Nonresponse Bias Analysis

Barbara Felderer

2024, Version 1.0

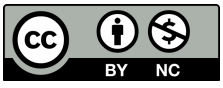
Abstract

Declining response rates increase the fear of nonresponse bias. This guideline discusses the relationship between nonresponse and nonrespones bias and gives an overview of indicators that are frequently used to determine the risk of nonresponse bias. The indicators are illustrated in a simulated data example.

Citation

Barbara Felderer (2024). Nonresponse Bias Analysis. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines).

DOI: 10.15465/gesis-sg_en_047



1. Introduction

Declining response rates all over the world increase the fear of nonresponse bias, i.e., that the respondents to a survey do not well represent the group of individuals who has been invited to participate in that survey. In the presence of nonresponse bias, raw survey estimates can not be used to draw valid conclusions on the population of interest.

High nonresponse does not necessarily imply high nonresponse bias. This guideline discusses the relationship between nonresponse and nonresponse bias and gives an overview of methods to determine nonresponse bias for a specific survey or survey variable of interest. Talking about survey nonresponse, we can in general distinguish between item nonresponse and unit nonresponse. Unit nonresponse means that an individual who is sampled and invited to participate in a survey does not participate in that survey at all. Item nonresponse occurs if an interviewed person does not give an answer to a specific question. This guideline captures unit nonresponse, for a discussion of the handling of item nonresponse we refer to the GESIS survey guideline on imputation (Bruch, 2023). Adjustment methods might be applied to reduce nonresponse bias but they only work under certain conditions that are discussed in this guideline. This guideline does, however, not address the treatment nor the prevention of nonresponse bias. For the former, we recommend the survey guidelines on weighting (Gabler, Kolb, Sand, & Zins, 2015; Sand & Kunz, 2020), for the latter the survey guideline on nonresponse bias (Koch & Blohm, 2015).

Nonresponse is by far not the only source of potential survey error (Groves & Lyberg, 2010). For simplicity, this guideline ignores all other sources of error, for example, we assume simple random sampling and measurements without error or item nonresponse. The next section discusses the relationship between survey nonresponse and nonresponse bias. Several univariate and multivariate indicators for the risk of nonresponse bias are discussed in Section 3. In Section 4, we illustrate some of the nonresponse bias indicators using a synthetic data example. A R-file to replicate the example is made available online. We conclude with a discussion in Section 5.

2. Relationship of survey nonresponse and nonresponse bias

In the following, we denote the survey variables by X, Y, Z, and the matrix and vector of their observed values by x, y, z. Unit nonresponse can occur for several reasons and is not necessarily a problem for the quality of the survey. Let Y be the survey variable of interest that one plans to analyse, X and Z be two distinct sets of personal characteristics of the invited individual and $\rho \in [0, 1]$ be response propensities.

2.1 Nonresponse mechanisms

Groves (2006) distinguishes between three nonresponse mechanisms that can be explained by different models:

Separate Cause Model: The response propensity ρ depends on personal characteristics Z that are not associated with the variable of interest Y. Y is associated with personal characteristics X that do not affect the response propensity ρ. In this situation, Y and ρ are not associated. This means that nonresponse does not lead to nonresponse bias in the analysis of Y.

As Groves (2006) notes, completely unrelated causes are hard to imagine in practice. The separate cause model is, however, very useful when thinking of the relationship of nonresponse and nonresponse bias and to contrast the other models against it.

Common Cause Model: The same individual characteristics Z affect the response propensity ρ and the

variable of interest Y. The common cause generates an association between ρ and Y thus potentially biasing the analysis of Y.

If Z is known for respondents and nonresponents, it can be used to perform nonresponse adjustments in the analysis of the survey variable Y and to reduce nonresponse bias. Like the separate cause model, the common cause model is a simplified model. In practical applications, there will most likely be unboserved Z-variables that can not be included in the nonresponse adjustment.

Survey Variable Cause Model: The variable of interest Y directly affects the response propensity ρ. Since ρ and Y are associated, analysis of Y will suffer from nonresponse bias, and this can not be completely removed by any weighting or adjustment method.

It is important to note that different nonresponse models might hold for different variables Y of the same survey. The nonresponse mechanism and thus nonresponse bias is always variable-specific.

2.2 Different perspectives on nonresponse bias

Nonresponse bias can be viewed from several perspectives that highlight different facets.

We differentiate between three groups: The overall target population (with size N^*), the sampled individuals (with size N), and the survey respondents (with size n), where $N^* > N \geq n$. In this guideline, we assume that the survey sample is randomly drawn from the target population. For ease of exposition we further assume a simple random sample, i.e., sampling with equal inclusion probabilities.

With y_i being the value for survey variable Y for individual i, the population mean of Y is given by $\bar{y}_P = 1/N^* \sum_{i=1}^{N^*} y_i$, the mean of the sampled individuals by $\bar{y}_S = 1/N \sum_{i=1}^{N} y_i$, and the mean of the survey respondents by $\bar{y}_R = 1/n \sum_{i=1}^{n} y_i$. For more complex survey designs with unequal inclusion probabilities, survey estimates must to be design-weighted.

Nonresponse bias (NRB) in the estimated mean of a survey variable Y is given by the difference between the mean value of the survey respondents $\bar{y}_R$ and the mean of the target population $\bar{y}_P$. Assuming random sampling, $\bar{y}_P = \bar{y}_S$ holds, such that

$$NRB_Y = \bar{y}_R - \bar{y}_S. \tag{1}$$

As can easily be seen, we do not have to expect nonresponse bias in $\bar{y}_R$ if respondents do not differ from the target population in Y on average. Nonresponse bias gets larger as the difference increases.

Looking at this relationship more closely, usually the deterministic and the stochastic view on nonresponse bias are distinguished. Even though they refer to the exact same concept, they highlight different aspects making it worth to look at both of them.

