes. These include, among others, exploratory studies where generating hypotheses is the primary aim, survey experiments where internal validity is prioritized over external representativeness, and studies targeting hard-to-reach populations for which probability surveys are impractical or even impossible to conduct. This survey guideline aims to provide guidance for researchers to critically assess whether non-probability surveys are fit-for-purpose for their own research projects and to evaluate whether the use of non-probability surveys aligns with the specific goals and constraints of the studies at hand.*

## 1 Introduction

In recent years, non-probability surveys have risen in popularity in empirical social science, which is not least due to the increasing use of the Internet as a means of conducting surveys. Although non-probability surveys are not tied to the online mode, it is particularly easy to quickly survey many people online. Despite their various limitations, non-probability (online) surveys are frequently used because of their low cost per interview and shorter field periods compared to most probability-based surveys. This guideline aims to provide an overview of the limitations associated with non-probability surveys and helps readers to make informed decisions on when and how to use them appropriately. We start by outlining fundamental principles of probability surveys,

followed by a definition of non-probability surveys accompanied by examples. Next, we discuss various problems that are inherent to non-probability surveys. We then consider situations in which non-probability surveys might be fit for a specific purpose. Finally, we provide a summary as well as warnings for conducting or using non-probability surveys.

# 2 What are non-probability surveys?

In this section, we first give some common examples of non-probability surveys. We then give an initial overview of the characteristics of probability surveys and conclude with the differences from non-probability surveys, as non-probability surveys are best defined by distinguishing them from probability surveys.

## 2.1 2.1 Examples of non-probability surveys

The following section provides an overview of often-used types of non-probability surveys in social science research. The typology is mainly based on the AAPOR task-force report by Baker et al. (2013) on "Non-Probability Sampling."

### 2.1.1 Volunteer Sampling

In Volunteer Sampling, advertisements for the survey are fielded on web pages, in newspapers, or by using similar tools, and interested individuals select themselves into the set of respondents (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013; Statistics Canada, 2021, p.34). Individuals who are interested in participating are invited to register on a survey's website or directly to participate in a survey voluntarily.

Volunteer Sampling can be conducted in any survey mode. It is frequently used to recruit members of *opt-in* or *online access panels* (Callegaro et al., 2014; Statistics Canada, 2021, p. 35). Registered respondents' information is stored by the survey agency, and this list is used as some kind of replacement for a sampling frame to invite them to additional surveys in the future. However, this replacement of a sampling frame is based on an arbitrary selection and may differ systematically from the target population in many characteristics. A second application of Volunteer Sampling is *crowd-sourcing* (e.g., Amazon Mechanical Turk), where registered individuals are paid to complete different tasks, such as participating in a survey. Crowd-sourcing is different from an opt-in panel in that the volunteers are not mainly recruited on the platform to participate in surveys but to do any possible online tasks, one of which could be filling in a survey. Another practical example of volunteer samples is *student samples*, where surveys are conducted at universities with university students who are recruited online or in person. Volunteer sampling can also be conducted in malls or other public places where interviewers approach passers-by and ask them to participate in the survey.

### 2.1.2 River Sampling

River Sampling is a form of Volunteer Sampling that is exclusively conducted online. Advertisements are fielded during visits to websites or when conducting other online activities. For example, river samples may be advertised in apps or on social media. Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau (2013, p.18) describe river samples as opt-in web samples that intercept possible respondents online, doing other things (Baker et al., 2010, p. 14). Potential participants are often approached via banners or pop-ups to invite them to participate in the survey. Callegaro, Baker, Bethlehem, Goritz, Krosnick, & Lavrakas (2014) differentiate river samples from access panels: In contrast to non-probability panels, the advertisement leads directly to a survey without any registration step. Initial screening questions can still be implemented to allow quota sampling. The next time the same respondents see a similar ad, it may lead them to a different survey.

*Social media sampling* is a common type of River Sampling. Previously gathered information on the users by the social media platform may be used to send survey advertisements specifically to target population members.[1]

### 2.1.3 Snowball Sampling and Respondent Driven Sampling

Snowball Sampling is another method to generate a non-probability survey that is particularly suitable to sample persons of hard-to-reach societal groups (Bacher et al., 2019; Lohr, 2022; Statistics Canada, 2021, p.35), such as minorities or respondents with special health conditions. Here, an initial sample is collected using either non-probability or probability methods. In the following step, the respondents of the survey themselves hint at or invite additional respondents. Snowball Sampling is often used by asking respondents of hard-to-reach populations (e.g., refugees) to provide contact information of other members of the same group. One follows this approach several times until the desired number of respondents is reached. Respondent Driven Sampling is a method very similar to Snowball Sampling, with the difference that respondents (seeds) invite additional respondents with the help of coupons or special links. This allows researchers for every further respondent to know who invited them (Lee et al., 2017). When certain assumptions are met, it is possible that Respondent Driven Sampling can lead to a probability survey (Lee, Suzer-Gurtekin, Wagner, & Valliant, 2017). However, one of these assumptions requires long enough chains of survey invitations, which are often not given due to selective nonresponse, leading to some chains ending far too early (Gile et al., 2015; Lee, Suzer-Gurtekin, Wagner, & Valliant, 2017). Additionally, for both Snowball Sampling and Respondent Driven Sampling, it may make a difference whether they start using a probability or non-probability survey (Goodman, 2011), although, even when starting with a probability survey, it likely leads to a non-probability survey.

