Misreporting Among Reluctant Respondents

Ruben L. Bach[*] Stephanie Eckman[†] Jessica Daikeler[‡]

Word Count: 6,346 (excluding tables, figures, appendix, reference list)

Acknowledgements: The authors thank Frauke Kreuter, members of FK²RG at the University of Mannheim, Brady West, and the anonymous reviewers for their helpful feedback. The LISS panel data was collected by CentERdata (Tilburg University, The Netherlands) through its MESS project funded by the Netherlands Organization for Scientific Research.

Published in: *Journal of Survey Statistics and Methodology*, 8(3): 566-588

https://doi.org/10.1093/jssam/smz013

[*]University of Mannheim, Germany
[†]RTI International, Washington, D.C.
[‡]GESIS–Leibniz Institute for the Social Sciences, Mannheim, Germany

Abstract

Many surveys aim to achieve high response rates to keep bias due to nonresponse low. However, research has shown that the relationship between the nonresponse rate and nonresponse bias is small. In fact, high response rates may lead to measurement error, if respondents with low response propensities provide survey responses of low quality. In this paper, we explore the relationship between response propensity and measurement error, specifically motivated misreporting, the tendency to give inaccurate answers to speed through an interview. Using data from four surveys conducted in several countries and modes, we analyze whether motivated misreporting is worse among those respondents who were the least likely to respond to the survey. Contrary to the prediction of our theoretical model, we find only limited evidence that reluctant respondents are more likely to misreport.

1 Background

Many surveys aim to achieve high response rates to keep bias due to nonresponse low, but increasing the response rate by bringing in reluctant respondents may lead to measurement error. That is, respondents who are the least likely to become respondents may provide survey responses of low quality when they do respond (Curtin et al., 2000, 2005; Groves et al., 2004; Groves, 2006; Groves and Peytcheva, 2008; Keeter et al., 2000; Merkle and Edelman, 2002; Tourangeau et al., 2010; Peytchev et al., 2010; Olson, 2013). Thus, researchers who use extraordinary measures to increase the response rate may in fact increase total error (Biemer, 2001; Groves, 2006).

We study this relationship between respondents' reluctance and measurement error in this paper. To do so, we must operationalize both reluctance and measurement error. We estimate response propensities, i.e. the probability of each person who was selected for a survey to respond to the survey to measure respondents' reluctance. Respondents with the lowest response propensities are reluctant respondents. We operationalize measurement error through motivated misreporting, a phenomenon whereby respondents deliberately give inaccurate or false responses to reduce the burden of the survey. This response behavior is often observed in questions used to determine respondent eligibility for follow-up questions. Asking such questions in certain formats allows respondents to learn how follow-up questions can be avoided by giving inaccurate or false answers, thus introducing measurement error (Tourangeau et al., 2015). The motive behind this motivated misreporting is respondents' desire to reduce the burden of the survey (Eckman et al., 2014). Respondents who have a low propensity to respond to the survey at all may be more interested than other respondents in reducing the burden of the survey when they do respond. Thus, reluctant respondents should show more motivated misreporting, supporting the hypothesis that response propensity affects measurement error. We elaborate on these operational definitions and the hypothesis in more detail in the next section.

To study this hypothesis empirically, we use four surveys that were conducted in three countries (the Netherlands, the U.S. and Germany) and in three modes (Web, CAPI and CATI). Each contained experimental manipulations of filter questions, a type of eligibility questions that are prone to motivated misreporting. These experimental manipulations allow us to study the connection between response propensity and measurement error. Before we review the data in more detail, we present the theoretical reasoning underlying the hypothesis that nonresponse influences measurement error.

