DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 93–115, http://dx.doi.org/10.2478/JOS-2019-0005

The Effect of Survey Mode on Data Quality: Disentangling Nonresponse and Measurement Error Bias

Barbara Felderer[1], *Antje Kirchner*[2], *and Frauke Kreuter*[3]

More and more surveys are conducted online. While web surveys are generally cheaper and tend to have lower measurement error in comparison to other survey modes, especially for sensitive questions, potential advantages might be offset by larger nonresponse bias. This article compares the data quality in a web survey administration to another common mode of survey administration, the telephone.

The unique feature of this study is the availability of administrative records for all sampled individuals in combination with a random assignment of survey mode. This specific design allows us to investigate and compare potential bias in survey statistics due to 1) nonresponse error, 2) measurement error, and 3) combined bias of these two error sources and hence, an overall assessment of data quality for two common modes of survey administration, telephone and web.

Our results show that overall mean estimates on the web are more biased compared to the telephone mode. Nonresponse and measurement bias tend to reinforce each other in both modes, with nonresponse bias being somewhat more pronounced in the web mode. While measurement error bias tends to be smaller in the web survey implementation, interestingly, our results also show that the web does not consistently outperform the telephone mode for sensitive questions.

Key words: Mode effects; telephone survey; web survey; combined bias.

1. Introduction

Researchers often use evidence of bias in survey estimates to assess and compare data quality among different modes of survey administration. There are two major problems with this approach. First, by assessing combined bias as a measure of data quality, researchers mix different sources of bias, for example, bias due to differential coverage, nonresponse or measurement error which might each differ in magnitude and direction, and do so differently for different survey modes. In the worst case, seemingly unbiased estimates in one survey mode might actually be more biased compared to another survey mode, when each source of bias is investigated individually. Hence, investigating combined bias leaves researchers guessing about the sources of bias and makes it hard to

[1] University of Mannheim, Mannheim, 68131, Germany. Email: felderer@uni-mannheim.de
[2] RTI International, 3040 E. Cornwallis Road, RTP, NC 27709, U.S.A. Email: akirchner@rti.org
[3] University of Maryland, College Park, Maryland 20742, U.S.A. Email: jkreuter@umd.edu

Acknowledgments: This material is partly based upon work supported by the National Science Foundation under Grant No. SES-1132015. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Stephanie Eckman, Julie Korbmacher, Renae Reis, and Joe Sakshaug for helpful comments on earlier drafts of this article.

derive practical implications and inform survey designs. Understanding the individual contributions of, for example, nonresponse and measurement error bias to the total survey error and potential bias, and how these differ by survey mode is of utmost importance (Biemer 2010). Another common challenge that researchers face is how to actually measure bias. Often, researchers compare sample estimates to other aggregate population estimates, or they rely on assumptions, such as the "more-is-better assumption" for undesirable behaviors, to assess which survey mode performs better. This approach does not necessarily inform researchers about which mode is the least biased and comes closest to the "truth." Ideally, bias is assessed by comparing survey estimates with auxiliary and gold standard data for the same sampled individuals. This measure can then be used to inform the research community about the overall effects of survey mode on data quality. However, often researchers do not have access to this kind of information; either because the data are nonexistent, for example when investigating attitudes, or, the data are unavailable for reasons of data confidentiality.

The key features of this study that allow us to address both challenges are its experimental design and the use of a unique combination of large scale survey data and administrative records from German social security records – containing rich information on a variety of labor market related and demographic characteristics. The specific design of our study enables us not only to measure bias directly but also to disentangle different sources of error contributing to bias among two commonly used survey modes, telephone and web. More specifically, we separate bias due to nonresponse and measurement error, which ultimately helps researchers to understand the nature and relative contribution of each bias source to combined bias. These results are particularly relevant for researchers planning to use a mixedmode design, in which it is generally assumed that the strengths of one mode will compensate the weaknesses of another, and thus enhance data quality, while at the same time potentially reducing costs.

The following section will provide a brief overview of why differences in bias between survey modes are to be expected, how bias has been assessed in past studies, and the research questions of our article. Section 2 will introduce the design, data and methods used in our analyses. The results are described in Section 3. The article concludes with a summary and discussion of the main results in Section 4.

1.1. Why Do We Expect Differences in Bias Across Survey Modes?

Both data collection modes, telephone and web, have their particular strengths and weaknesses with respect to achieving survey participation and response accuracy. Survey mode can affect the sample composition, as different modes have different coverage error. In order to participate in a telephone survey, sample members have to have access to a telephone, whereas for web surveys, access to the Internet is a prerequisite. In list-assisted sampling designs this might lead to differential coverage error if for example, more sample members have access to a telephone than the Internet. To the extent that coverage error is systematic and differs by mode, this might introduce differential bias. Sample composition may also differ, as the ability to establish contact with the target person and respondents' willingness and capacity to complete the survey differs across survey modes (Dillman et al. 2002, 6). Self-administered surveys, for example, tend to have lower response rates compared to interviewer-administered surveys (De Leeuw 2005). Additionally, if relevant subgroups self-select depending on the survey mode, this can introduce differential nonresponse bias, should the selection mechanism be related to survey variables of interest (Groves and Couper 1998; De Leeuw et al. 2008; Biemer 2010; Kreuter et al. 2010). Measurement error results from a difference in the respondents' survey report and the (unobserved) true value, for example, due to misunderstanding a question, failure to retrieve the correct information or incorrect reporting (for a review, see, for example Biemer and Lyberg 2003). If the misreporting mechanism is related to the survey outcome of interest and systematically differs between survey modes, this again will result in differentially biased estimates.

In addition to coverage, nonresponse and measurement error, there are other potential sources of error that may bias survey estimates, including specification or adjustment errors (Biemer 2010). In line with existing research, we will focus on bias due to nonresponse and measurement error when investigating mode differences, as those can be expected to be the main drivers of differential bias. Previous empirical research on mode differences shows that response rates tend to be generally lower in web surveys (Lozar Manfreda et al. 2008). This increases the potential for selectivity and *nonresponse bias* in the web survey compared to the telephone survey (Fricker et al. 2005; O'Neill and Dixon 2005; Abraham et al. 2006; Groves 2006; Letourneau and Zbikowski 2008). Although response rates are known to be lower in web surveys, web surveys have several benefits over more traditional surveying methods. Web surveys are generally less cost intensive, the data are available almost immediately, respondents can take the survey at their own pace and convenience, and it provides a more private survey setting (Callegaro et al. 2014). Due to this latter fact, *measurement error bias* might be less pronounced for certain types of questions in the self-administered web mode compared to interviewer-administered modes (Kreuter et al. 2008; Chang and Krosnick 2009, 2010; Sakshaug et al. 2010). While we would expect to see little difference between web and telephone for factual items that are less prone to misreporting (Atkeson et al. 2014), survey mode might influence response accuracy for items that are sensitive in nature or those that evoke concerns of social (un)desirability (Kreuter et al. 2008; Chang and Krosnick 2009, 2010; Sakshaug et al. 2010; Atkeson et al. 2014).

In line with the literature on sensitive questions (Lee 1993; Groves 2004; Bradburn et al. 2004), traits that are positively valued – such as regular employment – should be overreported, while undesirable traits – such as welfare receipt or marginal employment – should be underreported. While a respondent might choose to give a correct answer to a sensitive question in the web mode due to the increased privacy, they might respond differently to the same question in a telephone interview with an interviewer present (Kreuter et al. 2008; Chang and Krosnick 2009, 2010; Malhotra et al. 2014; Roberts et al. 2014). Hence overall, the combined bias due to nonresponse and measurement error might actually be smaller in the web administration compared to the telephone mode. Again, if one were to investigate combined bias only, researchers would never know how the bias terms interact. The main focus in our article will be the interaction of nonresponse and measurement error and how each contributes to bias. We will discuss other potential sources of bias as appropriate.

1.2. How has Bias been Assessed in the Past?

Mode effects studies are usually not able to directly differentiate nonresponse bias and measurement error bias in a survey estimate, but instead rely on indirect indicators, benchmarks or assumptions (such as “more-is-better” for sensitive questions) for a comparison of data quality in the absence of gold standard validation data. External population benchmarks for some variables are often used to assess nonresponse bias (Fricker et al. 2005; Yeager et al. 2011; Malhotra et al. 2014). Bias due to measurement error is often assessed using indirect indicators of survey satisficing, including non-differentiation, item missingness, the use of extreme values, acquiescence or socially desirable responses (McCabe et al. 2002; Duffy et al. 2005; Chang and Krosnick 2009, 2010; Atkeson et al. 2011, 2014; Hope et al. 2014; Malhotra et al. 2014). Other mode effects studies use panel information from previous waves to assess bias due to nonresponse and measurement error (Sax et al. 2003; Duffy et al. 2005; Braunsberger et al. 2007; Chang and Krosnick 2009; Vannieuwenhuyze et al. 2010; Roberts et al. 2014). The results of these studies are not comparable with cross-sectional studies, since nonresponse or measurement error bias for the initial wave is usually not assessed and data provided in the initial wave need not necessarily be more accurate. Also, typically none of these studies analyze the combined effect of nonresponse and measurement error on survey estimates.