[1] See the survey guideline by Pötzschke et al. (2023) for more detailed information on social media sampling.

## 2.2 2.2 Characteristics of probability surveys

A probability survey is based on a random sample of units from the target population. The *target population* is defined as "the complete collection of observations we want to study" (Lohr, 2022, p.4). This could be, for example, all German citizens living in private households that are older than 17 years. If the target population is large, surveying the entire target population would be unfeasible within a given time frame, too expensive or inefficient, so usually only a sample of the units is selected. A *sample* is a subset of the target population. The units of the sample are studied to infer the target population. In some cases, lists of units of the target population are available from which the sample can be drawn. These *sampling frames* might be lists of pupils in a school or the residents of a specific city from official registers. In other cases, sampling is based on a small list of randomly selected addresses, which are a starting point, to conduct random routes sampling (Bauer, 2016). Random routes selection can either be done by generating a list of addresses that is afterward handed over to interviewers to contact the residents, or by the interviewers in the same step as the contacting. The *inclusion probability* is a target population unit's probability of being selected for a specific sample (Cochran, 1977). For a sample to be called a *probability sample*, units are selected randomly from the target population. As a crucial prerequisite for probability surveys, every unit in the target population must have a positive and known inclusion probability. If this is not fulfilled, inferences from the sample to the target population are not valid.

In the simplest form of a probability sample, a *simple random sample*, the inclusion probability is equal for every unit of the target population. Inclusion probabilities in complex random samples are non-zero and known but not equal, which must be accounted for by applying design weights to draw valid inferences. It is usually not possible to conduct an interview with every unit drawn so that the final sample of actual respondents (the net sample) differs from the sample originally drawn (the gross sample). Both can lead to errors in the inference from probability surveys. It is important to recognize possible errors in order to minimize errors when inferring from surveys to the population. The representation side of Groves' Total Survey Error framework [TSE; Groves & Lyberg (2010)] makes it possible to systematically consider errors in sampling and the realization of the net sample - i.e., the representation of the population. In the following section, we will go through the individual sources of error to discuss whether they are applicable to non-probability surveys later on.

First, coverage error is the source of error that stems from differences between the population and the sample frame. For example, if a sample frame (from which the sample is drawn) does not cover the whole population, meaning that some members of the population have zero probability to be selected into the sample (undercoverage) this is a form of coverage error. The opposite can also be true if the sample frame covers persons that are not part of the target population (overcoverage).

Second, sampling error (Groves et al., 2009) describes errors that arise when the sample is drawn from the sampling frame. In some instances, the sample may differ from the target population by design. For example, it could be that one region is oversampled to guarantee a higher sample size needed for statistical analyses. If the sample is not a simple random sample, for example, because

respondents have unequal probabilities of being selected into the sample, the statistical analysis needs to take that into account, for instance, by applying design weights.

The third error type of the TSE is called nonresponse error. Nonrespondents are sampled units who are eligible for participation but did not participate in the survey for any reason (e.g., non-contact, refusal). The amount of nonrespondents in relation to the number of eligible sample units is often quantified in terms of a response rate (RR[2]).

All of the types of errors can turn into bias, if they are systematic, meaning that some individuals have a higher propensity than others to be covered on the sampling frame, drawn for the sample, and/or to participate in a survey (Groves, 2006; Rubin, 1976). If the reasons for not responding to a survey are not associated with the variables of interest (i.e., the error is not systematic for the variable of interest), the *separate cause model* holds (Groves, 2006), and no nonresponse bias is to be expected for this variable. However, when the same characteristics are associated with responding to a survey and the variable of interest, the *common cause model* holds. In this situation, nonresponse bias in the estimation of the variable of interest is very likely but may be reduced by adjustment weights that incorporate the characteristics that cause nonresponse and relate to the variable of interest. If the propensity to respond depends on the variable of interest itself, the *survey variable cause model* is true, which leads to nonresponse bias in estimation that can not be eliminated by any adjustment method.[3]

To address nonresponse bias in probability surveys, there are well-established *adjustment weighting methods*. One method would be to derive adjustment weights that make use of benchmark data, which is an essentially error-free external data source (e.g., Census data), and weight the survey respondents in a way that the distribution of survey respondents matches the distribution of the target population for all variables that are available from the benchmark data and the survey data. These variables are usually called auxiliary variables that are usually not equal to the variables of interest. Another method would be to use auxiliary variables that are available for the survey respondents and nonrespondents to model the nonresponse process and derive adjustment weights. Both methods need auxiliary variables that are highly correlated to the variables of interest and the nonresponse mechanism. To be effective, nonresponse weights ideally need to incorporate all variables that are related to both the nonresponse and the variable of interest. In addition, the nonresponse must not depend on the variable of interest itself.[4]

## 2.3 2.3 Characteristics of non-probability surveys

The most important difference between probability and non-probability surveys is that in the latter the requirements for inclusion probabilities are violated. In the absence of a proper sampling frame, the probability of inclusion is unknown for the members of the stated target population, and large parts of the target population have a zero probability of inclusion (Callegaro et al., 2015). If an online survey is, for example, advertised on a news website, it can not be known which

[2]See the survey guideline by Stadtmüller et al. (2019) for more details on response rate calculation

[3]More detailed information about the concept and analysis of nonresponse bias can be found in the survey guidelines by Koch & Blohm (2015) and Felderer (2024).

