FORS GUIDES
to survey methods
and data management

Representativeness of Surveys and its Analysis

Michael Ochsner[1]

[1] FORS

FORS Guide No. 15, Version 1.0

December 2021

Abstract:

The analysis of representativeness of a data set belongs to the standard quality assurance procedures in survey research. This FORS Guide challenges current practices of the analysis of representativity and suggests a framework to analyse the risk for representation bias taking into account different uses of data.

Keywords: Bias, Inference, Population, Data Quality, Representation

How to cite: Ochsner, M. (2021). *Representativeness of Surveys and its Analysis*. FORS Guide No. 15, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences (FORS). doi: 10.24449/FG-2021-00015

The FORS Guides to survey methods and data management

The FORS Guides offer support to researchers and students in the social sciences, who intend to collect data, as well as to teachers at university level, who want to teach their students the basics of survey methods and data management. Written by experts from inside and outside of FORS, the FORS Guides are descriptive papers that summarise practical knowledge concerning survey methods and data management. The FORS Guides go beyond the documentation of specific surveys or data management tools and address general topics of survey methodology. They give a general overview without claiming to be exhaustive. Considering the Swiss context, the FORS Guides can be especially helpful for researchers working in Switzerland or with Swiss data.

Acknowledgement:

Michael Ochsner wants to thank Jessica M. E. Herzing for her editing work on this FORS Guide. As the editor of FORS Guides at the time, she initiated the theme, discussed the plan and commented on first drafts. The Guide was delayed by Covid, leading to the fact that Jessica left the Editorial Board before the Guide was finished. Michael also thanks Oliver Lipps for his useful comments and Monika Vettovaglia for lectureship.

1. INTRODUCTION

The aim of a survey is to find out some unknown characteristics of the population or how the population thinks about specific topics. For practical reasons, information is usually asked only from a small selection of the population under scrutiny. From the information gathered from such a sample one intends to make statements about the whole population, usually using inferential statistics. However, if the sample is substantially different from the population regarding what is to be measured, statements about the population are biased. Therefore, survey researchers are concerned about the appropriateness of the samples of respondents[1] with which they aim to make statements about the population: How much confidence can they have in the data? In such contexts, often the word "representative sample" is used. The idea is that representation bias "results from systematic differences between the group to which generalizations are to be made and the group from which data were obtained" (Wilcox et al., 1994).

Survey researchers have developed several techniques to assess the "representativeness" of a data set, such as response rates, deviations from population statistics or the more statistically sophisticated R-indicators. However, these techniques come with several problems. First, they are data-driven and therefore dependent on data availability (i.e., variables that are available for both the respondents and the population). Second, the data used to investigate representation bias are often not linked to the research questions at hand. Third, representation is presented as a property of the data, while it is a property of the estimate of interest, and the same data set can be biased for some research questions but not for others.

In this guide, I challenge the current concept of "representativeness" in survey research and develop a framework for the analysis of the "risk for representation bias" that draws on the Total Survey Error (TSE) framework (Biemer, 2010). However, I use the concept in the opposite way than is usually done: Whereas most methodological research uses the TSE framework to explain what part in the TSE framework the study is investigating in analysing errors, in this study, the total error is the focus of the analysis of representation bias. Studying isolated sources of error without integrating them into a context comes with the implicit assumption that the total survey error is the sum of all partial errors, which obviously is very unlikely to be true. Different sources of errors can accumulate but they can also compensate each other regarding bias of a specific estimate we are interested in. I suggest a framework to investigate the risk for representation bias using several indicators in relation to an actual research question (or estimate) at hand.

I start with a short introduction to the use and misuse of the term "representative" in survey research and the current practice in the assessment of representation bias. I then describe a framework for the analysis of the risk for representation bias that contextualizes the concept of representativeness by linking it to the analysis that seeks to make statements about the population. I further give an overview of the Swiss context and present some examples how the framework can be used in different situations faced by survey researchers. The guide ends with recommendations how to deal with questions of representativeness and useful links and sources for survey practitioners.

[1] Note that I write from the point of view of a *data user*. Along with most survey research literature, I use the term "sample" for the respondents available in the data set for which representativeness is under scrutiny. If I refer to the sample of persons invited to take part in a survey, I use the term "sampling frame".

2. WHAT DOES “REPRESENTATIVE” MEAN AND HOW IS IT USED?

The term “representative” is often used in scientific as well as non-scientific texts referring to surveys. However, what does “representative” mean? While many have an intuitive idea of what “representative” means, it gets more difficult when asked to define it clearly. Generally, people will answer something along the lines that a) 'representative’ means that a sample or respondent pool resembles the population in question (see, e.g., Schouten, Cobben & Bethlehem, 2009, pp.101-102) or b) a sample is representative if estimates drawn from the sample can be generalized to the population (see, e.g., Wilcox et al., 1994). Both statements, however, are problematic. The first will never be achievable because a sample would need to be of enormous size to actually resemble the population in all relevant aspects, the second is too abstract to be practically useful, as only known if the population parameter is known; additionally, it includes the assumption that an estimator cannot be biased. Obviously, the wide use of the term is in contrast with the vague definition of its meaning. Even though contested by statisticians for decades for being ambiguous and having no statistical foundation (see, e.g., Kruskal & Mosteller, 1979a, b, c), the term remains prominent among survey researchers and the analysis of “representativeness” of the sample or data is part of the standard procedure when conducting a survey or reporting results. Why is this the case? On the one hand, the intention is a valid one: for example, investigating which of several survey designs leads to a “better” sample is clearly useful (see Good & Hardin, 2009). On the other hand, claiming representativeness is also an act of communication putting a “quality stamp” on the data. In fact, uses of the concept of “representativeness” are diverse. In their encompassing analysis of the use of the term, Kruskal and Mosteller (1979a, b, c) identified nine types of meanings for “representative sampling” and its synonyms. In the following, I will present Kruskal and Mosteller’s typology to illustrate the diverse meanings and (mis)uses of the term – and question, along with them, the usefulness of the term. I will then turn to the issue at hand and discuss the occasions where survey researchers want to make statements about “representativeness” of their data. Finally, I will give a short overview of the state of the art in practice and techniques to evaluate the “representativeness” of a given data set.