2 A Nonresponse-Measurement Error Model

The idea that reluctant respondents may be worse reporters builds on the nonresponse-measurement error model developed by Groves (2006), shown in Figure 1. This model suggests a nexus between response propensity and measurement error. Let Y denote the reported value of some true value Y^*. Y equals then Y^* plus an error term ε: For each individual i, $Y_i = Y_i^* + \varepsilon_i$. The magnitude of the error for case i, ε, is determined by the response propensity of the case RP_i, introducing a covariance between Y and RP. Respondents with a high response propensity, for example, may be more inclined to giving accurate answers in a survey, thereby introducing this covariance.

In terms of the discussion above, motivated misreporting results in negative values of ε because respondents under-report the true value (i.e., the reported value Y is smaller than the true value Y^*). Larger *absolute* values of ε thus indicate more motivated misreporting and our model predicts that low response propensities lead to more misreporting (high $|\varepsilon|$), while high response propensities cause lower levels of motivated misreporting (small $|\varepsilon|$).

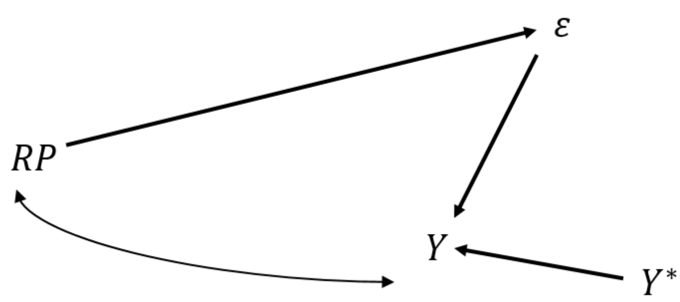

Figure 1: Nonresponse-Measurement Error model explaining a relationship between response propensity (RP) and measurement error (ε) in a reported survey variable Y (figure adapted from Groves (2006)).

This model does not specify how exactly response propensity influences the error term. Many possible mechanisms exist. For example, lack of interest in the survey topic can cause a case to have a low response propensity (Martin, 1994; Groves et al., 2004) and may also explain why low interest respondents who *do* participate in the survey put less effort into answering survey questions carefully and truthfully. Other motives such as a general reluctance to help out (Tourangeau et al., 2010) or a lack of motivation and cooperativeness (Cannell and Fowler, 1963; Bollinger and David, 2001) may also reduce RP and introduce more measurement error. In terms of motivated misreporting, the desire to reduce the burden of the survey may result in low response propensities because respondents are reluctant to participate in the survey in the first place. When they do participate, low response propensities result in large errors because respondents skip follow-up questions to keep the survey short. Thus, there may be some characteristics Z that explain both RP and ε and induce the relationship between the two shown in Figure 1. These external causes are excluded from the model. Nevertheless, we can use this model to test our hypothesis about the relationship between response propensity and motivated misreporting.

3 Previous findings

Two lines of literature are relevant for this study. The first one includes those studies that analyze the connection between (non)response propensity and measurement error. The second one concerns studies of motivated misreporting, one form of measurement error. We review studies that fall within these two strands of literature below.

3.1 Findings on the nexus between response propensity and measurement error

Empirical studies of a connection between (non)response propensity and measurement error have focused on several aspects of both response propensity and measurement error. Cannell and Fowler (1963), for example, assess the impact of nonresponse on errors in self-reported hospital stays. Comparing self-reports with administrative hospital records, they find that respondents who needed extensive follow-up, i.e., respondents who are nearly nonrespondents, tend to misreport both the number of hospital stays and their duration. However, it is unclear if the higher level of measurement error among late respondents is caused by response propensity or simply the result of the increased recall period for late respondents (Fricker and Tourangeau, 2010). Kreuter et al. (2010), also validating survey reports against administrative records, find that measurement error among respondents recruited with increased levels of follow-up offsets the reduction in nonresponse bias gained by including them. In other words, they find that nonresponse bias is reduced when additional, hard-to-recruit respondents are included. At the same time, however, measurement error among those respondents is high, leading to a net increase in total error when these respondents are included.