The *deterministic* view on nonresponse bias is given by (see for example Groves, 2006):

$$NRB_Y = (1 - RR)(\bar{y}_R - \bar{y}_{NR}), \tag{2}$$

where $RR = \frac{n}{N}$ is the response rate and the mean of the nonrespondents is given by $\bar{y}_{NR}$. Nonresponse bias is affected by the response rate and the difference between means for respondents and nonrespondents. This means two things: For a given difference between respondents and nonrespondents, an increasing response rate will lower nonresponse bias. For a given response rate, lower differences between

respondents and nonrespondents lead to lower nonresponse bias. If respondents and nonrespondents do not differ in Y at all, no nonresponse bias is to be expected in Y.

The *stochastic* approach takes the perspective that participants are not determined to be either respondents or nonrespondents, but characterized by a latent, stochastic propensity to respond. Taking this approach, nonresponse bias is computed over all elements of the target population, weighted by their unobserved response propensities ρ_i. This gives rise to the following definition (Bethlehem, 1988):

$$\begin{aligned} NRB_Y &\approx \frac{1}{\bar{\rho}} Cov(y, \rho) \\ &\approx \frac{1}{\bar{\rho}} Cor(y, \rho) \sigma_y \sigma_\rho \end{aligned} \tag{3}$$

where ρ is the vector (of length N^*) of response propensities with population mean $\bar{\rho}$. The population covariance of y and ρ is given by $Cov(y, \rho) = 1/N^* \sum_{i=1}^{N^*}(y_i - \bar{y}_P)(\rho_i - \bar{\rho})$. The population standard deviations of y and ρ are $\sigma_y = \sqrt{1/N^* \sum_{i=1}^{N}(y_i - \bar{y})^2}$ and $\sigma_\rho = \sqrt{1/N^* \sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}$, and $Cor(y, \rho) = \frac{Cov(y,\rho)}{\sigma_y \sigma_\rho}$ is the population correlation of y and ρ.

As the stochastic view highlights, nonresponse bias decreases with increasing $\bar{\rho}$ (which corresponds to the response rate), decreasing correlation of Y and ρ, and decreasing standard deviations of both ρ and y. If $Cor(y, \rho) = 0$, i.e., if the nonresponse mechanism is not related to Y at all, no nonresponse bias is to be expected. The same is true if $\sigma_\rho = 0$ (all individuals have the same propensity to respond) or $\sigma_y = 0$ (all individuals have the same value of Y).

2.3 Why is the response rate alone not a reliable indicator of nonresponse bias?

Surveys often report the response rate as an indicator for the quality of the survey. As can be seen from the formulas above, the response rate is part of the deterministic and stochastic perspective on nonresponse bias. Keeping the other factors constant, nonresponse bias is lower the higher the response rate is. There is, however, no clear relationship between the response rate of a survey and the other factors that constitute nonresponse bias. The response rate alone does not allow for an evaluation of nonresponse bias and is thus not a good nonresponse bias indicator (for empirical findings see for example Groves, 2006; Groves & Peytcheva, 2008).

3. Nonresponse bias analysis

Many methods to examine nonresponse bias have been developed based on Equations (1) to (3). There are too many methods available to be covered in this survey guideline. Thus, the following sections focus on the most frequently used ones.

When conducting nonresponse bias analysis, we need to estimate some of the components of the indicators that have been introduced above. In the survey methodological literature, the term representativeness is widely used to describe the quality of a survey. Representativeness is, however, not clearly defined and may address many different aspects. Focusing on nonresponse bias and assuming random sampling in this guideline, we will call a survey representative if it is not subject to nonresponse bias.

3.1 Components of nonresponse bias analysis

Many of the parameters discussed in Equations (1) to (3) are not known but can be estimated based on the survey information. The mean of the respondents can be estimated by the survey mean $\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$ where y_i ($i = 1, \ldots, n$) is the survey value for the i^{th} respondent. The response propensity of the invited individuals ρ is not known. In many cases, auxiliary variables X are available for respondents and nonrespondents that can be used to estimate the response propensity. They might be available from the sample frame (e.g., age and gender from official registers), administrative data, paradata from the sampling or recruitment process or, in the panel context, be survey answers from previous survey waves. To estimate the response propensity ρ, the participation indicator R (R=1 if the individual responds to the survey and zero otherwise) is regressed on multiple (v) X-variables, commonly using logistic regression such as

$$\hat{\rho}_i = P(R_i = 1) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{iv})}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{iv})} \tag{4}$$

where $\hat{\rho}_i$ is the estimated response propensity for the i^{th} individual, $i = 1 \ldots N$ (the full sample), $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1 \ldots \hat{\beta}_v$ are the slopes for the observed auxiliary variables $X_1 \ldots X_v$; $x_{i1} \ldots x_{iv}$ are the values of the X-variables of individual i. For large data sets, machine learning methods might be preferred over standard logistic regression, see for example Felderer, Kueck, & Spindler (2023).

The population parameter $\bar{y}_p$ is usually not known – that is why we conduct the survey in the first place – and can not be estimated from the survey. The same is true for the parameters $\bar{y}_S$ and $\bar{y}_{NR}$.

There are several indicators available to evaluate the risk of nonresponse bias that can be roughly put into two categories: Indicators that refer to the risk of nonresponse bias of a whole survey and indicators that focus on specific survey variables. For the latter, one can distinguish indicators that basically refer to auxiliary variables and indicators that refer to the variable of interest.

Nonresponse bias as introduced above can usually only be estimated for X-variables that are known for respondents and nonrespondents or for which population benchmarks are available. The Y-variable is usually unobserved for the nonrespondents and lacks a population benchmark. We thus can not study *nonresponse bias* in the Y-variable directly but rather the *risk of nonresponse bias* that we derive from knowledge about nonresponse bias in the X-variables and the relation between X and Y.

Many indicators that we introduce in the following section are generated to apporach nonresponse bias in different ways refering to single aspects of equation (3). They consequently do not estimate nonresponse bias in the strict sense but rather the risk of nonresponse bias (in X or in Y).