2.1 A TYPOLOGY OF THE MEANINGS AND (MIS)USES OF THE TERM “REPRESENTATIVE”

Kruskal and Mosteller (1979a, b, c) reviewed a broad range of scientific and non-scientific literature and categorized how the term “representative” is used or interpreted. The first five types were detected in non-scientific literature (news articles, popularization of research, government documents etc.). They identified the same five types in the scientific literature (excluding statistics) plus a sixth type. In the statistical literature they also identified the six types plus another three additional types. Interestingly, going from non-scientific to scientific to statistical texts, the authors did not find less confusion about the term but rather more. As the types seven to nine are subtypes of the sixth type, the following overview excludes them for readability. The interested reader is referred to Kruskal and Mosteller (1979c).

Kruskal and Mosteller (1979b, p. 111) identified the following types of use of the term “representative”:

1. General acclaim for data
2. Absence of selective forces
3. Miniature of the population

4. Typical or ideal case(s)
5. Coverage of the population
6. Vague term, to be made precise

The first usage, *general acclaim for data*, can be summarized as: “Take my word, without evidence, that my sample will not lead you astray” (Kruskal & Mosteller, 1979c, p. 246). It is often used in newspaper articles on voting forecasts but can also be found regularly in scientific literature (as a positive – e.g. “our sample is representative for the population” – or negative qualification – e.g. “our sample is not representative for the population but serves as preliminary evidence”). It is thus simply a claim that is not further proven.[2]

The second usage, *absence of selective forces*, is widely used in survey research and refers to the more precise term *response bias*. It reflects that taking part in a survey might be linked to certain characteristics of a potential respondent. Presence or absence of such selective forces could then be interpreted as signs of (non)representativeness. The problem with such an interpretation, however, is that absence of selective forces can never be proven because presence of selective forces might be found for any variable and it may, thus, distort the analytical potential of the sample. Similarly, presence of selective forces might be irrelevant for some usages of the data, for example in cases when the selection variable is unrelated to the variable of interest. In survey practice, R-indicators are a typical example of this use of the term combined with “general acclaim for data”.

The third usage, *miniature of the population*, is a common way of interpreting representativeness. The idea is that the sample resembles the population perfectly. A sample is tested against known parameters of the population and qualified as representative or not. Yet, we normally do not know the population characteristics, which is why we want to draw the miniature replica of the population in the first place. Even if we knew some characteristics of the population, the idea of it is not clear. Potentially, for general population surveys, there will be a very large number of variables to build the miniature. When thinking of cross-tabulations, it will lead to an infinite regress as the example of a doll house shows: to be a perfect miniature, the doll house needs to contain a doll house. This miniature doll house, if it is a perfect miniature, contains again a doll house that is a miniature of a doll house etc. As Gilbert et al. (1977, p. 218-220) point out: reaching a perfect miniature of a population for a simple example using only as few as eight variables consisting of two sexes, four marital statuses, five categories of family size, four regions of the country, six types of cities, five occupation classes, four educational levels and three income levels, would lead to 57’600 cells to fill in and one case per cell would likely not even be seen as adequate. Even if we assume that the miniature needs not to be perfect, an important question remains: how close is close enough? Quota sampling reflects this idea of representation.

The fourth usage, a *typical or ideal case*, refers to the idea of finding an average representative of a group, for example a “mean voter state”. The problems with this interpretation are obvious: a state might be a “mean voter state” in one election but not in the following, the case might be typical for the population for some aspects but not for others. Typicalness is an ambiguous concept; it might mean a sample whose members are all typical or that the sample is typical. The first comes down to the sample size of one and the second comes close to the idea of the miniature of the population. This type is less common in the survey literature but used more

[2] To avoid blaming, no citations of cases in the literature are provided. The reader is encouraged to do a quick literature search using the term “representative sample” and the results will show current examples for most types of use. The examples leading to the typology can be found in Kruskal & Mosteller (1979a, b, c).

often in qualitative studies. However, one common use in survey research is the selection of higher-order units, such as one or two countries "representing" different world regions in multilevel analyses, for example Eastern, Western and Southern Europe.

The fifth usage, *coverage of the population*, is similar to the miniature of the population, however, the population is perceived as consisting of several strata that need to be covered. The idea of the coverage of the population rests on the belief that a sample of respondents is representative if all strata are covered, i.e. at least one member of each stratum is present, though not proportionately. The goal is to show the heterogeneity of the population. Similar problems as to the miniature of population apply: How many variables are needed to describe the heterogeneity? Also, without knowing the proportions in the population or in the sample, the disproportionate coverage of members of the different strata leaves doubts on the ability to take into account the disproportion whenever making statements about the population. Examples for this usage of the term "representative" are selection of states having a collectivist or an individualist culture when studying family values or selecting rural, urban and agglomeration communities when investigating attitudes towards abortion.

The sixth usage, the *vague term later to made precise*, is commonly used in the scientific literature. In research involving statistics, other vague terms are fruitfully in use, such as "average" or "variability". They are not clear but can simplify a description or interpretation as they refer to standard language. However, a clarifying statement has to be provided to explicate what exactly is meant by "average" or "variability". In these examples, "average" can stand for an arithmetic or a geometric mean, a median or any form of weighted means, just as "variability" can stand for a standard deviation, an interquartile range or simply the range of a variable. Similarly, it could be argued that the term "representative sampling" is just a common language term for "probability sampling". Thus, in this type of use, the term is used to describe in a simple way what later is correctly specified. Only probability sampling methods fall into this type of the use, such as random samples, stratified samples, cluster samples etc. If non-probability sampling, such as quota sampling, are concerned, the use of the term "representative sampling" refers to one of the first five types, for quota sampling, it is *miniature of the population*.

Kruskal and Mosteller conclude that they would not use the term "representative" for any of the first five usages. Rather the ideas behind such situations need to be clearly specified to be precise (Kruskal & Mosteller, 1979b, p. 123). While the authors suggest caution in the use of the term "representative" (Kruskal & Mosteller, 1979a, p. 14; 1979c, p. 261)[3], they acknowledge the advantage of using the less technical notion of the term "representative" in the sixth usage (*vague term later to made precise*) to describe the result and purpose of a sampling method. Yet, they recommend to use this term only in relation to probability sampling and state that it comes with the obligation to provide a clear description of the procedure (Kruskal & Mosteller, 1979b, p. 126).