Other studies show that reluctant respondents, defined as late respondents (Willimack et al., 1995) and converted refusers (Triplett et al., 1996) have higher item nonresponse rates. Little evidence, however, is found by Keeter et al. (2000) regarding the effects of more rigorous recruiting strategies compared to standard recruiting strategies on item nonresponse. Studies using response propensity scores find that including low response propensity cases results only in a weak increase in measurement error (measured as the differences between self-reports of marriage duration/frequency and administrative records) that is offset by gains in reduction of nonresponse bias (Olson, 2006). Furthermore, low response propensity cases underreport abortion experiences (Peytchev et al., 2010), show more misreporting errors in voting behavior (Tourangeau et al., 2010) and higher item nonresponse rates (Fricker and Tourangeau, 2010). Low response propensity cases, however, do not show more acquiescence, extreme responses or non-differentiation (Yan et al., 2004) or provide answers of worse data quality to questions asking for well-being (Hox et al., 2012). To sum up, the majority of previous research examining the influence of nonresponse on measurement error finds that response reluctance, measured through various operationalizations (see review above) does affect measurement error. However, there are some studies where this effect is small or even nonexistent. The different operationalizations of respondents' reluctance and measurement error may explain some of the variation of the findings.

In this study, we use the estimated response propensity scores as the operationalization of respondents' reluctance. The advantage of this approach is that the estimated response propensity score is a comprehensive measure of different aspects of response propensity. That is, if the model is robust, the estimated response propensity scores should capture a variety of aspects of reluctance, such as the extent of follow-up needed (Cannell and Fowler, 1963), how early or late a case responded to the survey (Willimack et al., 1995) and interest in the survey (Martin, 1994). For these reasons, we prefer the response propensity score to the more specific measures of reluctance used in other studies. Regarding the operationalization of measurement error, we study motivated misreporting, which we explain in more detail below.

3.2 Findings on motivated misreporting

Three question types, filter questions, looping questions (Eckman and Kreuter, 2018), and screener questions, are prone to motivated misreporting, a response behavior causing measurement error. These questions are typically used to determine respondents' eligibility for follow-up questions. Filter questions, used for example in the National Survey on Drug Use and Health, the U.S. Consumer Expenditure Survey or the National Crime Victimization Survey, are usually asked either in the interleafed or in the grouped format. Respondents in the *interleafed* format are asked a filter question with the follow-ups, if triggered, right away. In the *grouped* format, however, respondents are first asked all filter questions before answering the follow-ups that apply (see Table 1 for an example).

Table 1: Example of Interleafed vs. Grouped format (filter questions)

Interleafed version	**Grouped version**
Have you ever held a full-time job? From when and until when did you hold your most recent full-time job? How many hours per week did/do you work in your most recent full-time job? In what industry was/is your most recent full-time job? Have you ever held a part-time job? From when and until when did you hold your most recent part-time job? *[...]* Have you ever been self-employed? *[...]* *[...]*	Have you ever held a full-time job? Have you ever held a part-time job? Have you ever been self-employed? *[...]* FOLLOW-UPS FOR EACH YES From when and until when did you hold your most recent (item)? How many hours per week did/do you work in your most recent (item)? In what industry was/is your most recent (item)?

Comparisons between the interleafed and the grouped format in filter questions have shown that respondents trigger fewer follow-ups in the interleafed format than in the grouped format (Kessler et al., 1998; Duan et al., 2007; Kreuter et al., 2011; Eckman et al., 2014). This motivated misreporting is not possible in the grouped format because there is no chance for respondents to learn how the questions work. Similar effects are observed for different formats of looping questions and screener questions (e.g., Eckman and Kreuter (2018); Tourangeau et al. (2012)), however, we do not review them in more detail as they are not included in our analysis.