The most powerful designs to study differential effects of survey mode on nonresponse bias and measurement error bias are those that, in addition to random assignment of survey mode, have auxiliary data with extraordinary data quality available for all sample cases for the characteristics under study. Few studies allow for a validation study comparing web surveys to other forms of survey administration (e.g., McCabe et al. 2002; Sax et al. 2003; Sanders et al. 2007; Kreuter et al. 2008; Dillman et al. 2009; Sakshaug et al. 2010; Atkeson et al. 2011; Stephenson and Crête 2011; Atkeson et al. 2014). Particularly relevant for our study are the results of the validation studies by Kreuter et al. (2008) and Sakshaug et al. (2010), as they focus on the interaction of both sources of bias for a variety of variables and question types. Kreuter et al. (2008) find significant differences in completion rates comparing telephone, interactive voice recording, and web. The initial screening interview was conducted by phone, and screener respondents assigned to the web had the lowest completion rates. However, the results do suggest that sensitive items are reported more accurately in the web mode. Regarding the interaction of both error sources, both studies find that bias due to measurement error dominates nonresponse error for sensitive items, while nonresponse error tends to be larger for neutral and socially desirable items (Kreuter et al. 2008; Sakshaug et al. 2010). For the most part, Sakshaug et al. (2010) find that the different error sources reinforce each other and do not cancel each other out.

The main contribution of our article is a systematic assessment of the relative contribution of nonresponse and measurement error to the combined bias in survey estimates in each survey mode using large scale validation data that is known to be of very good data quality and can serve as a gold standard. Building on past validation studies by Kreuter et al. (2008) and Sakshaug et al. (2010), we analyze the interaction of nonresponse and measurement error for demographic and sensitive questions. Whereas these studies focus on a very specific subpopulation of student alumni, the scope of our analyses is broader. Our analyses rely on a stratified random sample of the adult labor-force in Germany. As such, more generalizable inferences can be drawn regarding the implications of the choice of a particular survey mode. Existing validation studies also typically analyze nonresponse or measurement error bias for mean statistics of certain survey items, but do not assess bias in distributions. Our analysis also investigates and compares bias in distributions for two metric items – age and income.

2. Data and Methods

2.1. Study Design and Administrative Data

The Integrated Employment Biographies (IAB 2011) maintained by the German Federal Employment Agency (FEA) serve as the sampling frame for our study. This administrative register combines information on individuals' times of (un)employment in Germany and welfare, also called basic income support ("Unemployment Benefit II", abbreviated UB II). These registers cover approximately 86% of the German labor force, starting from 1975, including all employees who are subject to social security contributions, individuals seeking employment, and those on welfare, excluding only self-employed and civil servants. This sampling frame is comprehensive, up to date (with only a short time lag) and accurate, since it contains payment-relevant information, as their main use is by the German statutory pension insurance to administer and calculate pension claims, benefit claims, and payments. For the analysis, we use updated versions of the data sets that are used to generate the IEB (IAB 2012, 2013).

Sampling from the FEA registers provides us with detailed information on (un)employment and welfare benefit receipt for all sampled cases, that is, respondents and nonrespondents to the survey. Sampled individuals were randomly assigned to one of the two modes. We specifically designed and worded both surveys such that survey responses can be validated given the information in the administrative data, including socio-demographic (e.g., gender and age) and sensitive information (e.g., income and welfare benefit receipt). Furthermore, we only use data that are known to be accurate and complete, and can thus serve as gold standard (Jacobebbinghaus and Seth 2007).

More specifically, we investigate nonresponse bias, measurement error bias, and combined bias in gender (0 male, 1 female), mean age (and categories: 18–29, 30–39, 40–49, 50–59, 60+ years), currently employed (0 no, 1 yes), type of employment (marginal employment with an income of EUR 400 and less, regular employment with an income of EUR 401 and more), past receipt of UB II (past 12 months), and mean monthly labor income (and income terciles) in euros, if currently employed. In Germany, respondents think in terms of monthly income and not annual income. Thus, the survey items ask for monthly income. However, labor income in the administrative records is captured only as the total gross income in a given employment spell (typically one year). Thus, monthly income has to be derived from this measure. The basic assumption is that all income is equally distributed over the months of a certain spell. Also, income is top coded in the administrative data, the limit being a yearly income of approximately EUR 57,000 in the states of former East Germany and EUR 66,000 in the states of former West Germany, depending on the type of pension insurance. Since this affects all administrative data equally and survey mode was randomly assigned, inferences with respect to the relative comparison across modes are still valid.

Administrative data used for the analyses were extracted from either the last valid (employment) period or, in the case of an ongoing period, from the respective interview month for respondents. The date of the last interview in either mode is taken as the reference date for nonrespondents.

2.2. Survey Data

Overall, a sample of 24,236 eligible adults was drawn in June 2011 from the FEA registers and randomly assigned to one of two survey modes: 12,400 individuals were randomly assigned to the telephone mode, while 11,836 individuals were assigned to complete the survey online. Addresses, and in part telephone numbers, were available for all sampled individuals in the frame.

Only 9,332 of the individuals assigned to the telephone mode turned out to have valid phone numbers. 2,400 individuals completed the telephone survey, corresponding to an overall response rate of 19.35% among sample members in the telephone mode of the experiment (RR1 according to AAPOR 2011). In the web mode, 1,311 letters were returned to sender due to an incorrect address, leaving 10,525 individuals who received the invitation to the survey. Of those, 1,082 individuals completed the web survey. The overall response rate among sample members in the web survey was 9.14%. Table 1 provides an overview of the sample sizes and response rates.

The telephone survey was fielded during the months of August to October 2011 and the web survey from February to mid-April 2012. Prior to fieldwork, all sampled individuals received an advance/invitation letter inviting them to participate in the government survey "Work and Consumption in Germany", commissioned by the Institute for Employment Research (IAB), and carried out by the LINK Institute. The invitation letter for the web mode also contained all relevant login information and a conditional incentive of EUR 3. Two weeks after the start of fieldwork, a reminder was sent to all sampled cases of the web survey component.

Both questionnaires contained questions relating to employment biographies that are conceptually equivalent to the administrative data described above. We only analyze questions that were fielded in exactly the same way in both surveys, except for one question about past receipt of unemployment benefit and will provide more information below. Both surveys were kept as similar as possible and only differed in some of the

Table 1. Response rates across modes of data collection.

	Telephone survey	Web survey
Sampled	12,400	11,836
Valid contact information	9,332	10,525
Completed	2,400	1,082
Response rate (ref. sampled)	19.35%	9.14%

questions in other parts of the questionnaire. The average survey completion time was 21 minutes for the telephone interview and 15 minutes in the web mode.

2.3. Methods

In order to assess nonresponse bias, we only include individuals for whom we have valid contact information (see Table 1) as all other individuals never received the invitation to participate in the survey. For the telephone, this means that we include individuals for whom we have a valid telephone number. Given that we do not have this kind of information for individuals assigned to the web mode, we include those who actually received our invitation to participate in the survey, that is, individuals whose invitation letter was not returned to sender. This approach implies that while we can clearly separate bias due to coverage error and nonresponse for individuals assigned to the telephone; the same does not hold for those individuals assigned to the web mode. A small portion of the nonresponse bias that we investigate will actually be coverage bias, although we expect this to be minimal as the internet penetration in Germany is quite high. Approximately 79% of the German households had internet at the time of the survey, with an additional 14% of the noninternet households having internet access outside home, for example, at work (Statistisches Bundesamt 2013). Furthermore, as we are comparing survey packages, our results give a realistic assessment of relative biases in these two survey modes. For simplicity we will refer to this as nonresponse bias in both survey modes. In a sensitivity analysis, we replicated our analysis including all sample cases, for example, all individuals assigned to the telephone mode including those without valid telephone numbers and all individuals assigned to the web mode including those whose invitation letter was returned to sender. Results can be found in Subsection 5.2. Appendix B.

Respondents to both modes are part of the measurement error analysis. Due to data protection regulations we are not able to match the data on an individual level, but rather compare the proportion of respondents reporting a certain characteristic in the survey with the proportion of respondents who have this same characteristic in the administrative data. Because we analyze survey and administrative data separately and do not combine data sources, we are not restricted to those respondents who consented to data linkage, and can include all survey respondents. While this has the advantage that our study is not subject to potential linkage nonconsent bias (Sakshaug and Kreuter 2012), it has the disadvantage that we cannot examine measurement error at an individual level.