What becomes clear from this analysis of the usage of the term is that the notion of "representativeness" is context-dependent. There is no ultimate judgement whether a sample or a data set can be considered representative. Yet, as Kruskal and Mosteller acknowledge despite their criticism, there are some situations when researchers want to make statements about what can be called the "representativeness" of an achieved sample. However, it is

[3] Their reservations about the term "representative" becomes especially evident in their ironic, sometimes humoristic comments of its usage in the texts they cite.

advised that researchers are explicit and precise regarding how and which "representativeness" is assessed.

Because the use of the term "representative" is not adequate in situations where no random/probability sampling is involved, I exclude from the following analysis any other data collection method, such as big data, quota samples or access panels, that *by definition* cannot be considered "representative" of a population. They have other merits and other aims than being representative.

2.2 THE ISSUE AT HAND: WHEN DO WE WANT TO MAKE STATEMENTS ABOUT REPRESENTATIVENESS?

Occasions where researchers want to assess or compare the "representativeness" occur in different situations in survey research. First, survey administrators need to document how data were collected and what the outcome regarding participation is (see, e.g., AAPOR, 2016). One part of what belongs to a survey documentation is the participation rate and some response analysis, which are often interpreted as analysis of representativeness. A special case of this application is fieldwork monitoring. Some survey practitioners want to take measures already during fieldwork in case that biases are detected. For example, if respondents from the countryside participate more frequently, the survey agency might increase their efforts in urban surroundings.

Second, some surveys follow a sequential design, i.e., not the whole questionnaire is fielded at once, but several modules are fielded sequentially, or the questionnaire is repeated after some time (panel studies). This comes with attrition, i.e., some participants answering the first part will not answer the second. In cases of sequential design, a survey researcher wants to know in what way the two or more samples differ regarding the participants. Are they comparable? Is there a bias in the second, usually smaller sample?

Third, survey designers strive for improving surveys. To do so, they often field experiments with different survey designs, for example testing the effect of different orders of questions, different incentives for participation or different modes of data collection. To derive a conclusion, many aspects are relevant, such as costs, practicability, ethical considerations etc. But the idea of "representativeness" plays a key role: which design provides the users with the "best" data, i.e. the data that enables researchers to do accurate estimations and inferences?

Fourth, not only survey designers and administrators are concerned with such questions. Researchers looking for adequate data to investigate their research questions also want to evaluate different available data sets. They want to have a "representative" sample at hand. Following from the fact that the concept of "representativeness" is context-dependent, in fact, each researcher should be concerned by these questions and at least investigate potential biases affecting their estimates and inferences and consider measures to counterbalance them.

Hence, there are a few situations where the intention to investigate which samples resulting from different survey designs are "better representing" the population is a valid one (see Good & Hardin, 2009). However, an important caveat applies as shown in the previous section: the analysis needs to be context-dependent whereas in the first three cases survey designers usually want to make general statements about the data. There is a risk that the term "representativeness" is used in the meanings *general acclaim for data*, *absence of selective forces*, or *miniature of the population*, all of which have been proven to be problematic. How can the virtuous desire to assess the quality of the data with regard to estimation of unknown population parameters be best achieved? Before sketching out a framework to analyse the potential for representation bias, some current practices are described and critically examined regarding Kruskal and Mosteller's (1979a; b; c) typology.

2.3 CURRENT PRACTICE OF, AND TECHNIQUES FOR, REPRESENTATION ANALYSIS

Current practices of representation analysis rely largely on two types of data available to conduct analyses: data from the sampling frame and external data, such as official population statistics. The first data are linked with the survey collected for the same sample members (available for respondents and non-respondents), whereas the second data are based on census or other surveys not linked with the survey under scrutiny and, thus, stem from a different, usually more comprehensive sample.

Indicators based on the sampling frame

The best-known indicator using the sampling frame data is the *Response Rate*. It has been used as a proxy for response or representation bias or even data quality and to guide decisions about fieldwork procedures (Biemer & Lyberg, 2003; Wagner, 2012). For example, the European Social Survey (ESS) lists the response rate as the first among "the most important standards on data collection" on their homepage[4] and requests its members to target a response rate of 70% or to strive for achieving a higher response rate than in the last round (ESS, 2017). However, meta-analyses of methodological studies investigating nonresponse bias show that response rates are weak predictors of nonresponse bias (Groves, 2006; Peytcheva & Groves, 2009). Knowing about this problem, other indicators based on response rate have been suggested and are frequently applied: *Subgroup Response Rates* indicate differences in participation across categories of variables available in the frame, such as ethnicity, age or nationality. Often *t-tests* are used to indicate "statistically significant" differences between the given categories of a variable, for example education, in the achieved sample and the sampling frame (e.g., Lipps & Pekari, 2016). The *Coefficient of Variation* indicates the average deviation of subgroup response rates as a ratio of the overall response rate. To include information from several variables at the same time, the coefficient of variance can be calculated on the basis of poststratification weights using several variables in the frame (or from paradata). Finally, the so-called *R-indicator* or representativity indicator (Schouten, Cobben & Bethlehem, 2009) calculates the expected response rate, given other variables from the sampling frame or paradata for each category of those variables (so-called "predicted probabilities"). The standard deviations across categories and variables are combined into a single number between 0 and 1. It is higher the smaller the standard deviations are across categories of the predictor variables from the sampling frame or paradata.

Wagner (2012, p. 563) concludes about this type of indicators that the "more predictive these data are of the response indicator […], the better able we are to assess the risk of nonresponse bias". However, the assumption here is that the auxiliary variables from the frame (or the paradata) are related to the statistic of interest. If there is no relation, we do not learn anything about the bias of the statistic of interest. Yet, such relations are normally very weak (Peytcheva & Groves, 2009). The use of response rates and its derivates as indicators for "representativeness" or even data quality is highly problematic as they are a property of the

[4] https://www.europeansocialsurvey.org/methodology/ess_methodology/data_collection.html

survey while the nonresponse bias is a property of the statistic of interest (see also Groves, 2006; Wagner, 2012).