[4]More information on weighting can be found in the survey guidelines by Gabler et al. (2016) and Sand (2020).

members of the target population have a chance of seeing the advertisement and how high their propensity to see it is.

In non-probability surveys, the process of inclusion is unknown, as there is no sampling frame and thus no sampling stage. The inclusion is often highly (self-) selective and dependent on the characteristics of the participants, meaning that participants with some characteristics might be overrepresented in the survey, while others might be underrepresented or even completely missing. One example would be a survey of the German population, which is recruited only online, e.g., in an online access panel, underrepresenting respondents less familiar with the internet. Lacking a proper sampling process with a well-defined sampling frame, the error types of the TSE, such as coverage error, sampling error, and nonresponse error, cannot be distinguished, quantified, nor properly corrected by traditional adjustment procedures. Moreover, bias caused by the survey invitation process and self-selection, for example, getting access to the survey or refusing to participate in the survey, are impossible to disentangle. For instance, consider, a river sample recruited via Facebook advertisements. Population members who do not use Facebook can not participate in the survey, which can cause coverage bias if the target population is not defined as Facebook users. Facebook users who are shown the advertisement but decide not to participate might introduce selection bias. It is impossible to disentangle these biases because only information is available on survey participants and not on invited Facebook users. In addition, for many non-probability surveys, response rates are impossible to calculate. Even in cases in which the exact number of survey invitations as well as the number of refused invitations is known, for example, if members of an access panel are invited to participate in a specific survey, the unknowable selection probabilities in the panel recruitment steps make them an uninformative indicator of nonresponse (The American Association for Public Opinion Research, 2023, p.78). The high amount of selectivity when recruiting for a non-probability survey makes it highly likely for survey variables to be systematically different from the population value due to (self-) selection bias. In addition, it is very unlikely to know in advance and observe all variables needed to correct for selection bias caused by potentially multiple and largely unknown mechanisms. The success of adjustment weights to correct for this potentially big amount of self-selection bias in the estimation is not guaranteed and is rather doubtful, as we will discuss in Section 3.1.

# 3 Why non-probability surveys cannot be used to infer from the survey to the target population

## 3.1 3.1. Inference from non-probability surveys rely on (too) strong and untestable assumptions

As stated above, statistical inference from the sample to the target population is only valid if every respondent from the target population has a positive and known probability of being included in the sample. These requirements are not met for non-probability surveys. However, even for probability surveys, a random sample does not lead to a random sample of survey responents if the response rate is not 100% (Lohr, 2023, p. 5). Systematic nonresponse can lead to nonresponse bias.

For probabilistic surveys, however, a broad strand exists in the literature on under which conditions valid inference is possible, even for surveys that suffer from nonresponse. In the following, we discuss two key assumptions made to draw valid inferences if not every sampled individual responds to the survey. These are: the assumptions of *ignorability* and *positivity* (Mercer et al., 2017; Rosenbaum & Rubin, 1983).

*Ignorability (A1)* means in the survey setting that all mechanisms by which the respondents are selected for the survey are independent of the variables of interest, either unconditionally or conditional upon observed covariates that are associated with participation and the variable of interest, the so-called confounding variables. It is important to note that if only conditional ignorability can be assumed, the results must be weighted, using the confounders as auxiliary variables. Let us, for example, assume that our variable of interest (e.g., technical competence) dependends on Internet use, which also affects survey participation. Non-participation is only ignorable conditional on Internet use. When accounting for differences in Internet use between the survey respondents and the target population with the help of weighting, the ignorability assumption is fulfilled. However, if we did not measure Internet use, or if non-participation depends on other unknown or unmeasured characteristics that are not part of the weighting process, we will encounter non-ignorable bias in target population inference that can not be fully corrected.

For *positivity (A2),* all possible subgroups defined by confounding variables must be represented in the survey. Let us assume that technical competence is not only dependent on Internet use but also on age and that these characteristics also affect survey participation. Let us further assume that both confounders are measured in the survey. In such a case, adjustment weights can reduce bias in the variable of interest, e.g., technical competence, if we know the distribution of age and Internet use in the target population, for example, from a census (that has little or almost no error). If, however, there is, no Internet user of 70 years or older in the survey (but does exist in the target population) no weighting procedure can account for subgroups that are entirely missing from the survey.

For non-probability surveys, it is very likely that these assumptions are not fulfilled (Valliant & Dever, 2011, p. 134), but it is also impossible to examine whether they actually are fulfilled for any non-probability survey because there could always be an unknown uncontrolled confounder that might stem from self-selection or sample selection (Lohr, 2023, p.8).