We compare bias in mean statistics (and distributions) for the variables introduced above across both modes using the:

a) full sample administrative data (fs): $\frac{1}{N}\sum_{i=1}^{N} y_{i,\text{admin}}$ with N being the sample size;
b) respondent sample administrative data (resp): $\frac{1}{n}\sum_{i=1}^{n} y_{i,\text{admin}}$ with n being the number of completed interviews; and
c) respondent sample survey data (svy): $\frac{1}{n}\sum_{i=1}^{n} y_{i,\text{survey}}$.

For an assessment of the combined bias we will compare a) the true value from the full sample administrative data to c) the respondent sample survey data. We will then break this combined bias into its components: nonresponse bias is assessed by comparing estimates from the full sample administrative data (a) to estimates from respondent sample administrative data (b). Bias due to measurement error is assessed by comparing estimates based on the respondent sample administrative data (b) to estimates based on respondent sample survey data (c).

As information is available for all sample cases, the estimation of nonresponse bias is straightforward. Nonresponse bias (nr) is the difference of the true nonrandom sample value according to administrative records (adm) for the full sample (fs) and the mean computed using the respondents (resp) only. In order to compare nonresponse bias between variables and modes, nonresponse bias is standardized by the full sample mean of each variable of interest multiplied by 100 to obtain the relative nonresponse bias in percent:

$$\widehat{rel.bias}(\hat{\bar{y}}_{nr}) = \frac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} * 100. \tag{1}$$

As our analysis focuses on relative biases in mean statistics and not on the mean statistics themselves, we need to estimate the variances of the relative biases and adapt significance tests accordingly.

More specifically, the variance of $\widehat{rel.bias}(\hat{\bar{y}}_{nr})$ is given by:

$$Var(\widehat{rel.bias}(\hat{\bar{y}}_{nr})) = Var\left(\frac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} * 100\right). \tag{2}$$

As $\bar{y}_{adm,fs}$ is nonrandom, we can write:

$$Var(\widehat{rel.bias}(\hat{\bar{y}}_{nr})) = \frac{100^2}{\bar{y}^2_{adm,fs}} (Var(\hat{\bar{y}}_{adm,resp}) + Var(\bar{y}_{adm,fs}) + 2Cov(\bar{y}_{adm,fs}, \hat{\bar{y}}_{adm,resp})). \tag{3}$$

From $\bar{y}_{adm,fs}$ being nonrandom, it also follows that $Var(\bar{y}_{adm,fs}) = 0$ and $Cov(\bar{y}_{adm,fs}, \hat{\bar{y}}_{adm,resp}) = 0$. Thus, the variance reduces to:

$$Var(\widehat{rel.bias}(\hat{\bar{y}}_{nr})) = \frac{100^2}{\bar{y}^2_{adm,fs}} Var(\hat{\bar{y}}_{adm,resp}). \tag{4}$$

This leads to the test statistic for a one-sample Z-Test for evaluating the significance of individual relative biases:

$$z = \frac{\frac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} * 100}{\sqrt{\frac{100^2}{\bar{y}^2_{adm,fs}} Var(\hat{\bar{y}}_{adm,resp})}}. \tag{5}$$

Under the null hypothesis $Var(y_{adm,resp})$ equals $Var(y_{adm,fs})$ so we can substitute $Var(\hat{\bar{y}}_{adm,resp})$ by $Var(y_{adm,fs})/n$ with n denoting the respondent sample size:

$$z = \frac{\dfrac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}}}{\sqrt{\dfrac{Var(y_{adm,fs})}{\bar{y}^2_{adm,fs} n}}} \tag{6}$$

The two-sample Z-Test for comparing the relative biases in the telephone and web mode is then given as:

$$z = \frac{\left(\dfrac{\hat{\bar{y}}_{adm,resp,web} - \bar{y}_{adm,fs,web}}{\bar{y}_{adm,fs,web}}\right)*100 - \left(\dfrac{\hat{\bar{y}}_{adm,resp,cati} - \bar{y}_{adm,fs,cati}}{\bar{y}_{adm,fs,cati}}\right)*100}{\sqrt{\dfrac{100^2}{\bar{y}^2_{adm,fs,web}} Var(\hat{\bar{y}}_{adm,resp,web}) + \dfrac{100^2}{\bar{y}^2_{adm,fs,cati}} Var(\hat{\bar{y}}_{adm,resp,cati})}}. \tag{7}$$

Transforming the counter and substituting $Var(\hat{\bar{y}}_{adm,resp,cati})$ by $Var(y_{adm,fs,web})/n_{web}$ and $Var(\hat{\bar{y}}_{adm,resp,cati})$ by $Var(y_{adm,fs,cati})/n_{cati}$ we derive:

$$z = \frac{\left(\dfrac{\hat{\bar{y}}_{adm,resp,web}}{\bar{y}_{adm,fs,web}} - \dfrac{\hat{\bar{y}}_{adm,resp,cati}}{\bar{y}_{adm,fs,cati}}\right)}{\sqrt{\dfrac{Var(y_{adm,resp,web})}{\bar{y}^2_{adm,fs,web} n_{web}} + \dfrac{Var(y_{adm,resp,cati})}{\bar{y}^2_{adm,fs,cati} n_{cati}}}}. \tag{8}$$

Similar to the estimation of nonresponse bias, bias due to measurement error (me) is straightforward to calculate, as the true values are known from the administrative records. Bias due to measurement error is given as the difference of the mean estimate in the survey data (svy) and the true statistic according to administrative records for all respondents. Standardizing measurement error bias with the mean of the respondents based on the administrative data multiplied by 100 gives us an estimate of the relative bias in percent:

$$\widehat{rel.bias}(\hat{\bar{y}}_{me}) = \frac{\hat{\bar{y}}_{svy,resp} - \bar{y}_{adm,resp}}{\bar{y}_{adm,resp}} *100. \tag{9}$$

In the comparison of respondent sample survey data and respondent sample administrative data, the respondent sample administrative data are taken as the nonrandom gold standard, as they contain the true information of the full sample of respondents. This implies that for this analysis we assume $Var(\bar{y}_{adm,resp} = 0)$ and $Cov(\hat{\bar{y}}_{svy,resp}, \hat{\bar{y}}_{admin,resp}) = 0$ leading to:

$$Var(\widehat{rel.bias}(\hat{\bar{y}}_{me})) = \frac{100^2}{\bar{y}^2_{adm,resp}} Var(\hat{\bar{y}}_{svy,resp}). \tag{10}$$

Like all survey data, some of the survey items are subject to item nonresponse. Very few respondents do not report an employment status (telephone 0.2%; web 3.3%) and past receipt of UB II (telephone 0.2%; web 2.8%). Since we are estimating the proportion of respondents belonging to a certain employment or past UB II status (yes/no), missing information is implicitly treated as a “no” response (e.g., not employed, no UB II receipt) in the assessment of measurement bias. The proportion of item nonresponse is highest in the income information (telephone 13.8%; web 15.7%) which results in a reduction of the case base for the survey estimates which is used in the measurement error analysis. There is no item nonresponse in the reports of gender or age. In a sensitivity analysis, we drop cases with missing information in employment and past UB II status for the corresponding analysis. Neither of our results reported below change substantively.

The combined bias (combined) due to nonresponse and measurement error for a survey statistic is simply the difference between the estimate derived from the full sample administrative data and the respondent sample survey data. The combined bias estimate can be standardized similarly to the other biases to obtain the relative combined bias in percent.

$$\widehat{rel.bias}(\hat{\bar{y}}_{comb}) = \frac{\hat{\bar{y}}_{svy,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} *100. \tag{11}$$

Comparing respondent sample survey data and full sample administrative data, the full sample administrative data means are nonrandom, implying $Var(\bar{y}_{adm,fs} = 0)$ and $Cov(\hat{\bar{y}}_{svy,resp}, \bar{y}_{admin,fs}) = 0$ leading to:

$$Var(\widehat{rel.bias}(\hat{\bar{y}}_{comb})) = \frac{100^2}{\bar{y}^2_{adm,fs}} Var(\hat{\bar{y}}_{svy,resp}). \tag{12}$$

Test statistics for relative measurement error bias and combined bias can be derived in an identical manner as for relative nonresponse bias in Equation 5 and Equation 8.

For the subsequent analyses, we distinguish demographic information (such as gender and age), from potentially sensitive information (type of employment, past receipt of UB II and mean labor income from current employment) and report the results in that order. To reiterate our expectations: irrespective of question type, we would expect there to be a generally lower nonresponse bias in the telephone mode compared to the web. We expect little to no measurement error bias for demographic items in either mode, whereas sensitive questions should be reported more accurately in the web mode. The prediction for combined bias depends on whether both sources of bias enforce or compensate each other.

