Survey Research Methods (2018)
Vol. 12, No. 1, pp. 1-13
doi:10.18148/srm/2018.v12i1.7205

ISSN 1864-3361
http://www.surveymethods.org

Is there an association between survey characteristics and representativeness? A meta-analysis

Carina Cornesse
Political Economy of Reforms (SFB 884)
University of Mannheim, Germany
and
GESIS – Leibniz Institute for the Social Sciences
Mannheim, Germany

Michael Bosnjak
ZPID – Leibniz-Institute for Psychology Information
Trier, Germany
and
University of Trier, Germany

How to achieve survey representativeness is a controversially debated issue in the field of survey methodology. Common questions include whether probability-based samples produce more representative data than nonprobability samples, whether the response rate determines the overall degree of survey representativeness, and which survey modes are effective in generating highly representative data. This meta-analysis contributes to this debate by synthesizing and analyzing the literature on two common measures of survey representativeness (R-Indicators and descriptive benchmark comparisons). Our findings indicate that probability-based samples (compared to nonprobability samples), mixed-mode surveys (compared to single-mode surveys), and other-than-Web modes (compared to Web surveys) are more representative, respectively. In addition, we find that there is a positive association between representativeness and the response rate. Furthermore, we identify significant gaps in the research literature that we hope might encourage further research in this area.

Keywords: Meta-analysis, survey representativeness, R-indicator, descriptive benchmark comparisons, response rate, nonprobability sampling, mixed mode, web surveys, auxiliary data.

1 Background and Aims

One of the most important questions in the research field of survey methodology is how to collect high quality survey data that can be used to draw inferences to a broader population. Extensive research has been conducted from different angles, often creating an ambiguous picture of whether and how a specific survey characteristic might help or hurt in the pursuit of reaching high survey quality. An example is the current debate around the representativeness of probability-based surveys versus nonprobability surveys, where proponents and opponents of each approach provide new findings on a regular basis. Some of these studies suggest that probability-based surveys are more accurate than nonprobability surveys (e.g. Chang & Krosnick, 2009; Loosveldt & Sonck, 2008; Malhotra & Krosnick, 2007; Yeager et al., 2011). Other studies demonstrate that nonprobability surveys are as accurate as or even more accurate than the probability-based surveys (e.g. Gelman, Goel, Rothschild, & Wang, 2017; Wang, Rothschild, Goel, & Gelman, 2015).

Another example of a scientific discussion in the survey methodological community is the question of whether the response rate can be used as a representativeness indicator. Even though meta-analytic research shows that the response rate is only weakly associated with nonresponse bias (Groves et al., 2008; Groves & Peytcheva, 2008) the debate about how to stop the decrease in response rates goes on (e.g. Brick & Williams, 2013; de Leeuw & de Heer, 2002). Another topic continuously investigated in survey methodological research with seemingly contradictory findings is which survey mode and which mode combinations yield the most representative data (e.g. Luiten & Schouten, 2012)). Based on a systematic review summarizing the available meta-analytic evidence about mode effects and mixed-mode effects on representativeness, Bosnjak (2017, p. 22) concludes: "there is a lack of meta-analytic studies on the impact of (mixing) modes on estimates of biases in terms of measurement and representation, not just on proxy variables or partial elements of bias, such as response rates."

In light of the extensive research on different aspects of survey representativeness that partly leads to seemingly contradictory findings, the overall aims of this paper are to identify and systematically synthesize the existing literature on survey representativeness, and to answer the question if, and

Contact information: Carina Cornesse, SFB 884 "Political Economy of Reforms", University of Mannheim and GESIS – Leibniz Institute for the Social Sciences (Email: carina.cornesse@uni-mannheim.de)

to what extent, specific survey characteristics are associated with survey representativeness. Before we develop the specific research question and hypotheses of this study, we define the scope by specifying the two survey representativeness concepts considered.

2 Survey Representativeness Concepts

Survey representativeness is an ambiguous and controversial term. In survey methodology, it most commonly refers to the success of survey estimates to mirror "true" parameters of a target population (see Kruskal and Mosteller, 1979a, 1979b, 1979c for this and other definitions of the concept of representativeness). The fact that a survey can be representative regarding the variables of interest to one researcher, and misrepresent the variables of interest to another researcher at the same time, adds to the vagueness of the concept of a generally representative survey. Since some researchers reject the term "representativeness" entirely (e.g. Rendtel & Pötter, 1992; Schnell, 1993), research on the subject frequently emerges under the keywords "accuracy" or "data quality" in general (e.g. Malhotra & Krosnick, 2007; Yeager et al., 2011). Other researchers still use the term "representativeness" (e.g. Chang & Krosnick, 2009) or "representativity" (e.g. Bethlehem, Cobben, & Schouten, 2011). From a total survey error perspective, survey non-representativeness encompasses the errors of nonobservation: coverage error, sampling error, nonresponse error, and adjustment error (Groves et al., 2004; Groves & Lyberg, 2010).

Because of its ambiguity, there are various operationalizations of survey representativeness. Two common approaches prevail; each is associated with specific advantages and disadvantages. The first approach is to compare the survey response set to the target population in question. This is mostly done by using benchmark comparisons, estimating the absolute bias. For this type of representativeness measure, it is necessary to assume that the benchmark to which the response set is compared is unbiased and reflects the "true" values of the target population. Therefore, these representativeness assessments are mostly conducted on sociodemographic characteristics only (e.g. Keeter, Miller, Kohut, Groves, & Presser, 2000; Wozniak, 2016). The goal is in many cases to show whether a specific data source is trustworthy and can therefore be used to answer a substantive research question, for example on health (Potthoff, Heinemann, & Güther, 2004) road safety (Goldenbeld & de Craen, 2013) or religion and ethnicity (Emerson, Sikkink, & James, 2010).

The other approach to operationalizing the concept of representativeness is to compare the survey response set to the gross sample including respondents as well as nonrespondents. This allows the inclusion of a much broader set of variables in the assessment compared to the absolute bias measure. For instance, data from the sampling frame, fieldwork, and from data linkage can be used for the representativeness assessment. Examples of such measures include R-Indicators (Schouten, Cobben, & Bethlehem, 2009), balance and distance measures (Lundquist & Särndal, 2012), and Fractions of Missing Information (Wagner, 2010). However, since only those characteristics can be assessed that are available for respondents as well as for nonrespondents, data availability is an issue also for measures operationalizing this approach. In addition, the underlying assumption is that if the response set perfectly reflects the gross sample, including both respondents and nonrespondents, this constitutes a representative survey. This assumption leaves out the fact that the sample might already be biased compared to the target population due to coverage error and sampling error. Therefore, these measures operationalizing the second approach to representativeness are – technically speaking – nonresponse bias measures (Wagner, 2012).