3.2 Multivariate nonresponse bias indicators

Several measures for the risk of nonresponse bias have been developed that are based on auxiliary information on respondents and nonrespondents. At their essence, these measures attempt to estimate the extent to which individuals in the survey resemble those in the gross sample or population with respect to the auxiliary X- variables. These results are then used to infer possible bias in the Y- variables of interest. The usefulness of the indicators to evaluate the risk of nonresponse bias in a specific variable of interest heavily depends on the association of this variable and the auxiliary variables. If both variables are not related at all, nonresponse bias in X is not a good indicator for nonresponse bias in Y. The stronger the variables are related, the more we expect Y to show nonresponse bias if X does. The following indicators are multivariate in a way that they account for several auxiliary variables and their relationships in

preducing survey nonresponse. Two of them allow to analyse the effect of nonresponse on specific survey variables of interest. The interpretation of the multivariate nonresponse bias indicators is limited to the specific auxiliary variables that they are built on. An excellent comparison of indicators for the risk of nonresponse bias can be found in Wagner (2012).

R-indicator

The R-indicator $R(\rho)$ (Schouten, Cobben, & Bethlehem, 2009) takes the variation of the response propensity ρ as a measure for the risk of nonresponse bias. With σ_ρ being the population standard deviation of ρ, the R-indicator is given by:

$$R(\rho) = 1 - 2\sigma_\rho \tag{5}$$

The R-indicator can take values between 0 and 1. If all sampled individuals have the same propensity to respond, the standard deviation of the response propensities σ_ρ is 0 and the R-indicator takes the value of 1. This means that the survey is perfectly "representative". The R- indicator is zero if $\sigma_\rho = 0.5$ which is the highest value σ_ρ can take. This means that the response propensities are very different and the risk of nonresponse bias is high. Only if the auxiliary variables are correlated with both ρ and the variable of interest, the R-indicator can give a good impression on the risk of nonresponse bias in the variable of interest.

In practical applications, the R-indicator can be estimated by $\hat{R}(\hat{\rho}) = 1 - 2\hat{\sigma}_{\hat{\rho}}$ where $\hat{\rho}_i$ are the estimated response propensities based on auxiliary information as described in Equation (4). Its standard deviation is estimated by $\hat{\sigma}_{\hat{\rho}} = \frac{1}{N}\sum_{i=1}^{n}(\hat{\rho}_i - \hat{\bar{\rho}})^2$ where $\hat{\rho}_i$ is the estimated response propensity for the i^{th} individual and $\hat{\bar{\rho}}$ is the mean of the estimated propensities.

The R-indicator gives an impression whether the different population subgroups, characterized by a combination of X-variables, are well represented in the survey. The R-indicator does not allow to determine how the individual auxiliary variables that are used to estimate the R-indicator contribute to representativeness. Partial R-indicators have been developed to overcome this limitation (Schouten, Shlomo, & Skinner, 2010).

As for all multivariate nonresponse bias indicators, the usefulness of the R-indicator depends on the auxiliary variables that are used to estimate $\hat{\rho}$. The R-indicator can only be interpreted with regard to the auxiliary information that it builds on (see for example Roberts, Vandenplas, & Herzing, 2020). R-Indicators are frequently used to compare different (sub)-samples. Comparisons of R-Indicators are only meaningful if the R-indicators are build in the exact same way including the same auxiliary variables. Higher values of the R-indicators indicate better representativeness. However, there is no agreement in the literature on the threshold value above which one can speak of good representativeness. As a guide, an R-indicator of 0.7 is considered to be rather low (Lugtig, Roth, Schouten, et al., 2022).

Goodness of fit of the propensity model

The propensity model for $\hat{\rho}_i$ (see Equation (4)) can be further analysed. The coefficients of the X-variables allow for an interpretation of which variables influence the response propensity. If the response propensity does not depend on X-variables, we expect all coefficients to be insignificant and close to zero. Measures for the goodness of fit of the propensity model, like the (pseudo) R^2 or the area under the curve (AUC) are taken as indicators for the risk of nonresponse bias (see for example Groves et al., 2008). Higher values in these measures means that more variation in $\hat{\rho}$ can be explained by the X-variables. This means

that respondents and nonrespondents differ for their values of X. A higher goodness of fit of the propensity model thus is taken as an indicator for a higher risk of nonresponse bias in estimates of the variable of interest.

Like the R-indicator, the goodness of fit of the propensity model, however, is only a good indicator for the risk of nonresponse bias in the variable of interest if this variable is associated with X. Finding that the *observed* X-variables that are included in the propensity model do not explain variation in $\hat{\rho}$ does, on the other side, not necessarily mean that the risk of nonresponse bias is low as there may exist *unobserved* variables that systematically affect nonresponse.

Variation of nonresponse weights

Propensity models like in Equation (4) can be used to create adjustment/nonresponse weights by taking the inverse of the estimated response propensity ($1/\hat{\rho}_i$) (Little, 1986). The rough idea behind adjustment methods is to give the groups of respondents who have a lower propensity to respond to the survey (given X-variables) a higher weight in the analysis of the survey data.

The variance of nonresponse weights can be taken as an indicator for the risk of nonresponse bias (Groves et al., 2008): If all individuals have the same propensity to respond, the variance of the nonresponse weights is zero. The higher the variance is, the larger are the differences in the response propensities and the higher the risk of nonresponse bias. This sort of analysis can be conducted for other kinds of nonresponse weighting methods as well and is not limited to propensity weights. Of course, the limitations concerning the relations between X-variables and Y that we discussed above also hold for this and other multivariate nonresponse bias indicators.