3. Results

We report the results for each error source by indicator and only report differences in the text that are statistically significant at an alpha level of 0.05 ($p < 0.05$), based on the adapted Z-Tests. Figures 1 to 4 display the relative bias in mean estimates for different variables separated by horizontal lines, by survey mode each due to nonresponse (nr), measurement error (me) and combined bias (combined), including 95%-confidence intervals. Solid triangles indicate bias for the telephone mode and hollow squares indicate bias for the web mode. The dashed vertical line indicates zero percent relative bias. Relative bias estimates, including confidence intervals and test statistics can be found in Subsection 5.1. Appendix A.

Females are significantly overrepresented in both survey modes with biases significantly differing between the two modes (see Figure 1). Not surprisingly, there is virtually no measurement error bias for gender in either mode. The very small discrepancies might be due to the fact that some individuals might identify with a gender other than the sex originally recorded in the administrative data. An overrepresentation of

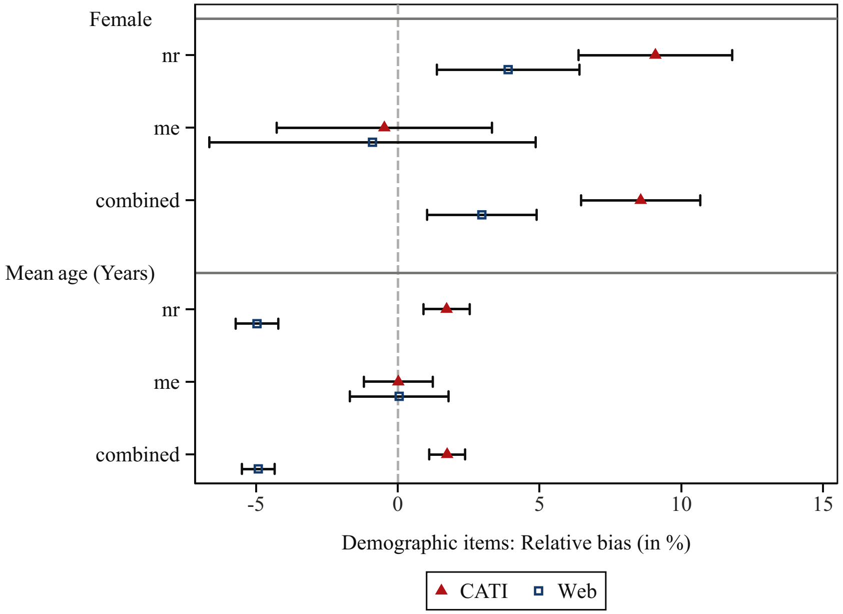

Fig. 1. Relative combined bias for socio-demographic variables, including 95%-confidence intervals.

females together with a very small measurement error bias leads to a relative combined bias that is dominated by nonresponse bias, that is, a significant overestimation of the proportion of women in both survey modes. The relative combined bias is significantly higher in the telephone mode than the web mode.

Our results also show a significant negative nonresponse bias for mean *age* in the web mode and a significant positive nonresponse bias in the telephone mode, although smaller in magnitude. Substantively, this means that younger individuals are overrepresented in the web mode, whereas older individuals are overrepresented in the telephone mode. The difference in biases between the two modes is statistically significant. As expected, there is virtually no relative measurement error bias in mean age for either mode. Combined bias is therefore almost identical to nonresponse bias and implies a significant overestimation of mean age in the telephone mode and significant underestimation in the web mode compared to the population. Relative combined bias differs significantly between the two survey modes.

To study bias in age in more detail, we investigate biases in several *age categories* (see Figure 2). In line with our expectations, younger individuals are overrepresented in the web mode while middle-aged and older individuals are overrepresented in the telephone mode, although relative nonresponse bias is not always significantly different from zero. Except for the middle-aged category “aged 40–49 years” biases differ significantly between both modes. Similarly to mean age, there is no evidence for significant measurement error in any of the age categories, with relative measurement error biases being very close to zero. Again, this results in a combined bias that is almost identical in magnitude to that of nonresponse bias: the proportion of younger individuals is overestimated in the web survey, whereas the proportion of individuals in the middle-aged and older age categories are overestimated relying on telephone survey estimates.

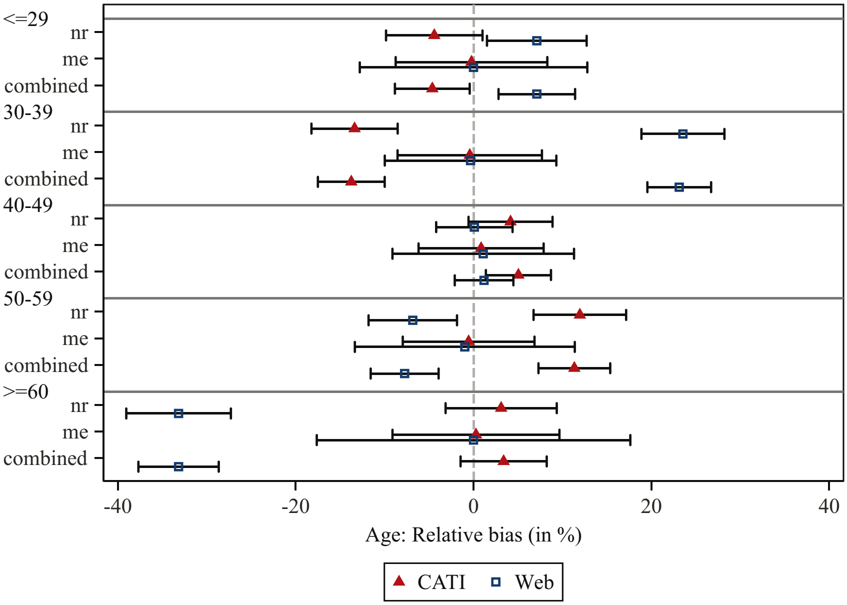

Fig. 2. Relative combined bias for age distribution, including 95%-confidence intervals.

Although combined bias is not significantly different from zero for every mode and age category, relative combined bias differs significantly between survey modes for all age categories except "age 40–49."

We now turn to those items potentially subject to social desirability concerns and sensitivity displayed in Figure 3. Our results suggest that relative nonresponse bias in *employment status* points in the same direction for both modes such that employed

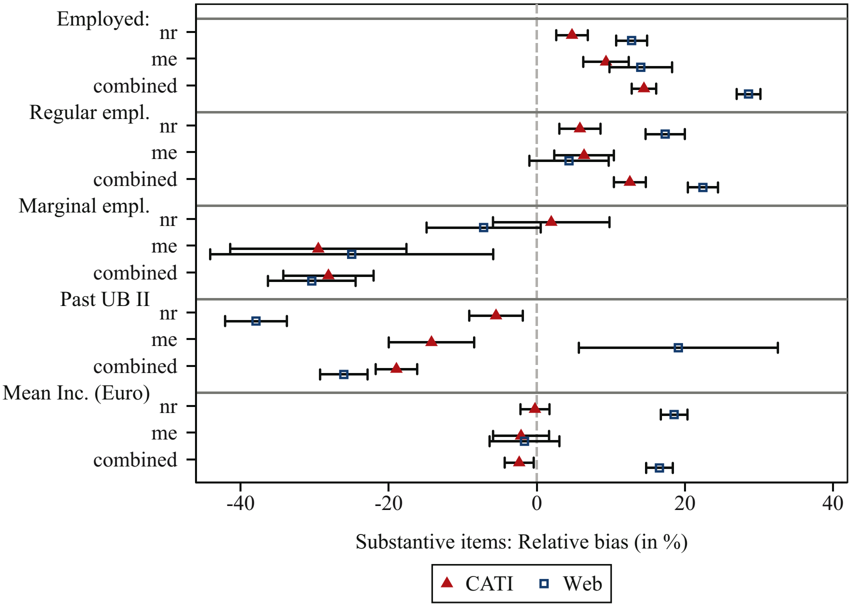

Fig. 3. Relative combined bias for substantive variables, including 95%-confidence intervals.

individuals are significantly overrepresented. This overrepresentation is significant for both survey modes and is significantly higher in the web mode than in the telephone mode. Investigating the different types of employment, we see that individuals in a regular form of employment are significantly overrepresented in both modes. This overrepresentation is, again, significantly higher for web mode than the telephone mode. There is no evidence of significant nonresponse bias among marginally employed individuals. Relative nonresponse bias in the employment variables tends to be larger in the web mode compared to the telephone mode.