Because there is no agreed-upon measure of survey representativeness, we focus on two conceptualizations that are common in the survey methodological literature. First, we investigate R-Indicators that assess representativeness by comparing respondents to the gross sample of a survey, which includes respondents as well as nonrespondents. Second, we examine descriptive benchmark comparisons that measure survey representativeness by comparing respondent characteristics to an external benchmark that is supposed to reflect the characteristics of the target population.

3 Conceptual development of research question and expectations

In this paper, we assess whether there is an association between survey characteristics and the reported degree of representativeness of a survey. We focus on a selection of survey characteristics that are commonly considered to affect survey representativeness in the current research literature and for which we could gather sufficient information during the data extraction process. Specifically, we expect that the following five survey characteristics are associated with the reported degree of representativeness: probability surveys versus nonprobability surveys, response rates, mixed-mode surveys versus single-mode surveys, web surveys versus single-mode surveys, and the number of auxiliary variables (i.e., data that is available for respondents as well as nonrespondents) used in the representativeness assessments. In the following, we describe our expectations for each survey characteristic considered in more detail. These associations should hold regardless of whether representativeness is operationalized using R-Indicators (an indicator based on response propensity models using auxiliary data; see section 4.3.1 in this paper for more details) or descriptive benchmark comparisons.

3.1 Probability surveys versus nonprobability surveys

Probability sampling theory states that valid inference and the assessment of an estimate's uncertainty can only be guaranteed by random selection of units of analysis from an accurate sampling frame (e.g. Kish, 1965; Lohr, 2010). Some researchers have, however, suggested that true probability sample surveys of the population do not exist in practice due to non-random survey nonresponse (e.g. Gelman et al., 2017).

Over the last decades, the number of nonprobability surveys, especially commercial opt-in online panels, has increased immensely. Their advantages include that they are fast and relatively cheap. Furthermore, they often reach high numbers of respondents (e.g. Bethlehem & Biffignandi, 2012). They also often have respondent pools containing millions of people, sometimes even across several countries, as for instance Google 360[1], Research Now[2], Omnirussia[3], or Toluna[4]. In many cases, little information is available on how these nonprobability panels recruit people, how many people are in their respondent pool, how they sample from the respondent pool to conduct an individual survey, and how many people in their respondent pool are actually active, i.e. respond to survey requests regularly. Furthermore, it has been observed that active panelists often participate in multiple non-probability panels (e.g. Tourangeau, Conrad, & Couper, 2013; Vonk, van Ossenbruggen, & Willems, 2006), especially since there are databases that contain many opt-in panels from which people who would like to earn money by completing surveys can choose as many panels as they like (e.g., www.umfragenplatz.de; https://www.mysurvey.com/www.mysurvey.com/www.mysurvey.com/). In addition, a Google search performed in July 2017 for the term "money for surveys"' yielded 180.000 hits and the application Google Opinion Rewards, that offers Google Play credit in return for answering surveys, has been downloaded 10 million times on Android devices across the world (according to the information available in the Google Play Store[5]). People who actively participate in several panels are, however, suspected of satisficing and even forging data to increase their financial rewards (e.g. Toepoel, Das, & van Soest, 2008; Yan & Tourangeau, 2008).

There is evidence that nonprobability survey respondents are a highly selective subgroup of the general population (e.g. Yeager et al., 2011). This is because people who recruit themselves by reacting to survey advertisements, which are a common way of nonprobability survey recruitment, are likely to attract people with specific profiles in terms of demographics, personality, values, and habits. For instance, people who frequently use the Internet are more likely to be exposed to commercials and advertisements (e.g. OECD, 2014). In addition, previous research indicates that nonprobability survey participants show higher political knowledge and engagement (e.g. Chang & Krosnick, 2009; Duffy, Smith, Terhanian, & Bremer, 2005) and that they are unlikely to be older than 65 (e.g. Loosveldt & Sonck, 2008). We expect the self-recruitment procedure to negatively affect the representativeness of nonprobability survey data, even though there is some contradicting evidence (e.g. Gelman et al., 2017; Wang et al., 2015). Still, our first hypothesis is:

H1: Probability surveys are more representative than nonprobability surveys.

3.2 Response rates

Declining response rates are a widely-discussed issue in the survey methodological literature. In the German general social survey ALLBUS, for instance, response rates have been decreasing from 54% in 1994 to 38% in 2012 (Blohm & Koch, 2015). Similarly, Curtin, Presser, and Singer (2005) find a response rate decline of about one percentage point per year in the University of Michigan Survey of Consumer Attitudes. These findings are in accordance with earlier research on declining response rates (e.g., the international comparative study on household survey nonresponse by de Leeuw and de Heer, 2002). If less people participate in a survey, the risk of nonresponse bias increases (Bethlehem et al., 2011). This can be deduced from the well-known expression for nonresponse bias:

$$\text{Bias}(\bar{y}_r) = \frac{M}{N}\left(\bar{Y}_r - \bar{Y}_m\right) \tag{1}$$

"where $\text{Bias}(\bar{y}_r)$ = the nonresponse bias of the unadjusted respondent mean; $\bar{y}_r$ = the unadjusted mean of the respondents in a sample of the target population; $\bar{Y}_r$= the mean of the respondents in the target population; $\bar{Y}_m$ = the mean of the nonrespondents in the target population; M = the number of nonrespondents in the target population; and N = the total number in the target population"(Groves, 2006: 648).