## 3.2 3.2. Representation errors can not be differentiated

For non-probability surveys, as discussed in Section 2.2, different kinds of representation error (i.e., bias due to coverage error, sampling error, or nonresponse error) can, in general, not be distinguished. In the absence of a sampling frame, let alone a sample, coverage error or nonresponse error are not meaningful concepts. Therefore, many researchers use the term (self-)selection bias when discussing the representation error in non-probability surveys (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013; Kohler et al., 2019). As the selection into the sample in non-probability surveys often depends mainly on self-selection (Callegaro, Lozar Manfreda, & Vehovar, 2015, p.8) or arbitrary selection by the recruiter or interviewer, unequal selection probabilities are common but cannot be mitigated because they are unknown. This is

also the case for respondents sampled randomly from a source that suffers from selection bias (Baker, Blumberg, Brick, Couper, Courtright, Dennis, Dillman, Frankel, Garland, & Committee, 2010, p.36). In many non-probability surveys, one source of selection bias lies in the decision to visit the place or website within the exact time frame when one advertises or conducts the survey. To take a survey sampled on a news website, for example, respondents who visit the news website daily for one hour might have a higher chance of seeing the advertisement and participating in the survey than respondents who visit the website only once a week. Similarly, respondents who do not have access to the Internet or the specific website, do not have a chance to participate in the survey.

Nonresponse is a form of reverse self-selection (Lehdonvirta et al., 2020, p. 139), caused, for example, by declined or unanswered survey invitations. However, in many non-probability surveys, for example using river sampling, invitations are not sent directly to individuals, so knowing who has seen the invitation (e.g., the ad in social media surveys) and did not participate is impossible (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013). Even if it would be possible to determine who has seen an invitation, the characteristics that lead to ignoring a survey advertisement are likely different from those that lead to ignoring a direct survey invitation. In general, the participants can be expected to be a very small and rather selective group of those who could participate, potentially leading to larger and more complex biases than in probability surveys, where the potential number of participants is much smaller and limited to the sample (Valliant & Dever, 2011, p. 134). In addition to some potential respondents not answering the survey, the opposite problem can also exist in many non-probability surveys. As Lohr (2022) states, in a probability survey, target population members are (normally) only able to participate once and only if they are eligible respondents. However, in a non-probability survey, it is often possible to malevolently participate more than once or motivate others to participate and answer in a certain way to influence the results of the survey toward a specific outcome (Bethlehem, 2015), and those malevolent participants may be hard to identify. Regarding online surveys, it is even possible for malevolent actors to program bots to participate in the survey either with the intention of receiving the incentive or to influence survey results (Goodrich et al., 2023). With the recent technologies, those bots might be very hard or even impossible to differentiate from real respondents (Höhne et al., 2024).

In some non-probability surveys (e.g., online access panel or student samples), potential participants are asked directly to participate by mail or another method. Here, calculating a response rate may be possible, but we recommend avoiding the term response rate in the context of non-probability surveys and instead suggest different terms, such as a "participation rate" (The American Association for Public Opinion Research, 2023; Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013). As mentioned above, online access panel members are usually not randomly selected in the first place, and the fact that the access panel is nonprobability-based should be reflected in a terminology that is different than the terminology for probability-based surveys.

## 3.3 3.3. Weighting is likely to fail

In probability surveys, design and adjustment weights are applied to account for unequal selection probabilities and systematic nonresponse. However, for adjustment weights to completely remove nonresponse bias, a number of requirements must be met, for example, that all auxiliary variables related to nonresponse and the variable of interest are included in the weighting process. This assumption is only testable if the true target population value (or at least a strong benchmark approximating the true target population value) for all variables of interest is available (Lohr, 2023, p.8). In practical applications, it is very unlikely that the requirement is met. While weighting might still partially reduce bias, it is also possible that bias will increase (Yeager et al., 2011). Weighting is, however, more likely to be successful in probability surveys because, it only has to correct for bias due to nonresponse and capture characteristics related to the response decision, while the selection bias in nonprobability surveys might be affected by many decisions of potential participants, which each might be correlated with other characteristics. In other words, the ignorability (A1) assumption might be harder to fulfill in non-probability surveys because of the potentially higher number of mechanisms affecting survey participation. Additionally, the positivity assumption (A1) might also be harder to fulfill because efforts are made in probability surveys to target sampled individuals (Lohr, 2023, p.6). Therefore, one can assume that missing population subgroups entirely is less likely in probability than in non-probability surveys.

For non-probability surveys, design weights cannot exist per definition. However, adjustment weights are sometimes proposed to mitigate all kinds of bias simultaneously (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013, p.70; Callegaro, Lozar Manfreda, & Vehovar, 2015, p.183). This means that the auxiliary variables need not only to be correlated to the nonresponse mechanisms but also to an unknown number of other selection mechanisms, making the ignorability assumption (A1) even more complicated to fulfill. At present, there is little reliable knowledge about whether and which weighting methods generally work for non-probability surveys (Cornesse & Blom, 2020, p. 21). In general, there is large evidence that the success of weighting is far less affected by the specific weighting method than by the variables included in the weighting process (Mercer et al., 2018). More research is needed in this area. This is in contrast to probability surveys, where weighting is well-established and based on a large body of literature (Edelman, 2023).