Regarding the mechanisms that could explain the observed format effect, Eckman et al. (2014) have shown that motivated misreporting arises from respondents' desire to reduce the burden of the survey. This desire to reduce the burden of the survey may also affect the response propensity score. For example, respondents who want to reduce burden may be unlikely to respond to the survey at all. Thus, this desire would be a mechanism that affects both motivated misreporting, ε in Figure 1, and the response propensity score, RP. The level of measurement error associated with a reported survey outcome would therefore be related to the response propensity score, inducing the relationship shown in Figure 1.

Given this theoretical model and the evidence from previous studies, we analyze the connection between response propensity and motivated misreporting. That is, we study whether measurement error in the form of motivated misreporting is more pronounced among reluctant respondents, using several experimental surveys briefly described in the next section.

4 Data

Data for our analysis come from three surveys, conducted in different countries and modes. We briefly present key characteristics of each survey below and in Table 2.

The questions from each survey are shown in the Online Supplementary Materials.

The first survey was conducted as part of the Dutch LISS panel, a longstanding probability-based internet panel. Sample members complete online questionnaires of about 15 to 30 minutes on a monthly basis (Scherpenzeel, 2011). In 2012, we put several filter question experiments in two consecutive waves of the LISS panel using the same questionnaire in both waves. In the first wave (April), LISS participants (n=5,513) were randomly assigned to either the interleafed or grouped filter question format. In the second wave (May), participants (n=5,668) were again randomly assigned to one of the two formats. Respondents in both formats were asked 13 filter questions with two follow-up questions for each filter answered with "Yes". All filter questions asked about purchases of items such as groceries, clothes or movie tickets during the last month. About 68 percent (n=3,767) of the LISS panel members selected for the study participated in the first wave of the experiment (AAPOR RR1, AAPOR, 2016) and about 64 percent (n=3,601) participated in wave two. Participation in the second wave was open to all panel members, irrespective of participation in wave one. Since there is no evidence that measurement error increases from wave one to wave two due to panel conditioning (Bach and Eckman, 2018), we treat each wave separately in our analysis. We refer to the first wave of this survey as LISS-1 and to the second wave as LISS-2. Results regarding motivated misreporting in both LISS-1 and LISS-2 are reported in Bach and Eckman (2018).

The second survey, the Survey on Free Time (SOFT), was a CAPI survey conducted in 2013 in the US. 1,120 households were selected from the U.S. Postal Service's Delivery Sequence File using a three-stage sampling design. Primary sampling units (PSU) were compromised of individual cities or urban areas. Secondary sampling units (SSU) used ZIP codes or ZIP code fragments within sampled (PSU) and participants were then sampled within SSUs. The response rate (AAPOR RR1) was about 27 percent (n=304). Respondents were randomly assigned to answer 16 filter questions in the interleafed format or in the grouped format. Filter questions asked about interest in sports, clothing purchases and watching television, followed by up to six follow-up questions.

The third survey, “Employment and Purchase Behavior in Germany” (EPBG), was a CATI survey conducted in Germany in 2011. 12,400 adults were selected from German administrative labor market records. The response rate was about 19 percent (AAPOR RR1) and we use 1,200 out of 2,400 completed cases in this analysis. The remaining 1,200 respondents completed the survey, but were assigned to experimental conditions not used in this paper. Respondents of the EPBG survey were asked 18 filter questions either in the interleafed or in the grouped format, covering clothing purchases, employment history and income sources. Four follow-up questions were asked for each filter, if applicable. We refer to this survey as EPBG. Results regarding motivated misreporting in this survey are reported in Eckman et al. (2014).

Table 2: Summary of four datasets from three surveys

	LISS-1	**LISS-2**	**SOFT**	**EPBG**
Country	NL	NL	U.S.	Germany
Mode	Web	Web	CAPI	CATI
Data collection	April 2012	May 2012	April-June 2013	Aug-Oct 2011
Number of filter questions	13	13	16	18
n *respondents*	3,767	3,601	304	1,200
Response rate[a]	68%	64%	27%	19%

[a] *AAPOR RR1 (AAPOR, 2016).*

See Table 2 for an overview of the four datasets. All of these datasets contain filter questions and respondents were randomly assigned to the different filter question formats (interleafed or grouped) in each survey.