Correlation between nonresponse weights and Y

The correlation between the nonresponse weights and observed Y is an attempt to estimate the association of the auxiliary variables and the response propensity and Y (Groves et al., 2008). The association can only be estimated for respondents and relies on the assumption that the association is the same for respondents and the full sample. A higher correlation between nonresponse weights and Y indicates a higher risk of nonresponse bias. In this case, we will also observe differences between weighted and unweighted means and proportions of Y (see Gabler et al. (2015) and Sand & Kunz (2020) on the theoretical background and application on different kinds of nonresponse weights). The correlation between nonresponse weights and Y is an indicator for nonresponse bias *before* weighting adjustment. It does not allow any conclusions about nonresponse bias of the adjusted statistics or the usefulness of the weights. The correlation between nonresponse weights and Y is only a meaningful indicator for nonresponse bias under the MAR-assumption.

Fraction of missing information

The fraction of missing information (FMI) was developed in the multiple imputation context (see Rubin, 1987) dealing with item nonresponse and has been transferred to unit nonresponse bias analysis (Wagner, 2010). In the context of unit nonresponse, the missing survey information of nonrespondents is imputed using auxiliary data that is available for respondents and nonrespondents. The idea is to take the level of uncertainty when imputing the missing values for the variable of interest Y based on auxiliary (complete) X as an indicator for nonresponse bias. For the FMI, the missing observations in the Y-variables are imputed multiple (M) times. The mean or proportion of Y is then estimated based on the fully imputed data set (including observed values for respondents and imputed values for nonrespondents) for each imputation round m.

With M imputations for each missing value, $\bar{y}$ is estimated by $\hat{\bar{y}}_M = \sum_{m=1}^{M} \hat{\bar{y}}_m / M$ where $\hat{\bar{y}}_m$ is the estimated mean in the m_{th} imputed data set. The FMI gives a measure of uncertainty about the imputed values by computing the relation between the between-imputation variance and the total variance of $\hat{\bar{y}}_M$. Let $Var_m(\hat{\bar{y}}_m)$ be the variance of $\hat{\bar{y}}_m$ in the m_{th} imputed data set. The within-imputation variance is given by $Var_W(\hat{\bar{y}}_M) = \sum_{m=1}^{M} Var_m(\hat{\bar{y}}_m)/M$. The within-imputation variance describes the variance that is due to sampling. The between-imputation variance is given by $Var_B(\hat{\bar{y}}_M) = \frac{\sum_{m=1}^{M}(\hat{\bar{y}}_m - \hat{\bar{y}}_M)^2}{(M-1)}$ and is the part of variation that is due to imputation uncertainty. The total variance is given by $Var(\hat{\bar{y}}_M) = Var_W(\hat{\bar{y}}_M) + (M+1)M^{-1}Var_B(\hat{\bar{y}}_M)$.

The FMI is given as the ratio of the between-imputation and total variance and is estimated as

$$\widehat{FMI} = \frac{(1 + \frac{1}{M})Var_B(\hat{\bar{y}}_M)}{Var(\hat{\bar{y}}_M)} \tag{6}$$

The FMI ranges from 0 to 1 and can be interpreted as the proportion of variation in the estimation of $\bar{y}$ that is due to the missing data. The higher the uncertainty about the values of Y given X, the more different are the imputed values between the imputation rounds and the higher is the between-imputation variance. Or, in other words, if the data includes good predictors for Y, the between-imputation variance decreases (Rubin (1987)). A larger $\hat{FMI}$ is thus interpreted to indicate a higher risk of nonresponse bias in the estimation of $\hat{\bar{y}}$ caused by other variables than included in the imputation process.

If the imputation model perfectly explains Y, there are no differences expected between the imputation rounds and the between-imputation variance is close to zero. As Wagner (2010) shows, for correctly specified imputation models, the FMI is close to the nonresponse rate if Y and X are only weakly correlated and moves toward zero as the correlation increases. Highly correlated X variables can thus recover the missing information in Y.

As Andridge & Little (2011) note the FMI has the disadvantage to focus more on precision than on bias and is limited to MAR situations. The FMI also depends on the imputation methods and the specification of the imputation model (Wagner, 2010).

3.3 Univariate nonresponse bias indicators

The multivariate nonresponse bias indicators introduced in the previous section are measures including several X-variables and their relationships. With the exception of the interpretation of the coefficients of the nonresponse propensity model, they do not allow for an interpretation which specific X variables contribute to bias and in which direction. Although it is possible to run nonresponse models with just a single variable and use this model to estimate the nonresponse bias indicators discussed in the previous section, other methods are better suited to look at the effects of individual variables. Such univariate nonresponse bias indicators give more detailed information on nonresponse bias of certain variables and can be used to evaluate which of the auxiliary variables affect the risk of nonresponse bias in the variables of interest. This is done by comparing respondents to official population benchmarks or respondents to nonrespondents for these auxiliary variables. Univariate nonresponse bias indicators do not account for the interplay of different auxiliary variables. Like the nonresponse bias indicators discussed in the previous section, univariate nonresponse bias indicators are only useful under the assumption that the auxiliary variables are related to the variable of interest.

Benchmark comparisons

One way to evaluate nonresponse bias is to compare survey findings to benchmarks from official statistics for the general population (see for example Felderer, Kirchner, & Kreuter, 2019; Rohr, Silber, & Felderer, 2023). For surveys on the German population, the distribution of socio-demographic characteristics among the respondents to a survey might, for example, be compared to official statistics like the Mikrozensus or to a high-quality survey of the German population like the ALLBUS. If the survey respondents are similar to the general population, the risk of nonresponse bias is assumed to be low. Benchmark comparisons thereby implicitly rely on the assumption that the survey estimate $\hat{\bar{x}}$ is not subject to bias apart from nonresponse. Other sources of bias like sampling or coverage error are neglected. The advantage of benchmark comparisons is that no information on nonrespondents are needed. However, the number of variables that can be compared against benchmarks is usually rather limited. Usually, no benchmarks are available for the variables of interest Y.