The use of statistical significance tests (t-tests) comes with several problems: First, what does a significant difference in one of the three categories of a variable mean? Is it possible to have a bias regarding persons with a high education but no bias regarding persons with a middle or low education? To solve this theoretical problem, chi-square tests can be used to indicate a significant difference of the distribution of the entire variable (or joint tests for all categories of a variable in a logistic regression). Yet, a conceptual issue remains: a significance test indicates the likelihood that the two samples have been drawn from the same population or that an estimate from a random sample is likely to apply in the population. However, we are not interested in whether the participation we observe in our random sample (the frame) is true for the population (of all possible surveys) but simply whether the actual sample is biased, which is visible from the effect sizes and not the statistical significance.

Remembering Kruskal and Mostellers' typology, we see that indicators based on the sampling frame reflect the use of type 2, *absence of selective forces*. In the case of the R-indicator, a single number is used to judge the "representativity" of a whole data set, thus combining types one and two of Kruskal and Mosteller's typology, using a limited set of variables, generally not of interest to the data users for a *general acclaim of data*, which is scientifically inappropriate.

Indicators based on population data

For many surveys, there is no sample frame available, or it contains only a very limited number of variables; similarly, paradata might be limited and not strongly related to the statistics of interest. Therefore, survey practitioners usually also evaluate how different or similar the distributions of their achieved sample are compared to the population. To be able to evaluate this, the distributions need to be known for the population. Often, the variables are taken from a census or from official statistics.

Indicators used in this group are based on summary distributional statistics. The basic indicator is the *Absolute Bias*, i.e., the absolute value of the difference in the percentage between the population data and the achieved sample for each category of a variable, for example marital status. If a variable has several categories, the *Average Absolute Bias* gives an indication of the bias for the whole variable (e.g., Lipps et al., 2015). A more sophisticated measure is the *Absolute Relative Bias* that divides the Absolute Bias by the value of the population to account for the fact that a difference of 5 percentage points in a category that accounts for 60% in the population is a smaller bias than a difference of 5 percentage points in a category that accounts for 10%. The *Absolute Relative Bias* is thus the expression of the difference of the sample in percentages of the value in the population (e.g., Eckmann, 2016). Again, the *Average Absolute Relative Bias* gives an indication of the bias for the whole variable.

Often, t-tests or chi-square tests are used to indicate "significant" differences between the given categories of a variable, for example education, in the population and the achieved sample when sample-based census data is used (e.g., Lipps & Pekari, 2016). This comes with the same issues noted above that it does not make sense per category but only per variable. Yet, significance testing does make sense as we test whether two samples were drawn from the same population. Significance testing can even be a direct test for how precise the estimates drawn from the sample are, for example by examining whether the population data falls within the 95% confidence interval of the given estimate from the sample, an approach that is astonishingly rarely applied. I do not know of such an instance.

The basic idea behind this approach is the "miniature of the population" type in Kruskal and Mosteller's typology: it is the idea that the sample should reflect the population in its characteristics. However, there are so many characteristics of a population that following this approach leads to an infinite regress (Gilbert et al., 1977). Additionally, similar to the sample frame data, bias regarding variables from population data is only indicative of representation bias for a statistic of interest when they are strongly correlated to the statistic of interest, which they are usually not. If the result from the comparison to population data leads to a general conclusion for a data set to be "representative", the combination of *miniature of population* and *general acclaim for data* leads to an unscientific, inappropriate use of the term.

The state of current practices

While the list of indicators presented above is only a selection of indicators in use in practice, the general problems remain the same for any derivative indicator as either frame/para data or population data are used to compare mostly socio-demographic variables across samples. The goal of the critical examination of these indicators is not to question the intention of the exercise. Rather, it is to remind practitioners of the quest they are taking. While all these indicators might give useful information in a specific context, they all fail to give relevant information to describe the "representativeness" of the sample regarding the use of the data. This observation is not new (e.g., Koch & Blohm, 2016; Wagner, 2012), and there are only rare exceptions taking the survey topic into account, such as Lipps and Pekari (2016) who include political variables in their representation analysis for a political survey. However, bias depends also on the type of analysis as it is a property of the estimate, which is never taken into account. It is remarkable that the analysis of data, so central to this question, is missing from any approach to representation.

What, then, can survey practitioners or researchers do when they want to know whether estimates from the data can be trusted or how to decide which of several survey data sets are to be preferred? The answer is as simple as unsatisfactory: avoid the term "representative" as it is not a scientific concept. When it needs to be used, revert to type six, and use it as a "vague term to be made precise". However, "we used probability sampling and did not find response bias", does not do the trick as it basically still is going back to type two with the goal of type one. Rather, use the term only in context, linking it to the statistic of interest and avoiding a general acclaim for data. While this answer might seem unsatisfactory at first sight, it is scientific at the second: why would we want to make claims without a proper frame of reference? In the following chapter, I suggest a framework for the analysis of the risk for representation bias that helps avoiding the pitfalls described above.

3. A NEW FRAMEWORK FOR THE ANALYSIS OF THE RISK FOR REPRESENTATION BIAS

As noted earlier, there are several situations where survey practitioners and researchers want to make statements or an evaluation of how an achieved sample can be used to make inferences to a given population. As has been shown, an analysis of such a question requires contextualization: *for what* can the sample be used to make inferences to a given population? Therefore, I suggest using a framework for the analysis of the risk for representation bias that starts from the purpose of the data use, is based on several indicators linked to this purpose and rests vague regarding the diagnosis, as bias can never be excluded. Importantly, the approach remains clear and transparent regarding the procedure.