A difficulty with regard to response rates is that they are not always computed in the same way; for an overview, see e.g., American Association for Public Opinion Research (2016). In the scientific literature, it often remains unclear how exactly the response rates were calculated. This is especially true for the nonprobability surveys that usually provide participation rates where the number of the survey respondents is divided by the number of the invited members of

[1] https://www.google.com/insights/consumersurveys/homewww.google.com/insights/consumersurveys/home

[2] https://www.researchnow.com/audiences-data-collection/www.researchnow.com/audiences-data-collection/

[3] http://www.omirussia.ru/en/online_panels/consumer_panels/#consumerwww.omirussia.ru/en/online_panels/consumer_panels/#consumer

[4] www.toluna-group.com/de/umfrage-plattform/toluna-samplexpress-

[5] play.google.com/store/apps/details?id=com.google.android.apps.paidtasks

the respondent pool instead of dividing the number of survey respondents by the number of persons in a sample, which is more common in the computation of probability surveys. In the literature, these participation rates are often termed response rates although the AAPOR Task Force on Non-Probability Sampling, in an attempt to avoid confusion, recommends not calling these participation rates response rates (Baker et al., 2013).

Response rates are often reported as an indicator of nonresponse bias. It is, however, not necessarily true that high response rates lead to high degrees of survey representativeness. Indeed, the association between the response rate and nonresponse bias has been found to be small (Groves & Peytcheva, 2008). While it is true that the response rate places an upper bound on the potential nonresponse bias (Bethlehem et al., 2011), surveys with high response rates can still be heavily biased if the small number of nonrespondents systematically differs from the respondents. Conversely, surveys with relatively low response rates can accurately reflect population properties if the set of respondents only randomly varies compared to the set of nonrespondents.

Because the response rate places an upper bound on the potential nonresponse bias and because the response rate is often considered to be an indicator of the general quality of the survey data, we expect that surveys with high response rates more often achieve high degrees of representativeness. Therefore, our second hypothesis is:

H2: There is a positive association between the response rate and the measured degree of representativeness.

3.3 Mixed-mode surveys versus single-mode surveys

Mixed-mode surveys apply two or more modes of data collection (e.g Dillman, 2008). Reasons for using multiple modes are diverse. One reason is that offline survey costs can be decreased by adding a web survey version (e.g Couper, Kapteyn, Schonlau, & Winter, 2007; Jäckle, Lynn, & & Burton, 2013). Another reason is that with different modes it is sometimes possible to reach a more diverse set of people and thereby survey representativeness increases (e.g Klausch, Hox, & Schouten, 2015; Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011).

Mixed-mode survey designs vary in their purpose and type of implementation. Some surveys offer all respondents an option to choose their favorite mode from the onset (e.g Bosnjak, 2017). Other surveys assign modes to respondents based on estimated response propensities (e.g Schouten & Cobben, 2007). Even more flexible mixed-mode designs are called "responsive design surveys" (e.g Groves & Heeringa, 2006; Peytchev, Conrad, Couper, & Tourangeau, 2010). In responsive design studies, the survey mode assigned to a sample unit can change over the course of the survey fieldwork period depending on the development of subgroup response propensities. This approach has the advantage that modes can be targeted at sample subgroups that have a high risk of misrepresentation in the response set given one initial mode of data collection (e.g. singles in urban areas based on face-to-face survey data collection). These responsive techniques have become possible by the technological development in real-time fieldwork management (e.g Blom, 2016; Macer, 2014).

Regardless of the specific design, mixed-mode surveys have in common that they strive to improve survey data quality as compared to traditional single-mode surveys. This goal might not always be achieved, especially when survey requests to an offline mode offer a concurrent web option (e.g Medway & Fulton, 2012). Nevertheless, our third hypothesis is:

H3: Mixed-mode surveys are more representative than single-mode surveys.

3.4 Web surveys versus other single-mode surveys

Web surveys are a data collection mode that has become popular in survey research and practice in the early 2000s (e.g Couper, 2000; Couper & Bosnjak, 2010). Since then, the number of web surveys has increased dramatically reaching turnovers of around 15 billion US$ in 2014 in both Europe and North America[6]. Generally, web surveys are a valuable addition to the possibilities of designing a survey. They are fast, cheap, and easy to conduct. In addition, they have the advantage that respondents can fill out questionnaires whenever and wherever they like by using mobile devices (e.g Couper, 2013; Toepoel & Ludtig, 2015).

The rise of web surveys has, however, from the onset been accompanied by concerns about data quality, especially with regard to potential coverage error (e.g. Eckman, 2015; Blom et al., 2016). One reason for this is that web surveys are difficult to combine with probability sampling approaches, because often there is no sampling frame of Internet users available from which a sample could be drawn (e.g. Couper, 2000). Therefore, people who do not have an Internet connection and devices that enable logging into the Internet do not have the chance to be selected. In addition, people without access to the Internet have no chance to self-select into nonprobability web surveys. Research shows, however, that Internet users are systematically different from non-Internet users. For example, people with a low education level are usually underrepresented among Internet users (e.g., Blom et al., 2016).

Coverage error is more likely to be a problem in countries with relatively low Internet penetration rates than in countries with high Internet penetration rates (e.g OECD, 2014). Bosnjak et al. (2013) find, however, that coverage error is the

[6] www.esomar.org/news-and-multimedia/news.php?pages=1&idnews=150

most influential source of attenuating representativeness in web surveys. Therefore, our fourth hypothesis is:

H4: Web surveys are less representative than other single-mode surveys.

3.5 Auxiliary variables

The term "auxiliary data" usually encompasses all data that are available for both respondents and nonrespondents and can therefore be used to enhance post-survey adjustments (e.g Kreuter, 2013). Examples of this type of data include sampling frame data, survey paradata, and data linked to survey data from external sources such as population registers. These data are often, but not always, available on an aggregate level only, such as municipalities, city districts, or streets. Auxiliary data are commonly used to assess survey representativeness and adjust for misrepresentation (e.g Brick & Kalton, 1996; Kreuter & Olson, 2011).

Some studies on survey representativeness only include basic sampling frame information or socio-demographic variables in the representativeness assessment because it is often difficult to obtain more and other data for the assessment. Since these few basic variables are also commonly used to design quota samples or to construct nonresponse weights, many surveys cover these characteristics sufficiently. However, the representativeness measure is generally more informative the broader the range of auxiliary variables taken into account actually is. The more auxiliary variables a study aims to address, the larger is the risk that at least one auxiliary variable is misrepresented in the data. For these reasons, our last hypothesis is:

H5: The more auxiliary variables are used for the representativeness assessment the lower is the overall representativeness.