Let $\hat{\bar{x}}$ be the survey mean and $\bar{x}_{bench}$ the population benchmark for the auxiliary variables. To be able to compare nonresponse bias over variables and between surveys, the relative nonresponse bias for a variable x_j is estimated as[1]

$$\widehat{rel.Bias}(\bar{x}_j) = \frac{\hat{\bar{x}}_j - \bar{x}_{j\,bench}}{\bar{x}_{j\,bench}}. \tag{7}$$

In practice, the differences between $\hat{\bar{x}}$ and $\bar{x}_{bench}$ will be non-zero due to random sampling variation. Appropriate statistical tests can be performed to evaluate statistical significance of the differences (see for example Eckman, Unangst, Dever, & Antoun, 2022; Felderer et al., 2019).

In practical applications, the relative biases are often estimated for a number of available auxiliary variables. All relative biases might be aggregated to one single measure that can be compared across surveys or experimental subgroups of a survey. A commonly reported measure is the average absolute relative bias (AARB) (see for example Cornesse, Felderer, Fikel, Krieger, & Blom, 2021; Friedel, Felderer, Krieger, Cornesse, & Blom, 2023) which is given by the mean of the absolute relative biases. The absolute values are taken to make sure that negative and positive relative biases do not cancel each other out. For v auxiliary variables the AARB is given by

$$\widehat{AARB} = \frac{1}{v}\sum_{j=1}^{v}\left|\frac{\bar{x}_j - \bar{x}_{bench_j}}{\bar{x}_{bench_j}}\right| \tag{8}$$

with subscripts $j = 1 \ldots v$ indicating the j^{th} X-variable. Relative biases can also be aggregated to median absolute relative bias and the maximum absolute relative bias. These aggregated measures have the advantage to reduce the findings for several X-variables to one single value but they are not a multivariate indicator in our sense as they do not account for the interplay of different X-variables. As they do not allow to identify the contribution of the individual characteristics, we recommend to not only analyse the aggregate measures but also their single components.

The AARB is not standardized and its interpretation is only meaningful in comparison to relative biases found in other surveys or for experimental subgroups of the same survey. In order to meaningfully compare AARBs between surveys, one needs to make sure that the same set of auxiliary variables (that are

[1] For categorical variables, proportions are used instead of means.

measured and coded in the same way) are included in the analysis. High nonresponse bias in the auxiliary variables might indicate a high risk of nonresponse bias in the variables of interest. If both kinds of variables are not correlated, even high nonresponse bias in auxiliary variables is no indication of nonresponse bias in the survey variables of interest.

Other measures that are based on benchmark comparisons are the Duncan dissimilarity index (for example Bosnjak et al., 2018) and absolute difference or standardized absolute difference (for example Peytcheva & Groves, 2009).

Comparison of respondents and nonrespondents on auxiliary variables

Usually,the number of survey variables that can be compared to official statistics is very limited. In many applications, however, auxiliary variables X from the sample frame, paradata from the fieldwork process (for example Krueger & West, 2014) or, in a panel context, survey information from previous waves are available for all individuals who are invited to participate in a survey. Comparisons between respondents and nonrespondents can be performed by comparing means and proportions and determining significance using appropriate statistical tests.

Comparison of early and late respondents

A comparison of early or "easy-to-contact" respondents to late or "hard-to-contact" respondents is sometimes used to get an impression of the risk of nonresponse bias (see for example Green, 1991). Doing this, the late respondents are assumed to be similar to nonrespondents. While this approach relies on strong assumptions, it has the advantage that it can be performed on auxiliary variables as well as the survey variables of interest. To receive information on nonrespondents, sometimes a nonresponse follow-up survey (see for example Roberts et al., 2020) is conducted using a shortened questionnaire. The respondents to the main survey are then compared to the respondents of the nonresponse follow-up survey assuming that the latter are representative of all nonrespondents to the main survey.

Variation of subgroup response rates

The evaluation of the variation of subgroup response rates is a univariate indicator that follows the same idea as the (multivariate) evaluation of the goodness of fit and nonresponse weights (see for example Wagner, 2012). If different subgroups (defined by the categories of a specific auxiliary variable) show different response rates, this is taken as an indication for an increased risk of nonresponse bias in the survey variable of interest. For a categorical variable with c categories the subgroup response rates are given as $RR_{sub,c} = \frac{n_c}{N_c}$ where n_c is the number of respondents in category c and N_c the number of sampled persons in category c. RR_{sub} is the vector of the subgroup response rates for all categories of a specific variables. The variance of subgroup response rates can be estimated by:

$$\hat{Var}(RR_{sub}) = \frac{1}{N-1}\sum_{c=1}^{C} N_c\left(\frac{n_c}{N_c} - \frac{n}{N}\right)^2 \tag{9}$$

where $\frac{n}{N}$ is the survey's response rate. The subgroup response rate does not require information on nonrespondents on an individual level but only sample (or population) proportions of the auxiliary variable. To standardize subgroup response rates, the coefficient of variation of the subgroup response rates (see for example Nishimura, Wagner, & Elliott, 2016) can be calculated as

$$\hat{CV}(RR_{sub}) = \frac{\hat{Var}(RR_{sub})}{n/N}. \tag{10}$$

Higher values of $\hat{CV}(RR_{sub})$ are taken as a higher risk of nonresponse bias. $\hat{Var}(RR_{sub})$ is only a useful indicator if X is associated with Y.

4. Illustration of use of nonresponse bias indicators

To illustrate the use of the measures described above, we apply them to synthetic data sets that are generated to match the three nonresponse mechanisms discussed in section 2.1. The approach was strongly inspired by Nishimura et al. (2016). The details of the synthetic data example set up can be found in the attached R-script. We create five X-variables and one Y-variable (Y_1, Y_2, Y_3) for each mechanism. The X-variables are generated to match the distribution in the German population according to the Mikrozensus 2019. The response mechanism roughly mimics typical findings for *age, gender, education, household size* and *German nationality*. For example, sampled individuals who are highly educated participate more often than the less educated ones. Lastly, a variable Z is generated that can be seen as an unobserved variable that is related to Y_1 and Y_3. A random error term is added to the generation of all Y- variables and response propensities to avoid perfect relations between them and X and Z.