The framework is based on two dimensions: the type of analysis (the statistical method to be applied) and the type of variables. The type of analysis indicates the sensitivity to deviations from the population regarding the included variables. This first dimension distinguishes four different uses of data. First, a researcher might be interested in subpopulation analysis: one uses general population data but is only interested in one category of a variable, for example parents. Second, data can be used to compare univariate distributions or descriptive statistics: one might be interested in the percentages of agreement to the implementation of environmental taxes. While before the researcher was interested in one category, here all categories of one variable are of interest. Third, a researcher might want to calculate means or correlations between two variables. Here, deviations in one category are not relevant as long as they are balanced out. Fourth, researchers often use multivariate statistical models where researchers can control for the bias if a variable is added that explains the deviation.

The second dimension, type of variables, is strongly linked with the benchmark data to be used, respectively the type of bias one is addressing. Three types of variables are relevant for the analysis of risk for representation bias: First, we can analyse risk for bias regarding variables indicating the *likelihood of participation*. These variables thus indicate response bias. They can be analysed using data from the sampling frame or paradata that are available for both respondents and non-respondents, such as living in urban or rural areas, availability of a telephone, type of housing, or nationality. Second, *control variables* are variables that are often used to control for differences across population groups in statistical models, most often socio-demographic variables like education, work status, sex. Such data are often available from official statistics, i.e. population data. Third, *variables of interest* are those variables that are unknown for the population, they are the reason why the survey is conducted and in which a researcher is interested, such as attitudes or behaviour.

Crossing the two dimensions reveals a matrix for the analysis of risk for bias based on the purpose of the study. Most of the time, several fields are relevant: one might be interested only in subpopulation analysis but will apply a multivariate model on many variables of interest including a few control variables. In such cases, it makes sense to analyse the risk for bias of several fields of this matrix. Each field of a matrix can be examined by different indicators matching the purpose. Table 1 depicts the schematic matrix and provides a few ideas of indicators useful to examine each field.

3.1 VARIABLES FOR LIKELIHOOD OF PARTICIPATION

If we want to examine whether participation depends on specific variables, or in other words, to investigate response bias, we will use data from the sampling frame used for the survey, i.e. data that we have for both respondents and non-respondents. Such data can be scarce (like address, name) or richer (including variables not necessary to contact the person, such as nationality, age etc.) or paradata that is also available for respondents and non-respondents (e.g. description of neighbourhood if interviewers drop by).

If we are interested in using data only from a subpopulation, we are interested in whether this subgroup is adequately covered in the survey. For example, if we are interested in migrant women from the global south, we would like to investigate nationality and sex. If we find that female migrants from the global south respond much less, we identify a response bias for a relevant category. We have thus to be cautious for such a use of the data as there might be (but not necessarily has to be) a risk for bias of our estimates. To investigate such a risk for bias, we can simply apply the absolute relative bias in response rates of the interesting categories because we are only interested in one category. For example, if migrant men from Northern Europe are responding much less (or more), it is of no concern as they are excluded from the analysis.

Consider now a policy advisor intending to analyse descriptively the distributions of variables, for example the attitudes towards increasing ecological taxes. Here, the absolute relative bias is not sufficient as we want to have a measure for all categories of the variable because we want to have an estimate for the whole distribution. In this case, the standard deviation of the absolute relative bias in response rates across all categories of age groups is a useful indicator. Distributions are prone to bias when there is a relation between the variable of interest and a variable strongly linked to participation because the distribution is directly affected and interpreted.

If researchers are interested in correlations or means, i.e. not in the exact number within categories but relative relationships between variables, the size of the cells start to matter. If there is a bias in a cell with very few cases, the bias will not affect means or relationships between variables. For example, if we find a considerable response bias regarding religion in the category of Jews, with Jews making up 0.5% of the population, this will not impact means or correlations even if religious affiliation was strongly linked to the variables used. However, if we find a medium response bias regarding Catholics, who make 70% of the population, if the variables analysed are linked to religious affiliation, we are confronted with a high risk for bias of this estimate. In such a case, standard deviations of absolute relative bias will not differentiate. Rather, researchers can use Cramér's V that weighs the deviation of the observed from expected cell size by cell size (chi^2) and standardises it by the number of observations to make it comparable across designs (Cramér, 1946, p. 282).

Many researchers use multivariate analyses for their analyses. Here, they can control in their models for many variables. Thus, if there is a bias on age groups and age groups are correlated with the outcome variable, the bias is accounted for as long as age groups are included in the analysis as the result of a multivariate analysis is the relationship net of age groups (and the other variables in the model). The more variables we have in a model, the better we can control for potential bias. It is thus relevant to know whether the response bias can be explained by other variables. This is the concept of R-indicators: to identify the joint response bias of several variables.

Note that an analysis of response bias can only inform on the response bias of the variables available in the frame. Most of these variables are not relevant to the analysis (as they are already known) or they can be included in the multivariate models applied, thus controlling for response bias. Furthermore, a response bias does not mean that an analysis using other variables from the same data set is biased nor does no identification of bias mean that estimates from an analysis using other variables are unbiased. Therefore, if a bias is found, it needs to be linked to statistics of interest to become interpretable regarding its potential for representation bias.

3.2 CONTROL VARIABLES

Probably the most used technique to investigate representation bias is the comparison between known population data and the achieved sample of a survey. The downside to this approach is that it is only possible for known population parameters while the goal of a survey is to get information about unknown parameters. The idea behind this approach is in a way a “control variable” approach: A society can be divided into different subgroups, such as age groups, sexes, nationality, urban/rural. In analyses for research projects, such subgroups are often used as control variables.

For population data, we know the distribution and compare the distribution to the distribution of the same variable in the survey data. If we are interested in sub-population analysis, similar to the response bias analysis, we are only interested in one category (or a limited number of categories) of the variable. Therefore, if we are interested in pensioners, it does not matter much if we find that the 18-29 old are underrepresented in our sample while the 30-45 old are overrepresented. If the 65+ old make up more or less the same percentage, such is the assumption of this approach, there is a lower risk for bias of our analysis than if there is a big difference. A look at the absolute difference of the percentages divided by the percentage of the cell in the population, which is the absolute relative bias (ARB), thus gives us an indication.