4 Method

4.1 Literature search, study identification and data extraction

In order to answer our research question whether and how survey characteristics are related to survey representativeness, we have identified, reviewed, and coded the existing literature. In our dataset, we included all journal articles, book chapters, and scientific working papers that contain one of the two measures of representativeness that we focus on (general sample-based R-Indicators or descriptive comparisons between a survey and an external benchmark). Additional necessary preconditions for articles to be included in our analysis are that they need to be published in English, available in full text, and listed in an established database.

Our literature search and data coding were conducted in multiple steps by multiple coders. We used several data bases (Web of Science, EBSCO host, Jstor, and Google Scholar) and conducted full text searches wherever possible. Furthermore, we used the publications section of the website www.risq-project.eu as a database for articles on R-Indicators. Generally, we used a large number of search terms ("representativeness", "representation", "survey research", "nonresponse", and "R-Indicator") in different combinations. We also conducted a snowballing search where we started with crucial articles and searched the literature section for relevant sources. For an overview of the article identification and eligibility assessment processes see Figure 1.

As for the coding procedure, every data entry was checked by another coder and disagreements between coders were resolved by the authors of this paper. When the coded papers lacked necessary information (e.g., the survey mode), this information was searched on the Internet, requested from the authors of the papers in question via email, or, if possible, computed from the existing information.

4.2 Meta-analytic procedure

To compute the two overall mean representativeness scores across all eligible studies identified, we used Hedges/Olkin-type random-effects meta-analytic models (e.g Hedges & Olkin, 1985; Raudenbush, 2009). In Hedges/Olkin-type meta-analyses, observed effect sizes (i.e., representativeness estimates in our case) are synthesized using a weighted mean procedure, with the inverse sampling variance of each effect size (i.e., representativeness coefficient) serving as weights. This procedure ensures that more precise representativeness estimates, that is, those being associated with smaller study-specific sampling variances, are assigned a larger weight when computing the overall mean representativeness estimate across all studies considered compared to the less precise effect sizes. Hedges/Olkin-type meta-analyses are based on one of two statistical models. A fixed-effect meta-analysis assumes all primary studies are estimating the same mean (representativeness) value, with the only source of variability being study-level sampling error. A random-effects meta-analysis allows for unsystematic differences in the mean (representativeness) values from study to study, in addition to study-level sampling error. The selection of a model must be based on the question of which model fits the inferential goal, with random effects models being favoured for unconditional inferences, that is inferences going beyond the specific set of characteristics of the observed studies (Hedges & Vevea, 1998). Technically, in random effects meta-analyses, two sources of variability around the mean representativeness effect are estimated, namely unsystematic between-study variance ('T-square') in true effects, and within-study sampling error. As a next step, the homogeneity of the overall weighted representativeness means is estimated, answering the question if, and to what extent, the variability between observed repre-

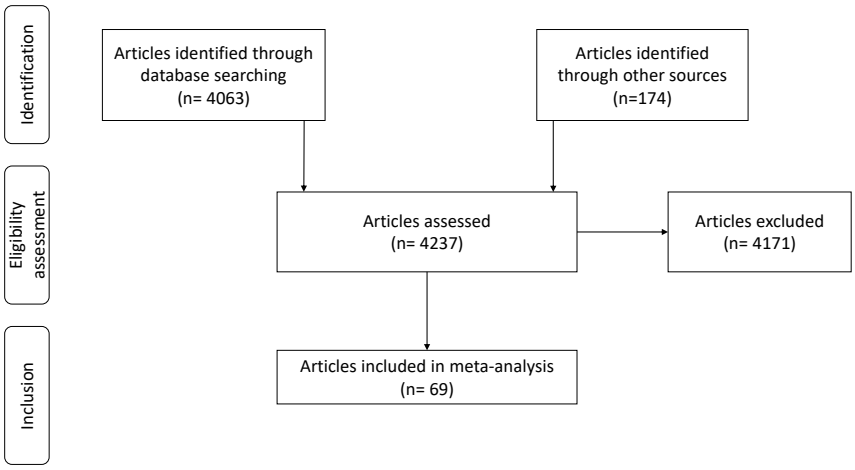

Figure 1. Article identification and eligibility assessment

sentativeness measures can be explained by sampling error and T-square alone, and/or by systematic differences among effect sizes (so-called moderators). We report the Q statistic of the model that indicates heterogeneity if significant. In case of heterogeneity, which we assume in light of our hypotheses, we perform moderator analyses aiming to explain the variability among representativeness estimates using mixed-effects models that combine random-effects meta-anaylsis with data on the moderator. For an in-depth treatment of meta-analytic procedures, we recommend Borenstein, Hedges, Higgins, and Rothstein (2009), Card (2012).

We use the rma function implemented in the R package *metafor*, version 2.0 (Viechtbauer, 2010). The rma function provides a general framework for fitting various meta-analytic models that are typically used in practice. Furthermore, we conduct moderator analyses using mixed-effects meta-regressions (van Houwelingen, Arends, & Stijnen, 2002) implemented in *metafor*. The R-script and analysis output are available as in the Supplementary Material of this paper.

4.3 Effect size measures

In meta-analytic terminology, dependent variables are called effect sizes and independent variables are called moderators. In the following, we describe the two effects sizes we use in this paper: R-Indicators and the Median Absolute Bias (MAB) derived from the descriptive benchmark comparisons. These measures are common in the existing literature and examine survey representativeness from different perspectives.

R-Indicators. General sample-based R-Indicators are a measure of survey representativeness that is based on logistic regression models of the propensity to respond to a survey (Schouten et al., 2009). In practice, these response propensities are unknown and therefore have to be estimated, which is usually done using a logistic regression model. The independent variables in the regression models are auxiliary variables that need to be available for both respondents and non-respondents to the survey. The individual response propensities from the regression model are aggregated using the formula

$$S(\rho) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\rho_i - \bar{\rho})^2} \tag{2}$$

where n is the number of cases, i is the case indicator, ρ is the propensity to respond, and $\bar{\rho}$ is the mean response propensity.

The aggregated results from the propensity model are rescaled to range between zero (not representative) and one (very representative) using the formula

$$R(\rho) = 1 - 2S(\rho) \tag{3}$$

where ρ is the standard deviation of the response propensities. For our meta-regressions, we compute inverse variance weights for the R-Indicators based on their confidence intervals that we find in the publications. In the primary studies, they are usually computed from standard errors using bootstrapping procedures. Where confidence intervals for the R-Indicators are not reported we impute them using a predictive mean matching procedure (e.g Schenker & Taylor, 1996).