For the *separate cause model*, the response propensity is generated to be a function of all X-variables ($\rho = f(X_1, \ldots, X_5)$) while Y_1 is a function of one Z-variable and does not depend on any X ($Y_1 = f(Z)$).

For the *common cause model*, the response propensity is again a function of all X-variables. The variable Y_2 is a function of three of the five X-variables ($Y_2 = f(X_1, X_2, X_3)$) which resemble gender, age and education.

For the *survey variable cause model*, neither Y_3 nor the response propensity depend on the X-variables. Y_3 depends on variable Z ($Y_3 = f(Z)$) and the response propensity is associated with the survey variable of interest ($\rho = f(Y_3)$).

For each scenario, samples of size $N = 2000$ are generated with response rates of 50%. All individuals who have a response propensity above the median response rate are considered respondents to the survey. As nonresponse depends on the same set of variables in the *separate cause model* example and the *common cause model* example Y_1 and Y_2 can be interpreted as being collected in the same survey. The variable Y_3 is assumed to be from a different survey with a different response mechanism.

The example is kept very simple and relationships are, of course, more complex in reality. For example, even in the *survey variable cause model* Y may depend on X. Moreover, many X-variables that affect ρ are usually not observed and/or not known or show missing values.

4.1 Multivariate nonresponse bias indicators

Let us consider two scenarios. In our first scenario, we have information on all five socio-demographic variables in our studies from the sample frame and use these information to conduct nonresponse bias analysis. In our second and more realistic scenario, some variables that affect nonresponse are not available from the frame and we only have a subset of the socio-demographic variables, gender, age and German citizenship, limiting our analysis to these variables.

To estimate multivariate nonresponse bias indicators, we run a logistic regression of the response indicator (yes/no) on all available frame information in each scenario. We estimate the R-indicator using the

predicted probabilities from the logistic regression and build nonresponse weights as the inverse of the predicted probabilities. Table 1 shows the R-indicator, McFadden's pseudo R^2 from the logistic regression and the variance of the nonresponse weights.

Table 1: Estimated multivariate nonresponse bias indicators for two scenarios characterized by different availability of auxiliary data.

indicators	full variable set			limited variable set		
	separate cause	common cause	survey variable cause	separate cause	common cause	survey variable cause
R-indicator	0.36	0.36	0.94	0.41	0.41	0.94
McFadden's R^2	0.35	0.35	0.00	0.29	0.29	0.00
$\sqrt{\text{variance of nr weights}}$	70.97	70.97	0.13	31.78	31.78	0.11

The measures in the first three columns make use of all variables that are part of the nonresponse process for the *separate cause model* and the *common cause model*. The limited variable set used in the second scenario does not include all relevant variables. None of the variables used in any scenario affect the nonresponse process for the *survey variable cause model*.

The indicators need to be interpreted with caution and we must be aware which conclusions can be drawn and which not. All these measures only indicate to what extent the distributions of the X-variables in the survey correspond to those in the population. For the full variable set scenario, all indicators show very low risk of nonrespnse bias in the survey variable cause model but high risk for the other two models. That is expected, as we know that the data was generated that way. We should, however, not naively conclude that we have a generally low risk of nonresponse bias in this survey. Knowledge of the relationship of X and Y, e.g., empirical evidence from other studies or theoretical considerations can help to judge whether to expect nonresponse bias in Y. If we, for example, know from other studies that the Y variable is highly correlated to the X- variables (like it is the case for Y_2), we would assume a high risk of nonresponse bias in this variable of interest but not for a variable that is not related to X, like Y_1 in the same survey. In our example, Y_3 is generated to not depend on the X-variables but to Y_3 itself. Thus, the indicators are not meaningful when it comes to the risk of nonresponse bias in Y_3. Likewise, there can always exist unknown or unobserved auxiliary variables that are related to Y and cause nonresponse bias. If these variables are not related to the X-variables that might capture parts of their effect, the high risk of nonresponse bias can not be detected. This illustrates the limitations of these indicators: they summarize the representation of different population subgroups (based on the X-variables) but do not allow to rule out effects of variables that are not part of the analysis, either because they can not be compared to some benchmark or are not observed at all.

Comparing the full variable set scenario to the limited variable set scenario, we find all indicators to "improve": the R-indicators are higher whereas McFadden's R^2 and the variation of the nonresponse weights are closer to zero. These findings show how the specification of the nonresponse model that is part of these three measures influences the results. Interestingly, the mis-specified models in the limited variable set scenario that excludes relevant auxiliary variables misleadingly indicates a lower risk of nonresponse bias. This makes sense as leaving out relevant predictors in the nonresponse model decreases the model fit and thus McFadden's R^2 and decreases the variation of the predicted values for the nonresponse propensity. Again, the indicators can only be interpreted in relation to the specific X-variables. We can never know for sure that there are no unobserved characteristics that are excluded from the estimation of the nonresponse bias indicator (like in the limited variable set scenario) that systematically affect survey nonresponse.

Both scenarios show that the *survey variable cause model* in our example exhibits higher representativeness (of the X-variables) than the other models. Finding that the indicators heavily depend on the model specification, it is important to note that the indicators should only be compared between surveys if they include the same set of X-variables. Also, they should only be interpreted with respect to the X-variables they make use of. In the full variable set we can draw conclusions about the surveys' representativeness with regard to gender, age, education, German citizenship and household size whereas in the limited variable set our conclusions are limited to gender, age and German citizenship.

Table 2 shows the correlation between nonresponse weights and Y and the FMI in estimating $\hat{\bar{y}}$. We show the results for the full information and limited information scenario described above. To estimate the FMI, we impute Y using all X-variable values that are available in the respective scenario. We conduct multiple imputations ($m = 10$) using predictive mean matching as implemented in the R package *mice* (Van Buuren & Groothuis-Oudshoorn, 2011). The results are shown in Table 2.