## 3.4 3.4. Quotas are no solution

A common attempt to introduce control into the sample selection of non-probability surveys is to apply quota sampling. In quota sampling, reference data are used to get information on the composition of the target population regarding selected quota variables, and the survey is conducted in a way that the composition of survey respondents regarding the quota variables matches their composition in the target population. The most popular quota variables are demographic variables, but some other variables might also be used, for example, the proportion of persons in the target population with Internet access. Then, the sample selection process aims to recruit a sample that matches those quotas. Quotas can be as simple as one or more univariate distributions but also cross-tabulated distributions (e.g., by age and gender).

In opt-in panels, quota sampling can be done by using known information on the panel members to only invite respondents according to those quotas. If we know, for example, that the target population consists of 55% female and 45% male persons a quota would demand that 55% of respondents are women.

Only when the quota variables are (highly) correlated with the selection mechanisms and the variables of interest, the accuracy of estimates for the variables of interest can be improved as well. Quotas implicitly assume that all characteristics that affect participation are captured in the quota. This assumption is very unrealistic, and it is nearly impossible to implement quotas for all needed variables, as it is impossible to know exactly what characteristics should be included. Only if all selection mechanisms are addressed in a quota (fulfilling the ignorability assumption), and the quota can be fulfilled (i.e., fulfilling the positivity assumption) accurate survey results can be achieved. One should be aware of these limitations when assessing results from surveys that are, for example, advertised as being representative of age, gender, and education.

## 3.5 3.5. A higher number of respondents does not guarantee lower bias

Proponents of non-probability surveys might argue that the possibility of reaching large sample sizes very quickly can make up for (some) problems with non-probability surveys, which is, however, not true (Meng, 2018). This misconception might stem from the fact that for probability surveys, larger sample sizes improve the precision of estimates. They also increase the ability to draw inferences even for subgroups that are rare in the target population.

The accuracy of an estimate depends on two factors: precision and bias. A typical measure of precision in a probability survey is the standard deviation, and a typical measure for the bias is the difference (e.g., in mean) of the survey estimate from the target population value. The Mean Squared Error (MSE) of an estimate, a common measure of accuracy, includes bias and precision (Kohler, 2019, Equation 5 on p. 153). The precision of a survey estimate is related to the sample size, that is, for a given estimate, the precision increases as the sample size increases. When the sample size is large, precision increases and the influence of a small number of outliers will be reduced.

While larger sample sizes increase precision, there is no guarantee that they reduce bias. If, for example, an estimate has a high precision due to a large sample size but also a large bias, that survey will only allow to precisely measure biased parameters(s). This is also true for a non-probability survey that covers especially large proportions of the target population (Meng, 2018). Meng (2018) argues that the required sample size for a non-probability sample to reach a similarly low MSE as a small probability survey mainly depends on the correlation of the variable of interest with the probability of being part of the sample. Even a very small correlation (e.g., 0.05) can be a huge problem, leading to the need for a tremendous-sized set of respondents (i.e., more than half of the target population size) to produce results as accurate as a much smaller simple random survey (i.e., 400 respondents, independent of target population size).

## 3.6 3.6. Estimates of precision need a lot of assumptions

There is no statistical formula that allows to estimate precision for non-probability surveys. Variance formulas and the like that are derived for probability sampling cannot simply be transferred to non-probability surveys. There are, however, some commonly used methods to cope with this problem. Examples include using jackknife or bootstrap procedures to estimate precision (Elliott & Valliant, 2017).

In jackknife variance estimation, the variance is calculated in a repeated procedure. One case is removed from the sample for the first step, and the estimate (e.g., mean, median, etc.) is calculated. Then, the same step is repeated by leaving the second case out of the original sample, and so forth. This procedure is repeated for every case in the survey. Finally, the gathered list of estimates is used to calculate the variance or confidence interval (Quenouille, 1956). Using the bootstrap method (Efron, 1979), repeated samples (around 2,000-10,000) are drawn from the original sample with replacement. Then again, the estimate is calculated for every bootstrap sample, and the measure of uncertainty is calculated from these estimates. Later, a Bayesian bootstrap was presented as an alternative to the standard bootstrap method (Rubin, 1981).

Both resampling methods were above described only in their most simple form, which assumes simple random sampling. In practical applications, every step in the sampling design, including quotas and all weighting steps, must be accounted for in every replication (McPhee et al., 2022). Such methods are commonly used to estimate the precision of estimates from probability surveys. For nonprobability surveys that lack a proper sampling design and thus design weights, and as their adjustment weights may not include all needed auxiliary variables, they are "likely to be too narrow." (McPhee, Barlas, Brigham, Darling, Dutwin, Jackson, Jackson, Little, Lorenz, Marlar, Mercer, Scanlon, Weiss, & Wronski, 2022).[5]

## 3.7 3.7. Conclusion on the generalizability of results

As we discussed in the previous sections, there are many potential sources of bias in non-probability surveys. Also, it is rarely possible to detect these biases or determine their extent. In contrast to probability surveys, no well-established theory, like sampling theory, allows for statistical inference (Callegaro, Lozar Manfreda, & Vehovar, 2015, pp. 52-56; Vehovar et al., 2016, p. 332). As we can never be sure whether the assumptions needed for an estimate to be accurate are fulfilled (A.1, A.2), the estimates of the survey may or may not be extremely biased, without us being able to know. This is even the case when quotas or weights are used. Under these conditions, it is impossible to generalize the results from a non-probability survey without an unreasonable risk of bias.