If we are interested in distributions, all categories matter. We thus use the mean of the ARB over all categories of a variable. As with the response bias analysis, this is rarely of interest as we only can identify bias of known parameters. It is not indicative of bias in other variables unless there is a strong correlation between the variable of interest and the control variable. In such a case one would better use more sophisticated analyses to account for those relationships. Still, if a researcher is interested in cross-tabulation of agreement to same-sex marriage across age groups and one finds a difference in the distribution of age groups, the difference would indicate a risk for bias in the distribution of agreement to same-sex marriage across age groups (if there is a correlation between the two).

If a researcher is interested in means or correlations, the size of the cells that are biased become relevant: Equivalent to what was mentioned regarding the likelihood of participation, a bias in a cell that is small, does not necessarily affect means and correlations. If we have 1% instead of 2% of respondents in a registered same-sex partnership, this is quite a large bias but if we assume that they have different levels of trust in institutions, this will still not change the mean of trust in institutions or correlations of trust with other variables as this 1% will not make a difference in our statistic of interest. Therefore, the ARB would overestimate the influence of this bias on our statistic of interest. Consequently, we can weigh the absolute bias of percentages in each category with the cell size of this category and use the average of this as an indicator, i.e., the average cell-size weighted absolute bias of relative frequencies.

In multivariate models, researchers can include control variables to account for confounding factors, one of which might be representation bias. The question, therefore, is equal to that which gave birth to the R-indicators: if we use the variables we have at hand, how much can they control for bias? Or in other words, how much of the bias in one variable is explained by the others? While we cannot calculate a response probability model like for the R-indicators based on population data, we can calculate a weight accounting for the variables we have and calculate a weighted Cramér’s V, based on the data weighted using the population data.

As for the participation variables, the analysis using control variables needs to be put into relation with the statistic(s) of interest. If a bias is detected for a variable, it is only so much a risk for bias as it is correlated to the statistic of interest. Furthermore, instead of simply looking at the difference, it is recommended to test whether the population statistic falls into the 95% confidence interval of the estimate from the data. For this approach, bootstrapping can be used to arrive at a more meaningful 95% “stability” interval, taking into account the distribution of the variable in the data rather than a normal or t-distribution.

3.3 VARIABLES OF INTEREST

Regarding the variables of interest, the indicators are straight forward: it’s the statistic(s) of interest. What is less evident is to find data for this, as we are interested in the statistic when it is unknown, hence when we do not have the data. However, there are possibilities to use data that is similar. For example, we can use older data, or data that only includes a part of the variables in the model that is to be calculated.

In case we are interested in a subpopulation, i.e., only one category (or a range) of a variable, and this variable is neither available in the sampling frame nor for the population, for example people who experienced discrimination, we can simply test whether the percentages are likely to be drawn from the same population, i.e. a t-test between the two estimates. However, this assumes that the two data sets are indeed drawn from the same population. Most of the time, the data sets differ in time, which means that differences can be due to changes over time (it is not the same population). This needs substantial interpretation. A non-parametric way to test how different the two estimates are (not assuming that they are drawn from the same population and relaxing some distributional assumptions) would be to calculate the bootstrapped absolute relative bias between two data sets and calculate the 95% confidence interval. This indicator would give an idea of the size and stability of the difference.

If interested in distributions, a similar approach can be used, this time for all categories of the variable. Instead of just calculating the average of the absolute relative bias, or a chi-square test, the two indicators can be bootstrapped to have a non-parametric test for the stability of the difference.

If the statistics of interest are means or correlations, cell size needs to be taken into account. In the best case, one simply takes the means or correlation of interest and compares them to the benchmark. As we compare two estimates, we can also calculate a bootstrapped difference or the overlap of bootstrapped 95% stability intervals with the benchmark data as an indicator for the importance of the difference.

Finally, if we use multivariate analysis, we run similar models using benchmark data and compare the estimates. Again, bootstrapping can be a useful tool to understand the importance of the difference.

It is important that the indicators used are meaningful regarding the aim of the final analysis. If a researcher is interested in an estimate using a multivariate model, a comparison with benchmark data needs to be interpreted as the benchmark data is either not including all variables or is a previous edition of the data under scrutiny. Differences are thus to be expected and it is to be investigated whether these changes are plausible or not. If a survey practitioner is comparing different designs, bootstrapping might be a useful tool to understand how different the statistics of interest are: the coverage of bootstrapped 95% stability intervals would be a good indicator for such a situation. There are, thus, many options, and the researcher should be creative to find an indicator that is meaningful regarding the purpose of the analysis.

Table 1. Indicator matrix for the analysis of risk for response bias

		Types of Analysis			
Types of Variables		**Sub-population**	**Distributions**	**Correlations; Means etc.**	**Multivariate Analysis**
	Variables for likelihood of participation (sampling frame data)	Absolute Relative Bias of Response Rates (ARBrr)	Overall average ARBrr	Cramér's V	Controlled Cramér's V / R-indicators
	Control Variables (population data)	Absolute Relative Bias (ARB)	Overall average ARB	Average cell-size weighted absolute bias	Weighted Carmér's V
	Variables of Interest (external benchmark data)	Coverage of BS 95% CI (CovBS95CI) of Absolute Relative Bias	CovBS95CI of avg. ABRF / CovBS95CI of Statistic of Interest	CovBS95CI of Statistic of Interest	CovBS95CI of Statistics of Interest

4. USAGE IN THE SWISS CONTEXT

4.1 SAMPLING

In Switzerland, surveys of national relevance can draw their samples from the national sampling frame since 2010. In this case, the sample is a random sample of the residence population and the only issue regarding statistical inference remains with survey non-response and measurement error. However, also in other contexts, researchers might be able to draw a random sample from a full register of their population, for example if the population of the study consists of all members of an association, for example an NGO, and this NGO draws a random sample from their member list.

4.2 POPULATION DATA

The Swiss Federal Statistical Office (SFSO) offers a wide range of population data like the Structural Survey (former census) and STATPOP. The SFSO offers also large-scale surveys like the SILC (Statistics on Income and Living Conditions) and SLSF (Swiss Labour Force Survey) that could be considered as population data as they use large samples based on the Swiss sampling frame and use weights drawn from rich register data.