Turning to bias due to measurement error, in line with our theoretical expectations, the socially more desirable characteristic of regular employment is significantly overreported in the telephone mode, but does not show significant measurement error bias in the self-administered web mode. However, regular employment is only slightly more accurately estimated in the web mode than in the telephone mode, with the difference in biases not being statistically significant. The potentially more stigmatizing form of marginal employment is significantly underreported in both modes, although to a somewhat lesser extent in the web mode. Like for regular employment, biases do not differ significantly between the modes. Again, we attribute these results to social desirability: telling an interviewer that one has a regular job is more desirable and less of a norm violation than admitting to being "only" marginally employed. Hence, not surprisingly, relative bias due to measurement error is always slightly higher in the telephone mode compared to the web mode although these differences are not significant for any employment type across modes. With the exception of marginal employment in the telephone mode, relative nonresponse and relative measurement bias reinforce each other, leading to an even larger relative combined bias. Despite a marginally smaller measurement error bias, the web mode exhibits a consistently larger combined bias compared to the telephone mode (differences are not statistically significant for marginal employment).

Relative nonresponse bias in *past benefit receipt* is negative for both modes. This leads to a significant underestimation of the proportion of individuals who received welfare in the past year and this underestimation is significantly more pronounced in the web mode than in the telephone mode. Relative measurement error bias points in different directions for both modes such that the proportion of past benefit recipients is overestimated in the web mode and underestimated in the telephone mode (with differences being statistically significant). Surprisingly, the magnitude of measurement bias is strikingly similar across both modes. Relative nonresponse and measurement bias reinforce each other in the telephone mode and point in opposite directions in the web mode. Despite this compensation of both sources of bias in the web mode, relative combined bias is still significantly larger compared to the telephone mode.

Mean *income* is significantly biased due to nonresponse in the web mode, whereas there is no significant relative nonresponse bias in mean income in the telephone mode. The difference in relative nonresponse bias is significant. There is no significant relative measurement bias for mean income in either mode. Relative combined bias is statistically significant for both modes and is mostly driven by nonresponse bias. Whereas there is only a small negative combined bias in the telephone mode, this bias is much larger in the web mode, which results in an overestimation of mean income. We find that relative nonresponse and combined bias differ significantly between the two survey modes.

There is no significant bias due to nonresponse in the telephone mode in the *income categories* except for a slight overestimation of the lower income category. This differs in the web mode: individuals with a low income are significantly underrepresented, whereas those with a high income are significantly overrepresented. Relative nonresponse biases differ significantly between the modes. Although measurement error bias for mean income is statistically nonsignificant, both modes show considerable measurement error bias in the different income categories (Figure 4). While significantly more respondents claim to belong to the low income group (telephone) or the middle income group (web), in both modes too few respondents report that they belong to the highest income category. Measurement error bias does not differ significantly between both modes for any of the income categories. All income categories show significant combined bias for both modes pointing in opposite directions (and being significantly different between the modes) in the lowest and the highest income category. Combined bias is mostly driven by nonresponse bias in the web mode and measurement error bias in the telephone mode.

To summarize our results, the individual contributions of nonresponse and measurement error bias for those variables that show significant relative combined bias indicate that nonresponse bias exceeds measurement error bias in magnitude for gender in both modes, for mean age in both modes (and all categories except for 40–49 years in the web mode and ages older than 60 in the telephone mode), and mean income (as well as low and high income) in the web mode. On the other hand, measurement error bias is larger than nonresponse bias for employment, "marginal" employment in both modes and income (as well as income categories) in the telephone mode. Relative combined bias in regular employment and past unemployment benefit receipt differs in its composition across the modes: nonresponse bias is larger than measurement error bias for both characteristics in the web mode, whereas measurement error bias exceeds nonresponse bias in the telephone mode.

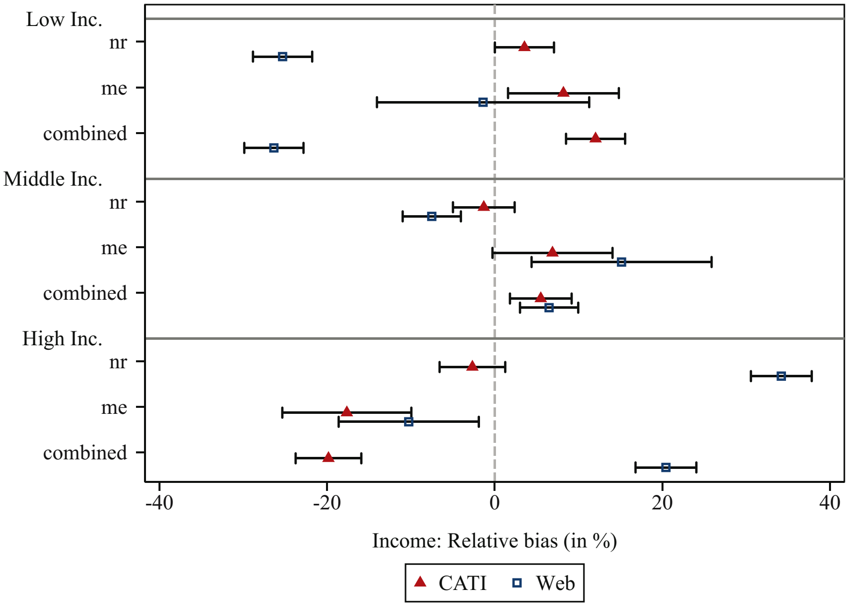

Fig. 4. Relative combined bias for income distribution, including 95%-confidence intervals.

With respect to the *interaction of the two sources of bias,* the results show that bias due to nonresponse and measurement error tend to reinforce each other, with the exception of past benefit receipt and the income categories in the web survey. Our results suggest that the relative combined bias is larger for the web mode compared to the telephone mode for mean age, employment status and employment type, past UB II receipt, mean labor income and the income categories. The combined relative bias is larger in the telephone mode than in the web mode for gender, whereas there is no consistent pattern across age groups. These results suggest that the data obtained via the web survey administration are, overall, more biased compared to the telephone mode.

4. Summary and Discussion

Our results show that the estimates obtained from the web survey are biased to a larger extent compared to the telephone survey when considering combined bias. In line with previous research, these results are mostly driven by a larger nonresponse bias in both modes for demographic items and by larger measurement error bias for sensitive items. Our results further suggest that potential social desirability concerns from respondents are somewhat alleviated in the web mode. However, the potential benefits of a smaller measurement error bias in the web mode are inconsistent across estimates and do not outweigh the comparatively larger nonresponse bias compared to the telephone mode.

The result for overreporting of past welfare benefit, UB II, receipt in the web mode is also somewhat puzzling. One potential explanation for the overreporting of welfare receipt could be slight differences in the question wording between the modes. The telephone survey asked for “welfare receipt in 2010”, that is, the previous year, while the web survey asked for “welfare receipt in the past 12 months”. The data collection period of the web survey was in the beginning of 2012, so we expect that some respondents did not refer to the last 12 months in their retrieval, but instead included the period since January 2011, thus leading to an overestimate. The administrative data can exactly differentiate these differing periods. Also, web survey respondents might have skipped reading the information regarding the reference period altogether (about 7% of those respondents who did not receive UB II in the reference period actually received benefits at some earlier point). Because respondents in the telephone component of the study received a series of filter questions about different earnings in 2011 and the respondents in the web mode only saw this one question, respondents in the web mode might be more prone to suffer from referring to the wrong reference period. Both kinds of error would result in web survey respondents reporting receipt prior to the 12-month reference period – since February 2011 – and thus explain the significant amount of overreporting. Another explanation for the underreporting in the telephone mode could also be due to a strategy to avoid follow-up questions. However, Eckman et al. (2014) find no significant filter effect for the income questions in this survey. These potential errors confound our results with respect to mode differences in UB II receipt. Nonetheless, web seems to outperform telephone for this item, in the sense that it is able to alleviate social desirability concerns.

We find substantial misreporting of income across different income categories, although this does not differ across modes. This is surprising in that we would have expected more accurate reporting in the web mode due to increased privacy and the fact that an individual can take the survey at their own pace, potentially spending more time to retrieve accurate information. In line with results reported in previous studies (e.g., Duncan and Hill 1985; Bound and Krueger 1991; Rodgers et al. 1993), measurement error seems to be correlated with true income; more specifically, that it is mean-reverting, which is a tendency for those with lower earnings to overstate these and those with higher earnings to understate. The tendency to overreport the middle category to the disadvantage of the extreme categories can clearly be seen for the web mode.