In contrast to generalizing the survey to the target population (e.g., expecting external validity), it can be less problematic to expect internal validity when using an experimental design within the survey (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013, p. 41; Kohler, 2019;

[5]More information and recommendations on transparent reporting of precision in non-probability surveys can be found in the AAPOR Task Force Report on "Data Quality Metrics for Online Samples" by McPhee, Barlas, Brigham, Darling, Dutwin, Jackson, Jackson, Little, Lorenz, Marlar, Mercer, Scanlon, Weiss, & Wronski (2022).

Mercer, Kreuter, Keeter, & Stuart, 2017, p. 254). Using non-probability surveys for experiments will be discussed in section 4.1.

# 4 When are non-probability surveys fit for purpose?

Non-probability surveys can be valuable in the social science tool kit if used for a proper purpose, for example, when statistical inference is not the aim of the project (Cornesse et al., 2020; Kohler & Post, 2023). Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau (2013) discuss that non-probability surveys may be viable for testing some theoretical concepts if internal validity is given. Even a non-probability survey is internally valid if an experimental design is used for which survey participants are randomly selected into treatment or control groups. In a correctly implemented survey experiment, the effect a treatment has for the survey participants can be estimated without bias. In theory, there might even be situations, where external validity is given. However, this is only the case if the treatment effect is homogeneous for the target group (Kohler, 2019), i.e., the treatment effect does not vary by any characteristics of individuals in the target group. In such a situation the ignorability and the positivity assumptions are fulfilled. Kohler (2019) suggested testing the homogeneity of the treatment effect with the data at hand. This can be done by splitting the set of respondents into various subgroups (e.g., male and female) and testing whether the treatment effect differs between these subgroups. To test this, additional information should be collected when conducting the survey to split the survey participants into several subgroups and the sample size must be large enough to reach a sufficient sample size in the subgroups. Nonetheless, this method can only falsify the homogeneity assumption and is restricted to the population subgroups that are captured in this subgroup analysis. Therefore, the method can only provide initial insights and indicate apparent violations, while the generalizability of the results can not be guaranteed. To confirm findings, experiments should be replicated with a probability survey, if possible.

Another fitness-for-purpose example would be an explorative study that aims to generate hypotheses rather than testing them. In such a case, a non-probability survey could provide first insights. In a subsequent step, a probability survey can be conducted to test the derived hypotheses for the general target population. Similarly, non-probability surveys are valuable for qualitative research, and they are frequently used to pretest survey questionnaires, e.g., the implementation of filtering and randomization of questions (Jerit & Barabas, 2023, p.10).

Additionally, in some cases, the target population of a study may be very hard to reach with probability samples. Hard-to-reach populations are populations that are either (1) *hard to sample,* (2) *hard to identify,* (3) *hard to find and contact,* (4) *hard to persuade, or* (5) *hard to interview* (Tourangeau, 2014, Chapter 1). To address the first three types, and potentially the fourth, Snowball Sampling & Respondent Driven Sampling could be a solution where the sampling process is in parts driven by the respondents (Lee, Suzer-Gurtekin, Wagner, & Valliant, 2017; Lohr, 2022, p. 505). This is because respondents of a special population, for example, members of the LGBTQIA+ community, might know and be able to invite other members of that population. Being invited by some other respondents who one trusts might also have positive effects on participation rates.

Although Snowball Sampling can reduce the cost of surveying hard-to-reach respondents it still leads to a non-probability sample, with biases of unknown magnitude. For example, the set of respondents might only include respondents who bond with others of the same population, leading to the exclusion or underrepresentation of more isolated individuals (Sadler et al., 2010). Therefore, generalizing the results is still highly problematic. In the case of hard-to-reach populations, it is likely even harder to evaluate potential bias than for surveys of the general population due to the limited amount of available benchmark information. However, for some hard-to-reach populations, non-probability surveys may be the only practical or affordable option to gather insights (Jerit & Barabas, 2023, p. 11).

Kohler & Post (2023, p. 84) add another aspect to the debate regarding the usage of non-probability surveys. The authors agree with the previous points, for example, that non-probability surveys may be viable for explorative studies or special populations in social sciences. However, the authors warn that even in the case of fitness-for-purpose it is not appropriate to naively communicate findings from non-probability surveys in media to a larger audience outside of the scientific fields not trained to evaluate the validity of such findings. The same is true when the results are expected to be used to inform political decision-making processes. The authors argue that accurate results are especially important in those processes because careless generalization of results from non-probability surveys could harm general trust in scientific results or lead to misguided political decision-making.