5 Methods

To test our hypothesis, we need an estimate of the response propensity score to identify reluctant and non-reluctant respondents. Furthermore, we need a measure of measurement error (that is, the extent of motivated misreporting). We describe how we estimate response propensity and motivated misreporting below.

5.1 Estimation of response propensity

The idea of the response propensity builds on the seminal work of Rosenbaum and Rubin (1983) on propensity scores. Originally introduced in the field of evaluation studies, the propensity score denotes the conditional probability that a unit (e.g. a person) receives a treatment, given observable attributes of the unit. Similarly, the *response* propensity is the conditional probability that a person responds to a survey or not, given the person's attributes (Bethlehem et al., 2011, chapter 11). This score, RP_i, varies between zero and one and is a latent variable. Although we cannot observe it, we can observe the corresponding response indicator, R_i, which allows us to estimate response propensity scores.

Logistic regression is the most common technique for estimating response propensities (Bethlehem et al., 2011, chapter 11). The dependent variable in these models is the binary response indicator, R_i, indicating whether a unit responded to a survey or not. All variables known or assumed to influence whether a unit is a respondent to a survey are included in the model as covariates, often in various functional forms (e.g. linear, quadratic, or interacted with other predictors). Predictions from this model then form the response propensity scores.

In recent years, however, nonparametric prediction algorithms from machine learning methods have been introduced in the response propensity score literature (McCaffrey et al., 2004; Buskirk and Kolenikov, 2010; Phipps and Toth, 2012). In our study, we use one of these approaches, specifically, an extended version of Friedman's (2001) gradient boosting machine as implemented in the "gbm" package (version 2.1.3) in R (Ridgeway, 2017; R Core Team, 2018). Boosting is a prediction method based on the combination of several classification or regression trees (Hastie et al., 2009, chapter 9). Technical details of this algorithm are beyond the scope of this paper, but we provide an intuitive explanation of the general idea of boosting below, following McCaffrey et al. (2004). For a full description of the boosting approach, see, e.g., Friedman (2001, 2002); McCaffrey et al. (2004); Ridgeway (2017).

The major advantage of the boosting algorithm (and other machine learning methods) is that we do not need to determine the (correct) functional form of the predictor variables in the propensity score model, including the decision about which variables to include in the model at all. Rather, boosting automatically select covariates that are predictive of the response variable based on the available data. That is, we provide the boosting model with a list of covariates and let the algorithm, driven by the data, decide which variables are highly predictive of response and which variables are less predictive of response. In addition, boosting can deal with many covariates even if the sample size is small. Last but not least, simulation studies have shown that methods such as boosting often outperform standard approaches such as logistic regression in the estimation of (response) propensity scores (e.g., Lee et al., 2010; Buskirk and Kolenikov, 2010).

In general, boosting algorithms with binary outcomes proceed as followed. In a first step, they use the log-odds of, in our case, being a respondent as an initial guess of the response propensity score. In a second step, the algorithm searches for a small adjustment model (in the form of a classification tree) to the initial guess. If the algorithm finds an adjustment model that increases the model fit (measured via the the bernoulli log-likelihood), then the algorithm adds this adjustment model to the initial guess and calculates new residuals based on a combined model of the initial guess and the adjustment model. These new residuals are then used to calculate additional adjustment models, until the maximum number of adjustment models specified in advance (i.e., the maximum number of trees) is reached. The final boosting model, i.e., the final response propensity model is then calculated as a linear combination of the initial guess and all adjustment models. In addition, each tree is calculated based on a random subset of all observations (similar to bootstrapping) as this has been shown to reduce variation in the final prediction without affecting bias (Friedman, 2002). To guard against overfitting, we train the boosting algorithm using 75% of observations and k-fold cross-validation. We evaluate final model performance using the remaining 25% of observations (for details on cross-validation and training-test-set performance evaluation, see, for example, Hastie et al. (2009), Chapter 7).