Our results are subject to some limitations. First of all, the question arises as to whether nonresponse adjustment techniques would alleviate bias in an identical manner in both modes and how this would affect the combined bias. Since nonresponse bias tends to point in opposite directions for both modes, the most obvious solution might be to pool the samples (De Rada and del Amo 2014). Another option is to rely on different weighting techniques (Bethlehem 2010). However, such techniques are only reducing bias if weighting variables are correlated with both nonresponse and survey variables of interest (Kreuter and Olson 2011). Studying nonresponse bias before adjustment is a topic in its own right, as Schouten et al. (2016) conclude that balanced samples are always advantageous, regardless of adjustment techniques that might be applied in retrospect. However, if the same mechanisms that lead to nonresponse bias are related to measurement error bias (Olson 2013; Malhotra et al. 2014; Roberts et al. 2014), combined bias might actually be inflated. The comparison of both modes after these adjustments are particularly interesting, that is, whether bias estimates are affected in a similar manner. While that is an interesting research question, these analyses are beyond the scope of this article. The second limitation is that we are comparing respondents across mode "packages" and cannot directly attribute measurement differences to mode effects. For example, differences in responses might not be causally attributed to different reporting schemes evoked by different modes, but can also be driven by different individuals (with differential reporting behavior) responding to different modes (and hence be due to sample composition). Third, data collection periods in both modes differed slightly and there is a potential time effect that we cannot rule out. However, with the exception of one characteristic – past receipt of welfare – we are confident that this does not jeopardize our results.

To reiterate, our results are in line with previous research: while younger, employed and more educated individuals participate in the web survey, there is less bias in the telephone mode and nonresponse bias tends to point in opposite directions in both modes. At the same time, measurement error bias tends to be equivalent or smaller in the web mode. Given that the web mode has several advantages over the telephone mode with respect to survey costs and immediate data availability, one implication that follows from our results could be to implement a sequential mixed mode design, especially since nonresponse biases in both modes tend to be in opposing directions. Thus, approaching respondents first by web and then following up on nonrespondents by the telephone seems to be a promising approach to reach different subgroups in the population and balance the respondent sample. Another promising strategy to reduce measurement error bias in the telephone could be to supplement the telephone component with a self-administration mode, either using IVR, T-ACASI or a web add-on, each with its own advantages and disadvantages.

5. Appendix

5.1. Appendix A – Biases

Table 2. Relative nonresponse bias, relative measurement bias and relative combined bias including 95% confidence intervals. Significant biases, based on one-sample Z-tests are indicated in boldface (p0.05). Z-values are reported for two-sample Z-Test of differences between the telephone and web mode.

	Rel. nonresponse bias			Rel. measurement error bias			Rel. combined bias		
Variable	CATI	Web	z-value	CATI	Web	z-value	CATI	Web	z-value
Female	**9.08** (6.37; 11.80)	**3.89** (1.38; 6.41)	**− 2.75**	− 0.48 (− 4.27; 3.32)	− 0.89 (− 6.65; 4.86)	− 0.12	**8.57** (6.46; 10.67)	**2.96** (1.03;4.9)	**− 3.85**
Mean age (years)	**1.72** (0.91; 2.54)	**− 4.97** (− 5.72; − 4.22)	**− 11.82**	0.02 (− 1.20; 1.23)	0.05 (− 1.69; 1.79)	0.03	**1.74** (1.11; 2.37)	**− 4.92** (− 5.50; − 4.34)	**− 15.24**
Age ≤ 29	− 4.42 (− 9.84; 1.01)	**7.12** (1.51; 12.72)	**2.90**	− 0.23 (− 8.76; 8.30)	0.00 (− 12.79; 12.79)	0.03	**− 4.64** (− 8.84; − 0.43)	**7.12** (2.81; 11.42)	**3.83**
Age 30–39	**− 13.38** (− 18.23; − 8.53)	**23.55** (18.88; 28.22)	**10.76**	− 0.43 (− 8.55; 7.69)	− 0.33 (− 9.98; 9.31)	0.01	**− 13.75** (− 17.51; − 9.99)	**23.14** (19.55; 26.72)	**13.92**
Age 40–49	4.16 (− 0.56; 8.88)	0.09 (− 4.20; 4.39)	− 1.66	0.85 (− 6.18; 7.88)	1.09 (− 9.12; 11.3)	0.04	**5.05** (1.38; 8.71)	1.18 (− 2.11; 4.48)	− 1.54
Age 50–59	**11.96** (6.75; 17.16)	**− 6.83** (− 11.81; − 1.86)	**− 5.12**	− 0.55 (− 7.96; 6.86)	− 0.98 (− 13.35; 11.39)	− 0.06	**11.34** (7.30; 15.37)	**− 7.75** (− 11.57; − 3.93)	**− 6.73**
Age ≥ 60	3.11 (− 3.14; 9.36)	**− 33.17** (− 39.06; − 27.29)	**− 8.28**	0.27 (− 9.12; 9.66)	0.00 (− 17.63; 17.63)	− 0.03	3.39 (− 1.46; 8.23)	**− 33.17** (− 37.69; − 28.65)	**− 10.82**
Employed	**4.74** (2.61; 6.88)	**12.79** (10.70; 14.88)	**5.29**	**9.33** (6.26; 12.40)	**14.03** (9.80; 18.25)	1.76	**14.46** (12.80; 16.11)	**28.58** (26.97; 30.18)	**12.02**
Regular employment	**5.81** (3.03; 8.59)	**17.33** (14.68; 19.98)	**5.87**	**6.37** (2.35; 10.40)	4.34 (− 1.01; 9.69)	− 0.59	**12.55** (10.40; 14.71)	**22.42** (20.38; 24.45)	**6.52**
Marginal employment	1.94 (− 5.92; 9.80)	− 7.19 (− 14.89; 0.52)	− 1.63	**− 29.51** (− 41.40; − 17.61)	**− 25** (− 44.10; − 5.90)	0.39	**− 28.14** (− 34.24; − 22.05)	**− 30.39** (− 36.31; − 24.47)	− 0.52
Mean income (EUR)	− 0.26 (− 2.22; 1.70)	**18.53** (16.73; 20.33)	**13.85**	− 2.13 (− 5.91; 1.64)	− 1.67 (− 6.38; 3.04)	0.15	**− 2.39** (− 4.34; − 0.43)	**16.55** (14.75; 18.35)	**13.96**
Low income	**3.54** (0.02; 7.07)	**− 25.31** (− 28.84; − 21.77)	**− 11.33**	**8.20** (1.59; 14.81)	− 1.39 (− 14.05; 11.27)	− 1.31	**12.03** (8.51; 15.55)	**− 26.34** (− 29.87; − 22.81)	**− 15.07**
Middle income	− 1.30 (− 4.98; 2.38)	**− 7.51** (− 10.98; − 4.04)	**− 2.40**	6.89 (− 0.27; 14.05)	**15.13** (4.39; 25.87)	1.25	**5.50** (1.82; 9.18)	**6.49** (3.02; 9.96)	0.38
High income	− 2.66 (− 6.58; 1.25)	**34.19** (30.56; 37.81)	**13.53**	**− 17.64** (− 25.34; − 9.94)	**− 10.26** (− 18.62; − 1.9)	1.27	**− 19.83** (− 23.75; − 15.92)	**20.42** (16.80; 24.05)	**14.78**
Past receipt of UB II	**− 5.52** (− 9.13; − 1.90)	**− 37.92** (− 42.09; − 33.75)	**− 11.51**	**− 14.23** (− 20.00; − 8.46)	**19.10** (5.67; 32.54)	**4.47**	**− 18.96** (− 21.76; − 16.16)	**− 26.07** (− 29.27; − 22.86)	**− 3.27**

5.2. Appendix B – Biases when Using All Sample Cases

In a sensitivity analysis, we replicated the bias estimation, including all sample cases, for example, all individuals assigned to the telephone mode, including those without valid telephone numbers and all individuals assigned to the web mode, including those whose invitation letter was returned to sender (see Table 3). In this second analysis, bias due to deployability and coverage cannot be separated from nonresponse bias. For simplicity, we will continue to refer to this as nonresponse bias. While relative nonresponse and relative combined biases might be affected, relative measurement error biases stay the same as they only refer to the survey respondents.

Although relative nonresponse biases change in magnitude, the relative difference in a comparison of the survey modes does not change for any of the variables compared to the analysis excluding the individuals for whom we do not have valid contact information. We do find some significant differences: when including all cases, relative nonresponse bias for *mean income* is now significantly different from zero for both modes as opposed to the web mode only. Relative nonresponse bias in *29 years and younger* is not significantly different from zero in any mode and modes do not significantly differ from each other when including individuals without valid contact information, whereas relative nonresponse bias for this variable is significantly different from zero in the web and biases significantly differ between the modes when excluding individuals without valid contact information. Also, relative nonresponse bias in *ages 50–59 years* is not significant in the web mode when including individuals without valid contact information. The age group *60 years and older* shows significant relative nonresponse biases in both modes, with significantly different relative nonresponse biases between the modes when including individuals without valid contact information, whereas the relative nonresponse bias for the telephone survey loses significance when excluding those individuals. This results in individuals aged *29 years and younger* being overrepresented when including all cases, but underrepresented when excluding individuals without valid contact information for the telephone survey, although relative nonresponse bias is not significant for this age group in any analysis of the telephone survey. Strikingly, the negative effect of past welfare receipt turns from a negative to a larger positive effect in the telephone survey when including the individuals without valid contact information, with differences between the modes being significant in both kinds of analysis.