# 5 Recommendations for fit-for-purpose settings

When considering using a non-probability survey it is crucial to make sure that it is fit for the desired purpose. If the survey is to be used to draw inferences to a general population, non-probability surveys are not an appropriate choice, as they are likely biased to an unknown and potentially substantial degree. This is especially true for topics that are of high interest to the broad population, that are likely to be discussed in the media, or that aim to influence political decisions (Kohler & Post, 2023). In these situations, it is especially important that the results are as accurate as possible because they can impact public opinion and public trust in scientific findings. However, if, for example, one aims to test questionnaires, develop hypotheses, include experiments for which internal validity is most relevant, or survey a special sub-population that is hard to reach, non-probability surveys may be an adequate option.

In case non-probability surveys are considered to be fit-for-purpose, there are several aspects to keep in mind when communicating the findings. Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau (2013) recommend being as transparent as possible about the survey process and carefully documenting every step toward the final set of respondents. This allows the expert to make an informed judgment about the quality of the data and the reliability of the findings. Providing extensive documentation can be complicated, as some of the organizations that collect the survey data do not report transparently how respondents were recruited, and some even outsource the data collection, making transparency even harder or impossible to achieve (Jerit & Barabas, 2023, p.14).

In addition, Callegaro, Lozar Manfreda, & Vehovar (2015, p.55) suggest that if using non-probability survey data for statistical inference (i.e., in fit-for-purpose situations such as hard-to-reach populations), clearly acknowledge that the inference may be poor or invalid. Another important recommendation from the literature is to use diverting terminology when talking about non-probability surveys instead of probability surveys (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013; Callegaro, Lozar Manfreda, & Vehovar, 2015; Vehovar, Toepoel, & Steinmetz, 2016). Examples are the use of "indication" or "approximate" instead of "estimate", or "participation rate" instead of "response rate."

The documentation of the data should include an assessment of how the set of respondents compares to the target population of interest (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, & Tourangeau, 2013; Cornesse, Blom, Dutwin, Krosnick, De Leeuw, Legleye, Pasek, Pennay, Phillips, & Sakshaug, 2020; Rohr et al., 2024). Importantly, the assessment should not be based on variables that were used for quotation as they are by definition distributed to exactly match the target population. Similarly, weighting variables should be excluded when comparing distributions after weighting. Additionally, checks for treatment homogeneity are recommended for experimental research (Kohler, 2019). In this vein, the scientific journal *Survey Research Methods* demands specific steps to assess the external validity of results before submitting a paper based on non-probability surveys (Survey Research Methods, 2023).

# Bibliography


The American Association for Public Opinion Research. (2023). Standard Definitions—Final Dispositions of Case Codes and Outcome Rates for Surveys. *10th Edition*. https://aapor.org/standards-and-ethics/standard-definitions/

Bacher, J., Lemcke, J., Quatember, A., & Schmich, P. (2019). Probability and Nonprobability Sampling: Representative Surveys of hard-to-reach and hard-to-ask populations. *Current Surveys between the Poles of Theory and Practice*, *1*(1), 1–18. https://doi.org/10.13094/SMIF-2019-00018

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., & Committee, w. m. i. P. f. t. A. E. C. b. a. T. F. o. u. t. a. o. t. A. S. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Report of the AAPOR Task Force on Non-Probability Sampling 128. *AAPOR Task Force on Non-Probability Sampling*, 1–128.

Bauer, J. J. (2016). Biases in Random Route Surveys. *Journal of Survey Statistics and Methodology*, *4*(2), 263–287. https://doi.org/10.1093/jssam/smw012

Bethlehem, J. (2015). Essay: Sunday shopping— The case of three surveys. *Survey Research Methods*, *9*(3), 221–230. https://doi.org/10.18148/SRM/2015.V9I3.6202

Callegaro, M., Baker, R., Bethlehem, J., Goritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look at the future. In *Online Panel Research: A Data Quality Perspective* (pp. 1–22). https://doi.org/10.1002/9781118763520.ch1

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. SAGE. https://doi.org/10.4135/9781529799651

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons.

Cornesse, C., & Blom, A. G. (2020). Response Quality in Nonprobability and Probability-based Online Panels. *Sociological Methods & Research*, *49*(1), 1–20. https://doi.org/10.1177/0049124120914940

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., & Sakshaug, J. W. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Edelman, M. (2023). Not Really a New Paradigm for Polling. *Harvard Data Science Review*, *5*(3), 1–10. https://doi.org/10.1162/99608f92.33fb61ba

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249–264. https://doi.org/10.1214/16-STS598

Felderer, B. (2024). Nonresponse bias analysis. *GESIS – Leibniz-Institut Für Sozialwissenschaften (GESIS Survey Guidelines)*.

Gabler, S., Kolb, J.-P., Sand, M., & Zins, S. (2016). Weighting. *GESIS Survey Guidelines*. https://doi.org/https://doi.org/10.15465/gesis-sg_en_007

Gile, K. J., Johnston, L. G., & Salganik, M. J. (2015). Diagnostics for Respondent-Driven Sampling. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *178*(1), 241–269. https://doi.org/10.1111/rssa.12059

Goodman, L. A. (2011). Comment: On Respondent-Driven Sampling and Snowball Sampling in Hard-to-Reach Populations and Snowball Sampling Not in Hard-to-Reach Populations. *Sociological Methodology*, *41*(1), 347–353. https://doi.org/10.1111/j.1467-9531.2011.01242.x