4.3 EXTERNAL BENCHMARK DATA

There is a rich selection of benchmark data that can be used in analyses of risk for representation bias on variables of interest. Firstly, the Swiss Federal Statistical Office offers a wide range of large-scale surveys, such as SILC, SLSF, but also some smaller surveys, dedicated to specific topics like the International Health Policy Survey (IHP). Furthermore, there are many surveys available in open access to which researchers and survey practitioners can compare their data and examine their plausibility. SWISSUbase offers access to many surveys, such as the Swiss Household Panel, the Swiss Election Study, the European Social Survey, the International Social Survey Program, the European Values Study, SHARE and many more.

5. PRACTICAL EXAMPLES

A framework is necessarily theoretical and abstract. To give the reader an idea how an analysis of risk for representation bias could be approached, I present a few small projects. None of these projects applied the framework as such fully. Quite the opposite. While working on these projects, I developed the ideas for this framework. Nevertheless, the projects show how different the aims of such an analysis can be.

5.1 EVALUATION OF DIFFERENT DESIGNS

Confronted with declining response rates, a Swiss city wanted to test whether and how its general population survey could be transferred from telephone mode to a web or a web/paper mixed mode. In parallel to the regular telephone survey, an experiment was conducted using the same questionnaire and sample frame to field different designs using web and paper modes. The goal was to evaluate whether the response could be improved and whether changing the mode would scotch the longitudinal aspects of the survey. Using sample frame data and population data, the response and the outcome samples were compared and finally, several tests were made regarding relevant variables across different designs. It was concluded that a web/paper mixed mode will improve response and that it will be possible to keep up the time series as in one year, both modes were fielded at the same time (see Ochsner, 2015).

5.2 PRE-FIELDWORK DECISIONS

Survey practitioners are confronted with the fact that societies become more multilingual due to increased migration. Therefore, they need to decide in which languages a survey is best presented. However, adding a survey language not only comes with additional work and cost, it also comes with methodological issues of comparability, for example whether translations are equivalent or whether the meaning of a concept changes when translated into another language. Thus, it is important to be able to estimate whether adding a language adds respondents who are otherwise underrepresented. Lipps and Ochsner (2019) present an approach to evaluate whether adding a survey language has the potential to reduce response bias. Using census or large-scale surveys which include information on language, helps estimating whether people who primarily speak a certain language are more likely to resemble persons who are generally underrepresented. They show that adding English in Switzerland, for example, increases response bias regarding some variables like educational level and can come with the risk of increasing rather than reducing existing biases.

5.3 EVALUATION OF A FOLLOW-UP SURVEY

Some surveys consist of two or more parts as there is a limit in length to what respondents can meaningfully answer at once in a satisfactory quality. However, not all respondents take part in a follow-up survey. This comes with the risk of response bias for those variables included in the follow-up survey. However, we already have quite a lot of information about the participants taking part in the main survey. Therefore, the information can be used to estimate the risk of response bias to the follow-up survey. Usually, the follow-up survey is on a similar topic as the main survey, hence an analysis of risk for bias can make use of a large number of "variables of interest", which helps to identify bias that is relevant to the aims of the data use. Furthermore, surveys often include variables that are known to be related to survey response, such as interest in politics or general trust, which can also be used for the analysis (see, e.g., the documentation for MOSAiCH 2018, Ernst Staehli et al., 2019, pp. 25–30). However, for scientific analyses of data from follow-up surveys, it is highly recommended to use advanced statistical methods to account for potential response bias, such as full information maximum likelihood or multiple imputation. The latter allows to include auxiliary variables that might be related to response in the imputation step (Collins et al., 2001). These methods include the level of uncertainty due to missing information in their estimation of the means and standard errors.

5.4 FULL POPULATION SURVEY OR BIG DATA

As already mentioned above, there are situations in which no random sample is drawn, but all available data is used. One example for such a situation are big data analyses, such as studying Tweets containing specific key words, another example are full population surveys or censuses where all members of a group receive a questionnaire. In such contexts, inference to the population is problematic as in one case (Twitter), we have all data qualifying the selection and thus no inference has to be made, the statistics derived from the data are the true ones for the selection; in the second case (census or full population surveys), similarly, all available data is used and only selection bias due to non-response leads to biased estimates and inference cannot help here. However, researchers still feel that they would like to show "significant" results. While substantial significance is shown by the effect size, statistical significance in the sense of probability that the effect holds true in the population does not make sense as, first, the estimate reflects the true effect in the population and, second, the assumptions for the inference do not hold (no random sample). Yet, what researchers seem to want to indicate is whether the effect size is "higher than chance", or "is stable and not dependent on selection bias". Such questions can be addressed in a simple way by applying bootstrapping to provide "stability intervals". This procedure selects a subsample of the data and calculates the estimate. This procedure is repeated several times, usually 1'000 times, and from the results, 95% stability intervals are calculated, i.e., the range in which 95% of the estimates fall. This gives an idea of how dependent the estimate is on selection of sample members. Ochsner et al. (2017) provide an example for such a strategy regarding a full population survey on the notions of research quality of humanities scholars in Switzerland.

6. IMPLICATIONS FOR SURVEY PRACTITIONERS

To conclude, a few recommendations can be drawn from this analysis for researchers and survey practitioners aiming at analysing the risk for representation bias of their data.

Recommendation 1 – Avoid the term "representative". If it needs to be used, explain clearly what is meant, revealing the context for which the statement is made. Only use it when it refers to probability sampling and do not make a general claim.

Recommendation 2 – Be creative. Instead of trusting one indicator, use several indicators linked to the analysis that is or will be made.

Recommendation 3 – Be specific. If having to inform generally on a data set, cover multiple uses of the data, never make general claims and base recommendations on the findings of the analysis.

Recommendation 4 – Be prudent. Reflect possible biases with regard to results of substantive analyses.

Recommendation 5 – Be scientific. Take plausible assumptions, be consistent, be simple and comprehensible, do not over-generalise, remain within the scope of the analysis.

Recommendation 6 – Stay focused. Keep an eye on what the goal is; what is the correlation of the test variables with the variables and statistics of interest? This correlation sets the limits of influence of the test variables on the statistics of interest.