MAB. We compute the Absolute Bias by category (AB_c) as the percentage point differences between proportions of a characteristic in a survey and an external gold-standard benchmark for each variable of a study with

$$\mathrm{AB}_c = \left| \left(\frac{n_{R_c}}{n_R}\right) - \left(\frac{N_C}{N}\right) \right| \tag{4}$$

where n_R is the number of respondents, N is the benchmark estimate, and C is the characteristic.

The raw estimates of the survey and benchmark for our computations are usually reported in tables in the publications. To aggregate the results, we compute the Median Absolute Bias MAB across all of the characteristics assessed in a study of these individual percentage point differences across all categories and variables in the study. We compute inverse variance weights for the MABs based on bootstrapped standard errors (e.g Efron & Tibshirani, 1986).[7]

5 Results

In this section, we summarize the meta-analytic findings. We start with some general results from random-effects meta-analyses, yielding a summary effect across all studies included for the two effect size measures considered, namely R-Indicators and MAB. Next, we give an overview of the moderator analyses conducted to test our five hypotheses, relating the following design features to representativeness: probability versus nonprobability surveys, response rates, mixed-mode surveys versus single-mode surveys, and the number of auxiliary variables. Finally, we present more detailed findings on each of our moderators in turn. In each of the analyses, significant outliers and missing values were excluded (see the Supplementary Material for detailed information).

5.1 General findings

Generally, our random-effects models show for each of the two effect sizes (R-Indicators and MAB) that most surveys in our sample achieve high degrees of representativeness. In addition, we find that there is a substantial amount of heterogeneity in the data worth to be explored further.

Table 1 displays the overall mean effect size, standard error, a heterogeneity estimate (Q-test results), and the number of studies (k) included in our analysis of the random-effects models on R-Indicators and the MAB. The mean effect sizes are 0.84 for the R-Indicators and 4.39% for the MAB. The confidence intervals (CI) around these mean effect sizes are from 0.82 to 0.85 for the R-Indicators and from 3.73% to 5.05% for the MAB. The Q-tests are highly significant for both effect sizes, indicating substantial heterogeneity in the data worth exploring with the aid of moderator analyses.

5.2 Moderator analysis

Overall, our moderator analyses show that there are significant associations between the degree of representativeness as measured using R-Indicators or the MAB and the survey design characteristics specified in our hypotheses. In Table 2, we display an overview of the findings. The table is structured as follows: The first column contains the names of the moderators and the categories of the dichotomous variables.

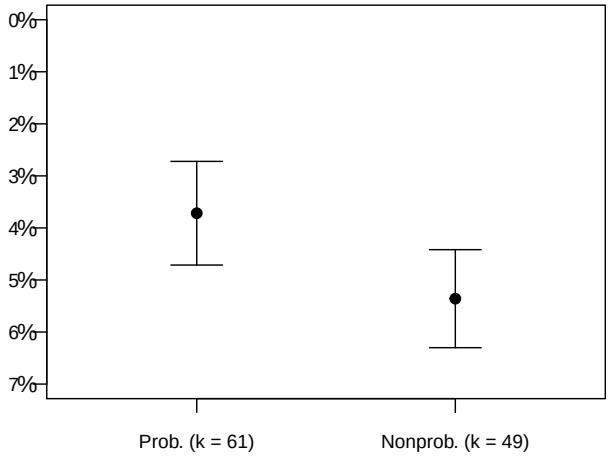

Figure 2. MAB subgroup comparison results by probability versus nonprobability surveys as a moderator

The rest of the table is divided into an overview of results from the moderator analyses on the R-Indicators and those of the MAB. For each of the effect sizes, we present the number of studies included in the moderator analyses (k). If the moderator is a dichotomous variable, we display the number of studies included for each of the moderator's categories separately. Furthermore, we present the regression coefficients, their significance level, and their standard errors of the mixed-effects meta-regressions of the effect sizes on the individual moderators as well as R^2-statistics as a measure of model fit.

Generally, we find in the mixed-effects meta-regression models that the degree of representativeness as measured using R-Indicators is associated significantly positively with the response rate, with mixed-mode surveys (as opposed to single-mode surveys), and the number of auxiliary variables that is used to compute the R-Indicators. In addition, the MAB is significantly negatively associated with probability surveys (as opposed to nonprobability surveys), the response rate, and other single-mode surveys (as opposed to web surveys).

Probability versus nonprobability surveys. The first moderator we examine is probability surveys versus nonprobability surveys. Overall, we find evidence in support of our expectation that probability surveys are more representative than nonprobability surveys (H1), although we can only identify this association using the MAB as a representativeness measure.

The computation of sample-based R-Indicators requires that a survey is based on a probability sample. In nonprobability samples, we are unable to identify the nonrespon-

[7] See the Supplementary Material for a sensitivity analysis using the Mean Absolute Bias instead of the Median Absolute Bias.

Table 1
Random-effects meta-regression models of R-Indicators and MAB

	R-Indicators	MAB
Mean	0.84	4.39%
CI	0.82–0.85	3.73%–5.05%
Q-test	20914.43***	8710.95***
k	109	108

* $p < 0.01$ ** $p < 0.001$ *** $p < 0.0001$

Table 2
Mixed-effects meta-regression models of R-Indicators on each moderator (standard errors in parentheses)

	R-Indicator			MAB		
Moderator	*k*	Coef.	Std. Err.	*k*	Coef.	Std. Err.
Sample						
Prob.	110	-	-	61	-2.18***	0.59
Nonprob.	0			49		
R^2		-			0.11	
Response rate (×100)	104	0.14**	0.04	90	-2.05*	1.14
R^2		0.08			0.02	
Mixed Mode						
Mixed	45	0.04***	0.01	8	-	-
Single	51			101		
R^2		0.13			-	
Web survey						
Web	1	-	-	56	1.57*	0.65
Other	50			45		
R^2		-			0.04	
No. auxiliary variables	104	0.05	0.00	104	-10.28*	0.00
R^2		0.09			0.04	

* $p < 0.01$ ** $p < 0.001$ *** $p < 0.0001$

dents, and we also do not have auxiliary data on the nonrespondents. A workaround might be the usage of population-based R-Indicators, which apply population aggregated auxiliary information, for example from benchmark data, instead of individual-level auxiliary data (e.g Shlomo et al., 2009). These population-based R-Indicators are not yet widely used, especially not with regard to nonprobability surveys. We therefore cannot consider them in our analysis. Therefore, we do not have any nonprobability surveys in the R-Indicator analysis that we could compare the probability surveys to.