Table 2: Estimated multivariate nonresponse bias indicators for two scenarios characterized by different availability of auxiliary data.

indicators	full variable set			limited variable set		
	separate cause	common cause	survey variable cause	separate cause	common cause	survey variable cause
correlation of weights and y	0.04	-0.25	0.02	0.04	0.03	0.03
FMI	0.80	0.31	0.81	0.90	0.92	0.92

Comparing the *separate cause model* and *common cause model*, we find that the correlation of weights and Y correctly identifies a higher risk in nonresponse bias for Y in the common cause model for the full variable set scenario. It does, however, not detect the high nonresponse bias in Y in the *survey variable cause model*. This was expected as there is no relationship between the X-variables and Y for this model. Consequently, findings are very similar for the *survey variable cause model* and the *separate cause model* for which X and Y are also not associated. Whereas for the *separate cause model* we are right to conclude that there is only a low risk of nonresponse bias in Y we would be wrong in the *survey variable cause model*. For the limited variable set scenario, the correlations are estimated to be of about the same size for all of the three nonresponse models, failing to detect the high risk of nonresponse bias in the *common cause model*. Again, the nonresponse bias indicators can only be interpreted in relation to the X-variables they are built on and they are not able to detect a correlation of Y and ρ or of Y and unobserved X-variables that are not part of the propensity model.

The FMI can be interpreted by comparing its value to the nonresponse rate. As Nishimura et al. (2016) point out, the FMI is bounded by the nonresponse rate under the MAR model. Observing values that are much higher than the nonresponse rate (50 % in our case) is an indication that the imputation model is mis-specified. If the imputation model is correctly specified, the FMI should not be larger than about 0.5. The higher the correlation of Y and X, the stronger the FMI decreases towards zero. For the limited variable set scenario we find all FMIs to be larger than 0.5 indicating that the models are mis-specified and do not caputure all X-variables that relate to the response propensity. For the full variable set scenario, the FMI is larger than 0.5 for the *separate cause model* and *survey variable cause model*. This again can be explained by the fact that X is not correlated with Y and using X for the imputation for Y has no positive effect on the between-imputation-variance of Y. As noted above, both models do not lead to data that are missing at random and thus do not meet the assumptions of the FMI. We are not able to distinguish between the *separate cause* and *survey variable cause* model based on the FMIs. Whereas for the *survey variable cause* model we are right to conclude that we observe a high risk of nonresponse bias the large FMI in the *separate cause model* is misleading. The FMI for the *common cause model* reflects the situation

very well. It is lower than the nonresponse rate indicating correct model specification. Due to random error in the generation of Y, X and Y are not perfectly correlated and the FMI is not exactly zero.

The multivariate nonresponse bias indicators are useful measures for the representativeness of a survey regarding the specific variables they are built on. They do, however, not allow for an evaluation of which variables are well represented and which are not. Most importantly, taking them as indicators for the risk of nonresponse bias in Y might be very misleading. The representativeness of X can only be used as an indicator for nonresponse bias in Y under the assumption that X and Y are highly correlated on the population level. This assumption usually cannot be tested. Only substantive knowledge on the relationships between X and Y can help to evaluate whether the assumption holds.

Several indicators that allow for an evaluation of the risk of nonresponse bias on the variable level are illustrated in the following section.

4.2 Univariate nonresponse bias indicators

Let us assume we know the means and proportions for the socio-demographics (X-variables) from official statistics such as the Mikrozensus. We can easily compare the survey estimates based on the respondents and the population values from the Mikrozensus for the X-variables for the three nonresponse models. We should be aware, however, that even for random sampling of the individuals, the survey estimates likely do not exactly match the population parameters by chance. Like in real applications, we are not able to separate random sampling error from nonresponse error.

To be able to compare the magnitude of nonresponse bias between X-variables and different surveys, we compute the relative biases using benchmarks from official statistics and the AARB (see Table 3). For illustration, we assume that we also know the true distribution of Y which is usually unknow. This allows us to estimate relative nonresponse bias in Y as well. The relative biases in X are the same for the *separate cause model* and the *common cause model* but differ from the ones for the *survey variable cause model*.

We can, for example, see that the younger age cohorts are under- and the older age cohorts are over-represented for all surveys. The mis-representation is strongest for the individuals aged 16 to 29 years. The AARB gives a summary of the biases and naturally shows the same trend as the indicators from the previous section: the *survey variable cause model* shows the highest representativeness or, in other words, lowest AARB for the socio-demographic variables under study.

In this example, we are able to evaluate bias in the Y-variables as well. The *separate cause model* and *common cause model* examples are generated to be the same survey with different Y-variables of interest. We can see that the Y-variable in the *separate cause model* example is under-estimated to a very small degree due to random sampling error whereas the Y-variable from the *common cause model* shows an over-estimation of more than 5%. This shows that variables from the same survey can be affected by nonresponse very differently depending on how they are related to the nonresponse mechanism. In our example, nonresponse bias is highest for the Y-variable from the *survey variable cause model*. Even though the survey was very representative for all X-variables, Y suffers from severe nonresponse bias. This example shows that a good/bad representation of X does not necessarily imply a good/bad representation of Y. This example made use of the population information of Y which is usually not available.

Table 3: Relative nonresponse bias (in percent) in estimated proportions of auxiliary variables and average absolute relative bias in the survey samples underlying different response mechanisms.

variable	separate cause	common cause	survey variable cause
age 16-29	-68.57	-68.57	9.05
age 30-39	-4.71	-4.71	13.53
age 40-49	-7.06	-7.06	-0.59
age 50-59	15.91	15.91	-14.09
age 60+	56.09	56.09	-4.35
female	-3.67	-3.67	-4.29
low education	-29.35	-29.35	-0.65
medium education	-5.59	-5.59	-0.59
high education	31.43	31.43	1.14
household size 1	6.36	6.36	12.27
household size 2	3.68	3.68	3.16
household size 3	-1.89	-1.89	-5.95
household size 4	-6.36	-6.36	-5.00
German citizenship	12.64	12.64	-1.72
Y	-2.26	5.70	18.91
AARB	18.09	18.09	5.45

5. Conclusion

Declining participation rates raise concerns of high nonresponse bias. However, as discussed in the previous sections, it is not so much the general willingness to participate that affects the risk of nonresponse bias, but rather how different this willingness is for different population groups. Nonresponse bias arises whenever respondents differ from nonrespondents in the characteristic of interest. Within the same survey, statistics for some variables may be completely accurate and others may be heavily biased. If the nonresponse mechanism depends on the variable of interest, estimates for this variable will be biased no matter how well the other variables are represented.