Recommendation 7 – Be inclusive. Use as much information as is available. Whenever possible use advanced statistical models to account for uncertainty due to (unit and item) nonresponse, such as full information maximum likelihood or multiple imputation.

Recommendation 8 – Big data are not representative for a general population. It is usually not the goal of analysing big data to draw conclusions regarding the general population. Rather, it is the analysis of all available data on a subject matter. It is not a sample and certainly not a probabilistic one, therefore inference cannot be made. Big data is very useful but not for claims regarding the general population. For example, an analysis of gender-neutral pronouns in Twitter data is very interesting but does not reflect the use of gender-neutral pronouns in other contexts, nor does the same analysis using the complete set of the most prestigious newspaper articles in the same time period. However, such data can be used to start formulating theories regarding the general population which then is studied using other data; or the results from Twitter and the newspaper can be compared and interpreted fully. This is highly interesting, but the concept of representativeness does not make sense in such contexts.

7. FURTHER READINGS AND USEFUL WEB LINKS

As mentioned throughout this FORS Guide, the three articles by Kruskal and Mosteller (1979a, b, c) are excellent starting points for any introduction into the topic, they are to the point, clear and entertaining to read.

For the demystification of response rate as quality indicator see Groves (2006) and Peytcheva and Groves (2009) and for overviews on different indicators used for the analysis of non-response bias, see Wagner (2012).

For different approaches how to address non-response bias, see Groves et al. (2002). For the use of multiple imputation see Collins et al. (2001) or Schafer and Graham (2002).

In Switzerland, the Swiss Federal Statistical Office offers a wealth of useful data sources, SWISSUbase also offers many surveys providing high quality data for benchmarking.

- Information on the Swiss Sampling Frame can be found here: https://www.bfs.admin.ch/bfs/en/home/basics/census/natonal-census-integrated-system/sampling-frame.html

- STAT-TAB offers an interactive interface to identify and download data on several topics, including STAT-POP, SLFS: https://www.pxweb.bfs.admin.ch/pxweb/en/
- Information on the Statistics on Income and Living Conditions (SILC) survey can be found here: https://www.bfs.admin.ch/bfs/en/home/statistics/economic-social-situation-population/surveys/silc.html
- SWISSUbase offers access to high quality surveys here: https://www.swissubase.ch

REFERENCES

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly, 74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. John Wiley & Sons. https://doi.org/10.1002/0471458740

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press

Eckman, S. (2016). Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?. *Social Science Computer Review, 34*(1), 41–58. https://doi.org/10.1177/0894439315572985

Ernst Stähli, M., Sapin, M., Pollien, A., Ochsner, M., Nisple, K., & Joye, D. (2019). *MOSAiCH 2018: Study on Religion and related topics*. [Dataset]. FORS – Swiss Centre of Expertise in the Social Sciences. https://doi.org/10.23662/FORS-DS-962-3

European Social Survey (ESS). (2017). *Round 9 Survey Specification for ESS ERIC Member, Observer and Guest Countries*. European Social Survey.

Gilbert, J. P., Light, R. J., & Mosteller, F. (1977). Assessing social innovations: an empirical base for policy. In W. B. Fairley & F. Mosteller (Eds.), *Statistics and policy* (pp. 185–242). Addison-Wesley.

Good, P. I., & Hardin, J. W. (2009). *Common errors in statistics (and how to avoid them)*. Wiley-Interscience.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646–675.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2002). *Survey Nonresponse*. Wiley-Interscience.

Koch, A., & Blohm, M. (2016). *Nonresponse Bias*. GESIS Survey Guidelines.

Kruskal, W. H., & Mosteller, F. (1979a). Representative Sampling, I: Non-Scientific Literature. *International Statistical Review / Revue Internationale De Statistique, 47*(1), 13. https://doi.org/10.2307/1403202

Kruskal, W. H., & Mosteller, F. (1979b). Representative Sampling, II: Scientific Literature, Excluding Statistics. *International Statistical Review / Revue Internationale De Statistique, 47*(2), 111. https://doi.org/10.2307/1402564

Kruskal, W. H., & Mosteller, F. (1979c). Representative Sampling, III: The Current Statistical Literature. *International Statistical Review / Revue Internationale De Statistique, 47*(3), 245. https://doi.org/10.2307/1402647

Lipps, O., & Ochsner, M. (2019). Additional Languages and Representativeness. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods. Multinational, Multiregional, and Multicultural Contexts (3MC*) (pp. 859–877). Wiley.

Lipps, O., & Pekari, N. (2016). Sample Representation and Substantive Outcomes Using Web With and Without Incentives Compared to Telephone in an Election Survey. *Journal of Official Statistics, 32*(1), 165–186. https://doi.org/10.1515/JOS-2016-0008

Lipps, O., Pekari, N., & Roberts, C. (2015). Undercoverage and Nonresponse in a List-sampled Telephone Election Survey. *Survey Research Methods, 9*(2), 71–82. https://doi.org/10.18148/srm/2015.v9i2.6139

Moore, J. C., Durrant, G. B., & Smith, P. W. F. (2018). Data set representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistical Society, Series A – Statistics in Society, 181*(1), 229–248. https://doi.org/10.1111/rssa.12256

Ochsner, M. (2015). *Systematische Methodenevaluation im Rahmen der Bevölkerungsbefragung der Stadt Zürich. Analyse und Empfehlungen*. FORS – Swiss Centre of Expertise in the Social Sciences.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2017). Assessment Criteria for Early Career Researcher's Proposals in the Humanities. In *Proceedings of the 16th International Conference on Scientometrics and Informetrics, 16.–20. October 2017, Wuhan* (pp. 105–111). Wuhan University.

Peytcheva, E., & Groves, R. M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. *Journal of Official Statistics, 25*(2), 193–201.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology, 35*(1), 101–113. https://doi.org/0894439317735307

Wagner, J. (2012). A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly, 76*(3), 555–575. https://doi.org/10.1093/poq/nfs032

Wilcox, J. B., Bellenger, D. N., & Rigdon, E. E. (1994). Assessing Sample Representativeness in Industrial Surveys. *Journal of Business & Industrial Marketing, 9*(2), 51–61. https://doi.org/10.1108/08858629410059834