Goodrich, B., Fenton, M., Penn, J., Bovay, J., & Mountain, T. (2023). Battling bots: Experiences and strategies to mitigate fraudulent responses in online surveys. *Applied Economic Perspectives and Policy*, *45*(2), 762–784. https://doi.org/10.1002/aepp.13353

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, *70*(5), 646–675. https://doi.org/10.1093/poq/nfl033

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

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

Höhne, J. K., Claassen, J., Shahania, S., & Broneske, D. (2024). Bots in web survey interviews: A showcase. *International Journal of Market Research*, *-270849935*, 1–10. https://doi.org/10.1177/14707853241297009

Jerit, J., & Barabas, J. (2023). Are Nonprobability Surveys Fit for Purpose?. *Public Opinion Quarterly*. https://doi.org/10.1093/poq/nfad037

Koch, A., & Blohm, M. (2015). Nonresponse Bias. *Mannheim, GESIS – Leibniz-Institut Für Sozialwissenschaften (GESIS Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_004

Kohler, U. (2019). Possible Uses of Nonprobability Sampling for the Social Sciences. *Survey Methods: Insights from the Field*, *1*(1), 1–12. https://doi.org/10.13094/SMIF-2019-00014

Kohler, U., & Post, J. C. (2023). Welcher Zweck heiligt die Mittel? Bemerkungen zur Repräsentativitätsdebatte in der Meinungsforschung. *Zeitschrift Für Soziologie*, *52*(1), 67–88. https://doi.org/10.1515/zfsoz-2023-2001

Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, *6*(1), 149–172. https://doi.org/10.1146/annurev-statistics-030718-104951

Lee, S., Suzer-Gurtekin, T., Wagner, J., & Valliant, R. (2017). Total Survey Error and Respondent Driven Sampling: Focus on Nonresponse and Measurement Errors in the Recruitment Process and the Network Size Reports and Implications for Inferences. *Journal of Official Statistics*, *33*(2), 335–366. https://doi.org/10.1515/jos-2017-0017

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2020). Social Media, Web, and Panel Surveys: Using Non-Probability Samples in Social and Policy Research. *Policy & Internet*, *13*(1), 134–155. https://doi.org/10.1002/poi3.238

Lohr, S. L. (2022). *Sampling: Design and Analysis*. Chapman, Hall/CRC.

Lohr, S. L. (2023). Assuming a Nonresponse Model Does Not Make It True. *Harvard Data Science Review*, *5*(3), 1–10. https://doi.org/10.1162/99608f92.2b901b7f

McPhee, C., Barlas, F., Brigham, N., Darling, J., Dutwin, D., Jackson, C., Jackson, M., Little, R., Lorenz, E., Marlar, J., Mercer, A., Scanlon, P. J., Weiss, S., & Wronski, L. (2022). *Data Quality Metrics for Online Samples: Considerations for Study Design and Analysis*.

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. https://doi.org/10.1214/18-AOAS1161SF

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*(S1), 250–271. https://doi.org/10.1093/poq/nfw060

Mercer, A., Lau, A., & Kennedy, C. (2018). For Weighting Online Opt-In Samples, What Matters Most?. *Pew Research Center Methods*. https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/

Pötzschke, S., Weiß, B., Daikeler, J., Silber, H., & Beuthner, C. (2023). A guideline on how to recruit respondents for online surveys using Facebook and Instagram: Using hard-to-reach health workers as an example. *Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_045

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, *43*(3), 353–360. https://doi.org/10.2307/2332914

Rohr, B., Silber, H., & Felderer, B. (2024). Comparing the Accuracy of Univariate, Bivariate, and Multivariate Estimates across Probability and Nonprobability Surveys with Population Benchmarks. *Sociological Methodology*. https://doi.org/10.1177/00811750241280963

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, *9*(1), 130–134. https://doi.org/10.1214/aos/1176345338

Sadler, G. R., Lee, H.-C., Lim, R. S.-H., & Fullerton, J. (2010). Research Article: Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. *Nursing & Health Sciences*, *12*(3), 369–374. https://doi.org/10.1111/j.1442-2018.2010.00541.x

Sand, T., Matthias und Kunz. (2020). Gewichtung in der Praxis. *Mannheim, GESIS – Leibniz-Institut Für Sozialwissenschaften (GESIS Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_030

Stadtmüller, S., Silber, H., Daikeler, J., Martin, S., Sand, M., Schmich, P., Schröder, J., Struminskaya, B., Weyandt, K. W., & Zabal, A. (2019). Adaptation of the AAPOR Final Disposition Codes for the German Survey Context. *Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS - Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_en_026

Statistics Canada. (2021). Statistics: Power from Data!. *12, 1-116*. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/nonprob/5214898-eng.htm

Survey Research Methods. (2023). *Journal of the European Survey Research Association*. https://ojs.ub.uni-konstanz.de/srm/about

Tourangeau, R. (2014). Hard to Survey Populations. In *Cambridge University Press* (pp. 3–21).

Valliant, R., & Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research*, *40*(1), 105–137. https://doi.org/10.1177/0049124110392533

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). Non-probability sampling. In *The Sage Handbook of Survey Methodology* (pp. 329–346).

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. https://doi.org/10.1093/poq/nfr020

Publisher