Since usually nonresponse bias in the variables of interest cannot be measured directly, measures based on auxiliary variables such as socio-demographic characteristics are used to estimate the risk of nonresponse bias. In order to draw conclusions to the variable of interest it is necessary to use auxiliary variables that are correlated to them. In addition to frame information, the collection of interviewer observations and paradata has shown to be useful to study nonresponse bias and to apply nonresponse adjustments (see for example Krueger & West, 2014).

Indicators that in addition to the auxiliary variables incorporate the variables of interest can be very helpful if the data follow the *common cause model*. In case of the *survey variable cause model* or if the nonresponse mechanism is mis-specified, however, they tend to be misleading.

All of the above indicators can provide information about the risk of nonresponse bias, but they have their limitations. Their interpretation should always be guided by considerations of contextual relationships. To get the most accurate picture of nonresponse bias, we highly recommend to use several indicators. The estimation and visualization of many of the proposed nonresponse bias indicators are implemented in the easy to use R package *sampcompR* (Rohr, 2023).

We propose to start with estimating the nonresponse model whenever auxiliary information is available

for respondents and nonrespondents. Modelling nonresponse requires a clearly defined sample and a response indicator that can be assigned to each unit of the sample. The model coefficients provide information on which characteristics influence participation and in which direction. Estimated response propensities can than be used to estimate several multivariate nonresponse bias indicators.

We propose to compare survey estimates to benchmarks from official statistics on all available characteristics. This is especially useful for all characteristics that are available for respondents only and are thus not available for the nonresponse models. Note that the differences between survey estimates and population benchmarks can only be attributed to nonresponse if the sample is drawn randomly from the population (and design weights are used if applicable) and the measurement is assumed to be error-free. Even if these assumptions are fulfilled, the survey estimate will randomly deviate from the population benchmark due to sampling.

References

Andridge, R. R., & Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, *27*(2), 153.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, *4*(3), 251–260.

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

Bruch, C. (2023). Imputation of missing values in survey data. *Mannheim, GESIS – Leibniz-Institut Für Sozialwissenschaften (GESIS Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_en_044

Cornesse, C., Felderer, B., Fikel, M., Krieger, U., & Blom, A. G. (2021). Recruiting a probability-based online panel via postal mail: Experimental evidence. *Social Science Computer Review*.

Eckman, S., Unangst, J., Dever, J. A., & Antoun, C. (2022). The Precision of Estimates of Nonresponse Bias in Means. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smac019

Felderer, B., Kirchner, A., & Kreuter, F. (2019). The effect of survey mode on data quality: Disentangling nonresponse and measurement error bias. *Journal of Official Statistics*, *35*(1), 93–115.

Felderer, B., Kueck, J., & Spindler, M. (2023). Using double machine learning to understand nonresponse in the recruitment of a mixed-mode online panel. *Social Science Computer Review*, *41*(2), 461–481.

Friedel, S., Felderer, B., Krieger, U., Cornesse, C., & Blom, A. G. (2023). The early bird catches the worm! Setting a deadline for online panel recruitment incentives. *Social Science Computer Review*, *41*(2), 370–389.

Gabler, S., Kolb, J.-P., Sand, M., & Zins, S. (2015). Gewichtung. *Mannheim, GESIS – Leibniz-Institut Für Sozialwissenschaften (GESIS Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_007

Green, K. E. (1991). Reluctant respondents: Differences between early, late, and nonresponders to a mail survey. *The Journal of Experimental Education*, *59*(3), 268–276.

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

Groves, R. M., Brick, J. M., Couper, M., Kalsbeek, W., Harris-Kojetin, B., Kreuter, F., … Wagner, J. (2008). Issues facing the field: Alternative practical measures of representativeness of survey respondent pools. *Survey Practice*, *1*(3), 2910.

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

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

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

Krueger, B. S., & West, B. T. (2014). Assessing the potential of paradata and other auxiliary data for nonresponse adjustments. *Public Opinion Quarterly*, *78*(4), 795–831.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique*, 139–157.

Lugtig, P., Roth, K., Schouten, B., et al. (2022). Nonresponse analysis in a longitudinal smartphone-based travel study. *Survey Research Methods*, *16*(1), 13–27.

Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *International Statistical Review*, *84*(1), 43–62.

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.

Roberts, C., Vandenplas, C., & Herzing, J. (2020). Validation of r-indicators as a measure of the risk of bias

using data from a non-response follow-up survey. *Journal of Official Statistics*, *36*(3), 675–701.
Rohr, B. (2023). *SampcompR: Comparing and visualizing differences between surveys. R package version 0.1.0.0*. Retrieved from https://github.com/BjoernRohr/sampcompR
Rohr, B., Silber, H., & Felderer, B. (2023). Comparing the accuracy of univariate, bivariate, and multivariate estimates across probability and non-probability surveys with population benchmarks. *SocArXiv. March 4. Doi:10.31235/Osf.io/N6ehf*.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys. hoboken*. New York: John Wiley & Sons.
Sand, M., & Kunz, T. (2020). Gewichtung in der Praxis. *Mannheim, GESIS – Leibniz-Institut Für Sozialwissenschaften (GESIS Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_030
Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101–113.
Schouten, B., Shlomo, N., & Skinner, C. (2010). *Indicators for monitoring and improving representativeness of response*.
Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03
Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly*, *74*(2), 223–243.
Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, *76*(3), 555–575.