As for relative nonresponse bias, the magnitudes of the combined bias differ slightly for the two kinds of analyses, but the directionality and relative differences comparing the survey modes are not affected for most of the variables. Differences in the relative magnitude of combined bias can only be found for *age 29 years and younger* and *middle income*. The differences in relative combined bias between the survey modes are not significant for *middle income* when the individuals without valid contact information are excluded, but are significant if they are included. However, the relative bias in middle income is significantly different from zero for both modes in both kinds of analyses. For *age 29 years and younger* we find the difference in relative combined bias between the two modes to be significant when excluding the individuals without valid contact information, but to be not significant if including these individuals. This is mostly due to a change in directionality for the telephone mode and a shift towards zero for the web mode. The

Table 3. Relative nonresponse bias, relative measurement bias and relative combined bias, including 95% confidence intervals. Significant biases based on one-sample Z-tests are indicated in boldface ($p < 0.05$). Z-values are reported for two-sample Z-Tests of differences between the telephone and web mode.

	Rel. nonresponse bias			Rel. measurement error bias			Rel. combined bias		
Variable	CATI	Web	z-value	CATI	Web	z-value	CATI	Web	z-value
Female	**8.07** (5.77; 10.38)	**4.37** (1.96; 6.78)	**− 2.18**	−0.48 (−4.27; 3.32)	−0.89 (−6.65; 4.86)	−0.12	**7.56** (5.75; 9.37)	**3.44** (1.61; 5.26)	**− 3.14**
Mean age (years)	**1.27** (0.58; 1.96)	**− 4.07** (−4.79; −3.34)	**− 10.46**	0.02 (−1.20; 1.23)	0.05 (−1.69; 1.79)	0.03	**1.29** (0.75; 1.83)	**− 4.02** (−4.57; −3.47)	**− 13.53**
Age ≤ 29	4.91 (−0.02; 9.84)	2.11 (−3.09; 7.31)	−0.76	−0.23 (−8.76; 8.30)	0.00 (−12.79; 12.79)	0.03	**4.67** (0.80; 8.53)	2.11 (−1.82; 6.05)	−0.91
Age 30–39	**− 16.51** (−20.57; −12.44)	**18.72** (14.38; 23.06)	**11.60**	−0.43 (−8.55; 7.69)	−0.33 (−9.98; 9.31)	0.01	**− 16.86** (−20.05; −13.68)	**18.32** (15.03; 21.61)	**15.07**
Age 40–49	−1.60 (−5.51; 2.30)	2.14 (−2.01; 6.30)	1.29	0.85 (−6.18; 7.88)	1.09 (−9.12; 11.30)	0.04	−0.77 (−3.83; 2.29)	**3.26** (0.11; 6.41)	1.80
Age 50–59	**12.9** (8.40; 17.39)	−3.49 (−8.34; 1.37)	**− 4.85**	−0.55 (−7.96; 6.86)	−0.98 (−13.35; 11.39)	−0.06	**12.27** (8.75; 15.79)	**− 4.43** (−8.11; −0.76)	**− 6.43**
Age ≥ 60	**5.77** (0.32; 11.22)	**− 30.86** (−36.59; −25.13)	**− 9.08**	0.27 (−9.12; 9.66)	0.00 (−17.63; 17.63)	−0.03	**6.05** (1.79; 10.32)	**− 30.86** (−35.20; −26.52)	**− 11.89**
Employed	**2.67** (0.88; 4.45)	**16.10** (14.03; 18.16)	**9.65**	**9.33** (6.26; 12.40)	**14.03** (9.80; 18.25)	1.76	**12.18** (10.79; 13.58)	**32.33** (30.77; 33.89)	**18.84**
Regular employment	**2.51** (0.19; 4.83)	**20.55** (17.95; 23.15)	**10.16**	**6.37** (2.35; 10.40)	4.34 (−1.01; 9.69)	−0.59	**9.04** (7.23; 10.85)	**25.78** (23.81; 27.75)	**12.26**
Marginal employment	3.02 (−3.77; 9.81)	−4.27 (−11.75; 3.22)	−1.41	**− 29.51** (−41.4; −17.61)	**− 25.00** (−44.10; −5.90)	0.39	**− 27.38** (−32.70; −22.06)	**− 28.20** (−33.87; −22.53)	−0.21
Mean income (EUR)	**− 6.55** (−8.19; −4.91)	**18.91** (17.18; 20.63)	**20.97**	−2.13 (−5.91; 1.64)	−1.67 (−6.38; 3.04)	0.15	**− 8.54** (−10.18; −6.90)	**16.92** (15.20; 18.64)	**20.96**
Low income	**11.24** (8.04; 14.44)	**− 25.38** (−28.75; −22.01)	**− 15.44**	**8.2** (1.59; 14.81)	−1.39 (−14.05; 11.27)	−1.32	**20.36** (17.16; 23.55)	**− 26.41** (−29.79; −23.04)	**− 19.73**
Middle income	**3.36** (0.09; 6.63)	**− 8.88** (−12.16; −5.60)	**− 5.18**	6.89 (−0.27; 14.05)	**15.13** (4.39; 25.87)	1.25	**10.48** (7.21; 13.75)	**4.90** (1.63; 8.18)	**− 2.36**
High income	**− 13.76** (−16.83; −10.69)	**36.51** (33.00; 40.02)	**21.12**	**− 17.64** (−25.34; −9.94)	**− 10.26** (−18.62; −1.90)	1.27	**− 28.97** (−32.05; −25.90)	**22.51** (19.00; 26.02)	**21.63**
Past receipt of UB II	**9.21** (5.76; 12.66)	**− 38.19** (−42.17; −34.22)	**− 17.65**	**− 14.23** (−20.00; −8.46)	**19.10** (5.67; 32.54)	**4.47**	**− 6.33** (−9.03; −3.63)	**− 26.39** (−29.39; −23.38)	**− 9.72**

effects of *age 40–49 years* in the web and *60 years and older* in the telephone mode increase and are significant in this second analysis. For *age 40–49 years* the relative combined bias in the telephone survey is less pronounced and not significant when including the individuals without valid contact information.

Even though we find some differences for relative nonresponse bias and relative combined bias between the two analyses for some age and income categories, these differences do not substantively change our findings and do not affect relative combined bias in mean age or mean income. The only substantive difference between the two sets of analysis is the change from underrepresentation to overrepresentation of past recipients of welfare benefit in the telephone mode when including the individuals without valid contact information. From this, we can conclude that more valid telephone numbers have been available for past benefit recipients than for nonrecipients. This makes sense, as individuals on UB II have to provide the German Federal Employment Agency with their telephone numbers to manage benefit claims. Even though this affects the relative combined bias in our survey, we do not expect this to be a general finding as this is very specific to the sample drawn using the data from the German Federal Employment Agency.

6. References

AAPOR, The American Association for Public Opinion Research. 2011. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition.

Abraham, K.G., A. Maitland, and S.M. Bianchi. 2006. "Nonresponse in the American Time Use Survey. Who is Missing from the Data and How Much Does it Matter?" *Public Opinion Quarterly* 70: 676–703. Doi: http://dx.doi.org/10.1093/poq/nfl037.

Atkeson, L.R., A.N. Adams, and M.R. Alvarez. 2014. "Nonresponse and Mode Effects in Self- and Interviewer-Administered Surveys." *Political Analysis* 22: 304–320. Doi: http://dx.doi.org/10.1093/pan/mpt049.

Atkeson, L.R., A.N. Adams, L.A. Bryant, L. Zilberman, and K.L. Saunders. 2011. "Considering Mixed Mode Surveys for Questions in Political Behavior: Using the Internet and Mail to Get Quality Data at Reasonable Costs." *Political Behavior* 33: 161–178. Doi: http://dx.doi.org/10.1007/s11109-010-9121-1.

Bethlehem, J. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78: 161–188. Doi: http://dx.doi.org/10.1111/j.1751-5823.2010.00112.x.

Biemer, P.P. 2010. "Overview of Design Issues: Total Survey Error." In *Handbook of Survey Research*, edited by P.V. Marsden and J.D. Wright, 27–57. Bingley: Emerald.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley.

Bound, J. and A.B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9: 1–24. Doi: http://dx.doi.org/10.3386/w2885.

Bradburn, N., S. Sudman, and B. Wansink. 2004. *Asking Questions. Revised Edition*. San Francisco: Jossey-Bass.