Regarding the MAB, we find that there are 61 probability surveys and 49 nonprobability surveys in our dataset. The right-hand side of Figure 2 shows an average MAB of 3.72% with a confidence interval from 2.72% to 4.71% in probability surveys and an MAB of 5.36% with a 95%-confidence interval from 4.42% to 6.30% in nonprobability surveys. Table 2 shows that the regression coefficient from the mixed-effects meta-regression model is -2.18 and highly significant, the R^2-statistic for this model is 10.93%. These findings suggest that probability surveys are more representative than nonprobability surveys, which is in accordance with our expectations.

Response rates. The second moderator we examine is the response rate, which is a continuous variable. Overall, we find that the response rate is highly significantly associated with the degree of representativeness as measured using R-Indicators and the MAB. This is in accordance with our second expectation (H2).

As displayed in Table 2, the number of R-Indicators in our

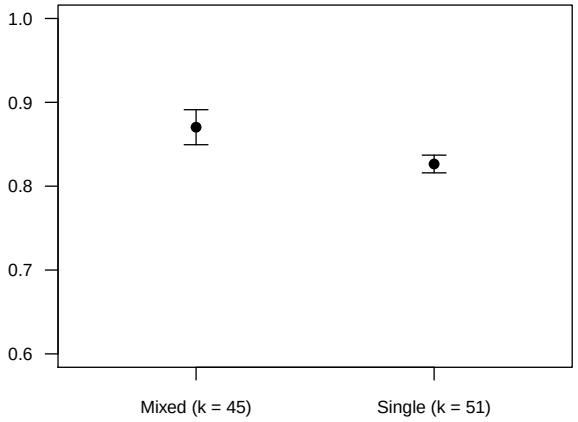

Figure 3. R-Indicator subgroup comparison results by mixed-mode versus single-mode surveys as a moderator

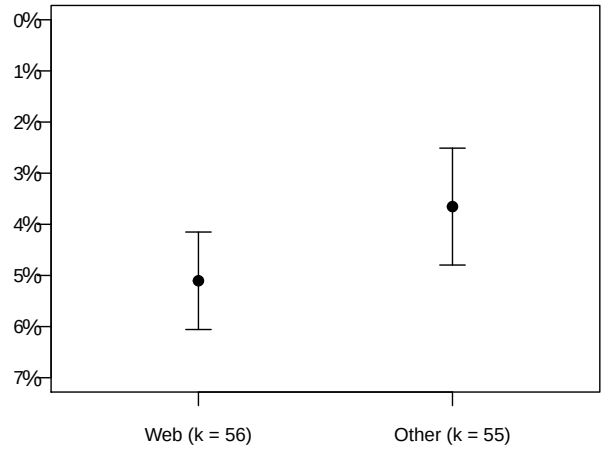

Figure 4. Subgroup comparison results by web surveys versus other single-mode surveys as a moderator

mixed-effects meta-regression model of the response rate on the R-Indicators is 104. The model's rescaled coefficient is 0.14, significant, and has a standard error of 0.04. The R^2-statistic of this model is 8.23%. The number of MAB studies that we use in our moderator analysis on the response rate is 90. The mixed-effects meta-regression produces a rescaled coefficient of -2.05 that is significant and has a standard error of 1.14 as well as an R^2-statistic of 2.22%.

Overall, we find across R-Indicators and the MAB that the results of our meta-regressions support our expectation that response rates are positively associated with the reported degree of representativeness.

Mixed-mode versus single-mode surveys. The third moderator we investigate is a dichotomous variable on mixed-mode surveys versus single-mode surveys. Our results suggest that mixed-mode surveys are more representative than single-mode surveys (H3) although there are too few mixed-mode surveys (eight surveys) among the MAB studies to allow for valid conclusions.

There are 45 mixed-mode surveys and 51 single-mode surveys among the R-Indicators in our dataset. Figure 3 shows that the average R-Indicators are 0.87 with a confidence interval from 0.85 to 0.89 in the mixed-mode surveys and 0.83 with a confidence interval from 0.82 to 0.84 in the single-mode surveys. The mixed-effects meta-regression reported in Table 2 has a coefficient of 0.04 with a standard error of 0.01 and an R^2-statistic of 12.89%.

The number of MAB studies is 8 on mixed-mode surveys and 101 on single mode surveys. Because of the insufficient number of cases in the mixed-mode survey subgroup, estimation of a mixed-effects meta-regression model on this moderator is not feasible.

Overall, we find support for our expectation that mixed-mode surveys are more representative than single-mode surveys on the R-Indicators while for the MAB, the available evidence at present does not allow us to conduct moderator analyses.

Web surveys versus other single-mode surveys. Next, we investigate the association between survey representativeness and web surveys versus other single-mode surveys as a moderator. In accordance with our expectation (H4), our findings indicate, that web surveys are less representative than single-mode surveys as measured using the MAB. There is only one R-Indicator study that examines web surveys. Therefore we cannot conduct a moderator analysis on the R-Indicator data.

The number of R-Indicator studies is 1 for the web surveys and 50 for other single-mode surveys. The remaining R-Indicator studies identified do not cover single-mode surveys. Because there is only one web survey in our R-Indicator data, we do not estimate a mixed-effects meta-regression model on this moderator.

There are 56 web surveys and 45 other single-mode surveys in our MAB data. The average MAB is 5.10% with a confidence interval from 4.15% to 6.06% in the web surveys and 3.65% with a confidence interval from 2.51% to 4.80% in the other single mode surveys. The mixed-effects meta-regression on this moderator estimates a regression coefficient of 1.57 that is significant and has a standard error of 0.65. The R^2 of this model is 4.37%.

In sum, the MAB data support our expectation that web surveys are less representative than other single-mode surveys while there is only one web survey among the R-Indicator studies in our data so that we cannot draw valid conclusions on this moderator for R-Indicators.

Number of auxiliary variables. The last moderator we examine is the number of auxiliary variables that enters the representativeness assessment. Overall, we find no signifi-

cant evidence in favour of our expectation (H5). This suggests that there is no negative association between the number of auxiliary variables included in a representativeness assessment and the reported degree of representativeness.