Braunsberger, K., H. Wybenga, and R. Gates. 2007. "A Comparison of Reliability Between Telephone and Web-Based Surveys." *Journal of Business Research* 60: 758–764. Doi: http://dx.doi.org/10.1016/j.jbusres.2007.02.015.

Callegaro, M., R.P. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick, and P.J. Lavrakas. 2014. *Online Panel Research. A Data Quality Perspective*. Chichester: Wiley.

Chang, L. and J.A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet. Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73: 641–678. Doi: http://dx.doi.org/10.1093/poq/nfp075.

Chang, L. and J.A. Krosnick. 2010. "Comparing Oral Interviewing With Self-Administered Computerized Questions: An Experiment." *Public Opinion Quarterly* 74: 154–167. Doi: http://dx.doi.org/10.1093/poq/nfp090.

De Leeuw, E.D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21: 233–255.

De Leeuw, E.D., D.A. Dillman, and J.J. Hox. 2008. "Mixed-Mode Surveys: When and Why." In *International Handbook of Survey Methodology*, edited by E.D. de Leeuw, J.J. Hox, and D.A. Dillman, 299–316. New York: Erlbaum/Taylor & Francis.

De Rada, V.D. and S.P. del Amo. 2014. "Two Are Better Than One: The Use of a Mixed-Mode Data Collection to Improve the Electoral Forecast." *Survey Practice* 7: 1–6. Doi: http://dx.doi.org/10.29115/SP-2014-0003.

Dillman, D.A., J.L. Eltinge, R.M. Groves, and R.J.A. Little. 2002. "Survey Nonresponse in Design, Data Collection and Analysis." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 3–26. New York: Wiley.

Dillman, D.A., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B.L. Messer. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR) and the Internet." *Social Science Research* 38: 1–18. Doi: http://dx.doi.org/10.1016/j.ssresearch.2008.03.007.

Duffy, B., K. Smith, G. Terhanian, and J. Bremer. 2005. "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research* 47: 615–639. Doi: http://doi.org/10.1177/147078530504700602.

Duncan, G. and D. Hill. 1985. "An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data." *Journal of Labor Economics* 3: 508–532. Doi: http://dx.doi.org/10.1086/298067.

Eckman, S., F. Kreuter, A. Kirchner, A. Jäckle, S. Presser, and R. Tourangeau. 2014. "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys." *Public Opinion Quarterly* 78: 721–733. Doi: http://dx.doi.org/10.1093/poq/nfu030.

Fricker, S., M. Galesic, R. Tourangeau, and T. Yan. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 6: 370–392. Doi: http://dx.doi.org/10.1093/poq/nfi027.

Groves, R.M. 2004. *Survey Error and Survey Costs*. Hoboken: Wiley & Sons.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. Doi: http://dx.doi.org/10.1093/poq/nfl033.

Groves, R.M. and M. Couper. 1998. Nonresponse in Household Interview Surveys. Wiley Series in Probability and Statistics: Survey Methodology Section. New York: Wiley.

Hope, S., P. Campanelli, G. Nicolaas, P. Lynn, and A. Jäckle. 2014. "The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration." *ISER Working Paper Series* No. 2014-20: 1–41. Available at: http://hdl.handle.net/10419/123808 (accessed December 2014).

IAB (Institut für Arbeitsmarkt- und Berufsforschung). 2011. Nuremberg: Integrierte Erwerbsbiographien (IEB) V09.00.

IAB (Institut für Arbeitsmarkt- und Berufsforschung). 2012. Nuremberg: Leistungshistorik Grundsicherung (LHG), Version 06.06.

IAB (Institut für Arbeitsmarkt- und Berufsforschung). 2013. Nuremberg: Beschäftigtenhistorik (BeH), Version 09.03.00.

Jacobebbinghaus, P. and S. Seth. 2007. "The German Integrated Employment Biographies Sample IEBS." *Schmollers Jahrbuch* 127: 335–342.

Kreuter, F. and K. Olson. 2011. "Multiple Auxiliary Variables in Nonresponse Adjustment." *Sociological Methods & Research* 40: 311–332. Doi: http://dx.doi.org/10.1177/0049124111400042.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T. Ezatti-Rice, C. Casas-Cordero, A. Petychev, R. M. Groves, and T. Raghuatan. 2010. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173: 389–407. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00621.x.

Kreuter, F., S. Presser, and R. Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys. The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly* 72: 847–865. Doi: http://dx.doi.org/10.1093/poq/nfn063.

Lee, R.M. 1993. *Doing Research on Sensitive Topics*. London: Sage.

Letourneau, P.M. and A.A. Zbikowski. 2008. "Nonresponse in the American Time Use Survey." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 4, 2008. 1283–1290. Denver, CO: American Statistical Association. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2008/Files/300982.pdf (accessed April 2018).

Lozar Manfreda, K., M. Bosnjak, J. Berzelak, I. Haas, and V. Vehovar. 2008. "Web Surveys Versus Other Survey Modes. A Meta-Analysis Comparing Response Rates." *International Journal of Market Research* 50: 79–104. Doi: http://dx.doi.org/10.1177/147078530805000107.

Malhotra, N., J.M. Miller, and J. Wedeking. 2014. "The Relationship Between Nonresponse Strategies and Measurement Error. Comparing Online Panels to Traditional Surveys." In *Online Panel Research. A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A.S. Göritz, J. Krosnick, and P.J. Lavrakas, 313–336. Chichester: Wiley.

McCabe, S.E., C.J. Boyd, M.P. Couper, S. Crawford, and H. D'Arcy. 2002. "Mode Effects for Collecting Alcohol and Other Drug Use Data: Web and U.S. Mail." *Journal of Studies on Alcohol* 63: 755–761. Doi: http://dx.doi.org/10.15288/jsa.2002.63.755.

Olson, K. 2013. "Do Non-Response Follow-Ups Improve or Reduce Data Quality? A Review of the Existing Literature." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 176: 129–145. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01042.x.

O'Neill, G. and J. Dixon. 2005. "Nonresponse Bias in the American Time Use Survey." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 10, 2005. 2958–2966. Minneapolis, MN: American Statistical

Association. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2005/Files/JSM2005-000193.pdf (accessed April 2018).

Roberts, C., N. Allum, and P. Sturgis. 2014. "Nonresponse and Measurement Error in an Online Panel. Does Additional Effort to Recruit Reluctant Respondents Result in Poorer Data Quality?" In *Online Panel Research. A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A.S. Göritz, J. Krosnick, and P.J. Lavrakas, 337–362. Chichester: Wiley.

Rodgers, W.L., C. Brown, and G.J. Duncan. 1993. "Errors in Survey Reports of Earnings, Hours Worked, and Hourly Wages." *Journal of the American Statistical Association* 88: 1208–1218. Doi: http://dx.doi.org/10.1080/01621459.1993.10476400.

Sakshaug, J.W. and F. Kreuter. 2012. "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data." *Survey Research Methods* 6: 113–122. Doi: http://dx.doi.org/10.18148/srm/2012.v6i2.5094.

Sakshaug, J.W., T. Yan, and R. Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items." *Public Opinion Quarterly* 74: 907–933. Doi: http://dx.doi.org/10.1093/poq/nfq057.

Sanders, D., H.D. Clarke, M.C. Stewart, and P. Whiteley. 2007. "Does Mode Matter for Modelling Political Choice? Evidence from the 2005 British Election Study." *Political Analysis* 15: 257–285. Doi: http://dx.doi.org/10.1093/pan/mpl010.

Sax, L.J., S.K. Gilmartin, and A.N. Bryant. 2003. "Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys." *Research in Higher Education* 44: 409–432. Doi: http://dx.doi.org/10.1023/A:1024232915870.

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does More Balanced Survey Response Imply Less Non-Response Bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179: 727–748. Doi: http://dx.doi.org/10.1111/rssa.12152.

Statistisches Bundesamt. 2013. *Wirtschaftsrechnungen. Private Haushalte in der Informationsgesellschaft – Nutzung von Informations – und Kommunikationstechnologien*. Wiesbaden, Germany: Statistisches Bundesamt.

Stephenson, L.B. and J. Crête. 2011. "Studying Political Behavior: A Comparison of Internet and Telephone Surveys." *International Journal of Public Opinion Research* 23: 24–55. Doi: http://dx.doi.org/10.1093/ijpor/edq025.

Vannieuwenhuyze, J., G. Loosveldt, and G. Molenberghs. 2010. "A Method for Evaluating Mode Effects in Mixed-Mode Surveys." *Public Opinion Quarterly* 74: 1027–1045. Doi: http://dx.doi.org/10.1093/poq/nfq059.

Yeager, D.S., J.A. Krosnick, L. Chang, H.S. Javitz, M.S. Levendusky, A. Simpser, and R. Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75: 709–747. Doi: http://dx.doi.org/10.1093/poq/nfr020.

Received April 2016
Revised May 2018
Accepted May 2018