There are 104 R-Indicator studies that we use in the moderator analysis concerning the number of auxiliary variables. The rescaled regression coefficient in the mixed-effects meta-regression model is 0.05, not significant, and has a standard error of 0.00. The R^2 of the model is 0.00%. Regarding the MAB, there are 104 studies included in the moderator analysis of the number of auxiliary variables. The rescaled mixed-effects meta-regression coefficient is -10.28, significant, and has a standard error of 4.48. The R^2-statistic of this model predicts a model fit of 3.81%.

Overall, the findings on this moderator are not entirely robust across two effect sizes. However Our findings show that neither on the R-Indicators, nor on the MABs there is a negative association between the auxiliary variables and the reported degree of representativeness.

6 Summary and conclusion

In this section, we revisit our expectations from above and interpret the empirical findings. Then, we discuss practical implications and limitations of our paper as well as avenues for further research.

In accordance with our first expectation (H1), we find that probability surveys are more representative than nonprobability surveys with regard to the MAB. In line with our second expectation (H2), we find that the response rate is positively associated with representativeness on R-Indicators as well as the MAB. In compliance with our third expectation (H3), we find that mixed-mode surveys are more representative than single-mode surveys on the R-Indicators. Furthermore, in accordance with our fourth expectation (H4), we find that web surveys are less representative than other single-mode surveys with regard to the MAB. Contrary to our fifth expectation (H5) we find that there is no negative association between the number of auxiliary variablesand the reported degree of representativeness.

We expected results to be consistent across the two measures of representativeness. Due to insufficient numbers of cases per category, however, some expectations could only be tested for one of the two measures. There are two moderator analyses (on the response rate and the number of auxiliary variables) that we could conduct on both representativeness measures that both yield consistent findings across the effect sizes. Both the R-Indicators and the MAB indicate that the higher the response rate is the higher is the reported degree of representativeness. In addition, we find on both the R-Indicators and the MAB that there is no negative association between the number of auxiliary variables and the reported representativeness. However, only the MAB indicates that in fact the association between the number of auxiliary variabels and the degree of representativeness might be positive.

Based on these findings and disregarding potential interdependencies between moderator variables in our data that might confound our results, we would recommend survey practitioners aiming to design representative surveys to use probability sampling, to put every effort into increasing response rates, and to consider mixed-mode survey designs encompassing the web mode. We would also recommend documenting the achieved degree of representativeness using the available auxiliary data to allow assessing the general quality of the data and to allow further research into the association between survey design characteristics and survey representativeness.

This study does, however, face some restrictions that limit the generalizability of the results. Firstly, the analyses do not allow for causal inferences. We cannot claim that probability sampling, high response rates, mixed-mode designs, and the other-than-web survey modes cause a survey to be representative. Our analyses are limited to identifying existing associations. These associations might be confounded with other survey characteristics that we cannot control for in our analyses, such as the data quality of the auxiliary data used in the primary research. Secondly, since our analyses are based on the existing literature, publication biases might influence our results (see the Supplementary Material for sensitivity analyses). This is especially the case when researchers do not publish results that show that a survey has a low degree of representativeness. Lastly, and potentially partly due to publication bias, the number of cases on some survey characteristics is too small to conduct all planned analyses. In the existing literature, there are few studies that assess the representativeness of mixed-mode surveys using benchmark comparisons that allow the computation of the MAB or that investigate the representativeness of web surveys using R-Indicators. In addition, there are no studies that assess the representativeness of a survey using R-Indicators as well as benchmark comparisons. Furthermore, there are a number of potentially influential moderators that we could not take into account either because too much information was missing from the publications identified, or the information was too ambiguous, for instance regarding whether and how post-survey adjustments were applied.

We would therefore like our recommendations to be taken with caution and this meta-analysis to be considered as encouragement for more systematic primary research in these areas, filling the gaps identified. If there were more primary research available, future meta-analytic replications could be conducted using more advanced modeling, such as multivariate regression analyses, simultaneously controlling for potentially influential moderators. In a future replication and extension of this meta-analysis, besides considering additional primary studies, one might want to add additional

moderators of practical importance, such as whether surveys apply a responsive mode design, which formula is used to compute the response rate, or whether adjustment weights were applied.

Acknowledgements

The authors gratefully acknowledge support from the Collaborative Research Center (SFB) 884 "Political Economy of Reforms"(project A8), funded by the German Research Foundation (DFG) and from GESIS – Leibniz Institute for the Social Sciences. The authors would like to especially thank Barry Schouten, Joe Sakshaug, and the participants of the International Workshop on Household Survey Nonresponse 2016 in Oslo, Norway, for their feedback on early versions of this paper. We would also like to thank Christian Bruch and Barbara Felderer for their valuable feedback on variance estimation and our student assistants Sophia Fauser, Margarita Kozlova, Linda Beck, and Thomas Alcock for their help with coding the data.

References

American Association for Public Opinion Research. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Retrieved from ttp : / / www . aapor . org / AAPOR _ Main / media / publications / Standard - Definitions20169thedition % 20final.pdf

Baker, R., Brick, J., Bates, N., Battaglia, M., Couper, M. P., Dever, J. A., ... Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. Retrieved from https : / / www . aapor . org / AAPOR _ Main / media / MainSiteFiles / NPS _ TF _ Report_Final_7_revised_FNL_6_22_13.pdf

Bethlehem, J. & Biffignandi, S. (2012). *Handbook of Web Surveys*. New York: NY: John Wiley & Sons.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. Hoboken, NJ: John Wiley & Sons.

Blohm, M. & Koch, A. (2015). Führt eine höhere Ausschöpfung zu anderen Umfrageergebnissen? Eine experimentelle Studie zum ALLBUS 2008. In J. Schupp & C. Wolf (Eds.), *Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen* (pp. 85–129). Wiesbaden: Springer Fachmedien.

Blom, A. G. (2016). Survey Fieldwork. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (382–397). London: UK: Sage Publications.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. New York, NY: Wiley.

Bosnjak, M. (2017). Mixed-mode surveys and data quality: Meta-analytic evidence and avenues for future research. In S. Eifler & F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 11–25). Wiesbaden: Springer Fachmedien.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition biases in different development stages of a probabilistic online panel. *Fields Methods*, *25*(4), 339–360.

Brick, J. M. & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*(1), 215–238.

Brick, J. M. & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36–59.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, London: The Guilford Press.

Chang, L. & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641–678.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, *64*(1), 464–494.

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

Couper, M. P. & Bosnjak, M. (2010). Internet Surveys. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 527–550). San Diego, CA: Elsevier.

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research*, *36*(1), 131–148.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*(1), 87–98.

de Leeuw, E. D. & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, A. D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse* (pp. 41–54). New York, NY: Wiley & Sons.

Dillman, D. A. (2008). The logic and psychology of constructing questionnaires. In *International handbook of survey methodology* (pp. 161–175). New York, NY: Lawrence Erlbaum Associates.

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research*, *47*(6), 615–639.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other mea-

sures of statistical accuracy. *Statistical Science*, *1*(1), 54–75.

Emerson, M. O., Sikkink, D., & James, A. D. (2010). The panel study on American religion and ethnicity: Background, methods, and selected results. *Journal for the Scientific Study of Religion*, *49*(1), 162–171.

Gelman, A., Goel, S., Rothschild, D., & Wang, W. (2017). High-frequency polling with non-representative data. In D. Schill, R. Kirk, & A. E. Jasperson (Eds.), *Political Communication in Real Time: Theoretical and Applied Research Approaches* (pp. 89–105). Routledge.

Goldenbeld, C. & de Craen, S. (2013). The comparison of road safety survey answers between web-panel and face-to-face: Dutch results of SARTRE-4 survey. *Journal of Safety Research*, *46*, 13–20.

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. P., 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), 1–6.

Groves, R. M. & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*, *169*(3), 439–457.

Groves, R. M., J., F. F., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

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. *Public Opinion Quarterly*, *72*(2), 167–189.

Hedges, L. V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysi*. London: UK: Academic Press.

Hedges, L. V. & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504.

Jäckle, A., Lynn, P., & & Burton, J. (2013). *Going Online with a Face-to-Face Household panel: Initial Results from an Experiment on the Understanding Society Innovation Panel*. Colchester, UK.: Economic and Social Research Council.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, *64*(1), 125–148.

Kish, L. (1965). *Survey Sampling*. New York, NY: John Wiley & Sons.

Klausch, T., Hox, J., & Schouten, B. (2015). Selection error in single-and mixed mode surveys of the dutch general population. *Journal of the Royal Statistical Society: Series A*, *178*(4), 945–961.

Kreuter, F. (2013). *Improving Surveys with Paradata*. New York, NY: Wiley.

Kreuter, F. & Olson, K. (2011). Multiple auxiliary variables in nonresponse adjustment. *Sociological Methods & Research*, *40*(2), 311–332.

Kruskal, W. & Mosteller, F. (1979a). Representative sampling, I: Non-scientific literature. *International Statistical Review*, *47*(1), 13–24.

Kruskal, W. & Mosteller, F. (1979b). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review*, *47*(2), 111–127.

Kruskal, W. & Mosteller, F. (1979c). Representative sampling, III: The current statistical literature. *International Statistical Review*, *47*(3), 245–265.

Lohr, S. (2010). *Sampling: Design and Analysis*. Brooks/Cole: Cengage Learning.

Loosveldt, G. & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, *2*(2), 93–105.

Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(5), 669–686.

Luiten, A. & Schouten, B. (2012). Tailored fieldwork design to increase representative household survey response: an experiment in the survey of consumer satisfaction. *Journal of the Royal Statistical Society: Series A*, *176*(1), 169–189.

Lundquist, P. & Särndal, C. E. (2012). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, *29*(4), 557–582.

Macer, T. (2014). Online panel software. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 413–440). London, UK: John Wiley & Sons.

Malhotra, N. & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. *Political Analysis*, *15*(3), 286–323.

Medway, R. L. & Fulton, J. (2012). When more gets you less: a meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly*, *76*(4), 733–746.

OECD. (2014). *Measuring the Digital Economy: A New Perspective*. OECD Publishing.

Peytchev, A., Conrad, F. G., Couper, M. P., & Tourangeau, R. (2010). Increasing respondents' use of definitions in web surveys. *Journal of official statistics*, *26*(4), 633.

Potthoff, P., Heinemann, L. A. J., & Güther, B. (2004). A household panel as a tool for cost-effective health-related population surveys: validity of the "Healthcare Access Panel". *German Medical Science*, *2*(1), 1–8.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–316). New York, NY: Russell Sage Foundation.

Rendtel, U. & Pötter, U. (1992). Über Sinn und Unsinn von Repräsentationsstudien. *DIW Discussion Papers, 61*.

Schenker, N. & Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, *22*(4), 425–446.

Schnell, R. (1993). Die Homogenität sozialer Kategorien als Voraussetzung für "Repräsentativität" und Gewichtungsverfahren. *Zeitschrift für Soziologie*, *22*(1), 16–32.

Schouten, B. & Cobben, F. (2007). R-indexes for the comparison of different fieldwork strategies and data collection modes. voorburg/heerlen: cbs. *Statistics Netherlands: Discussion paper 07002*.

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

Shlomo, N., Skinner, C., Schouten, B., de Heij, V., Bethlehem, J., & Ouwehand, P. (2009). Indicators for Representative Response Based on Population Totals. In *RISQ Work package 3, Deliverable 2.2*. Retrieved from http://www.risq-project.eu/papers/RISQ-Deliverable-2-2-V1.pdf

Toepoel, V., Das, M., & van Soest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly*, *72*(5), 985–1007.

Toepoel, V. & Ludtig, P. (2015). Online surveys are mixed-device surveys. Issues associated with the use of different (mobile) devices in web surveys. *Methods, data, analyses*, *9*(2), 155–162.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The Science of Web Surveys*. Oxford: Oxford University Press.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, *21*(4), 589–624.

Viechtbauer, W. (2010). Conducting Meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Vonk, T., van Ossenbruggen, R., & Willems, P. (2006). The effects of panel recruitment and management on research results, a study among 19 online panels. *ESOMAR Publication Services*, *317*, 79–99.

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.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991.

Wozniak, K. H. (2016). Perceptions of Prison and Punitive Attitudes: A Test of the Penal Escalation Hypothesis. *Criminal Justice Review*, *41*(3), 352–371.

Yan, T. & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51–68.

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