5.2 Predictors of response

Using the boosting approach described above, we estimate the response propensity of each selected case using a separate model for each of the four datasets. The dependent variable in each model is a binary variable indicating whether a case responded to the survey or not. The independent variables in these models are all created from information that is available for both respondents and nonrespondents in each dataset. That is, for every survey, we create as many covariates as possible from information that accompanied the survey (e.g., paradata, sampling frame information or administrative records). A complete list of covariates in each model is shown in the online supplementary materials. As shown in Table 2, the four surveys were conducted in different modes (i.e. web, CAPI, and CATI). Therefore, the information available to build the response propensity model differs between the surveys.

LISS-1 and LISS-2 were both conducted as part of the longstanding LISS online panel. Panel members have responded to several other waves of the panel before taking part in our two surveys, and thus the amount of information available for both respondents and nonrespondents from previous waves is large. We create and include 116 covariates in the response propensity model for LISS-1. These covariates cover socio-demographic information (e.g. age, gender, education, employment), attitudes, response behavior in previous waves, household composition, as well as paradata from the initial recruitment interview for the panel. The response propensity model for LISS-2 includes the same information as LISS-1 plus information that was collected as part of LISS-1, i.e. whether a person responded in LISS-1, the filter question format and the number of filters triggered in LISS-1.

The amount of information available about both respondents and nonrespondents in SOFT is much smaller, in part because it is a face-to-face interview. The response propensity model includes covariates derived from paradata that were collected during the CAPI interviews, such as the date and time of the first contact attempt and whether a person ever refused the interview, as well as covariates derived from the sampling design, for example primary and secondary sampling unit identifiers.

The sample of EPBG was selected from the German administrative labor market records. Therefore, propensity model includes several predictors derived from the administrative data, such as age, gender, employment and unemployment history, and education. Furthermore, the model contains predictors derived from the sampling frame (e.g. stratum identifiers) and paradata from the CATI interview, such as date and time of a call, interviewer IDs and assessments of the likelihood of a case to participate in the survey which were made by interviewers (see Sinibaldi and Eckman, 2015, for details on this variable).

Using the boosting algorithm and the covariates described above, we predict the response propensity scores, our measure of reluctance, for respondents and nonrespondents in each dataset. We discuss model performance in the Results section.

5.3 Measuring motivated misreporting

We use the differences in filters triggered between the formats (interleafed vs. grouped), the format effect, as our measure of motivated misreporting. Furthermore, we analyze motivated misreporting at the question level rather than at the respondent level, following Eckman et al. (2014). We prefer this approach to the respondent level approach (where the outcome would be defined as the number of filters triggered by each person) because it gives us more statistical power to detect a connection between reluctance and motivated misreporting. To account for the fact that filters are nested within persons, we cluster variances at the respondent level, following the literature on the analysis of data with group-level randomization (Murray et al., 2004; Abadie et al., 2017).

In formal terms, we define $Y_j \in [0, 1]$ as the outcome indicating whether a filter question j was triggered or not. Furthermore, we define $I_j \in [0, 1]$ as an indicator of whether a filter question was asked in the interleafed ($I = 1$) or grouped ($I = 0$) format. We estimate the format effect, our measure of motivated misreporting, as the difference in means between the two formats:

$$E(Y|I = 1) - E(Y|I = 0) \tag{1}$$

Strictly speaking, the format effect we estimate is not true measurement error, the ε term in Figure 1. However, comparisons of survey data with administrative records (see Section 3) have shown that motivated misreporting, i.e., the format effect, is due to measurement error in the interleafed condition. Thus, we can use the format effect to test our hypothesis.

5.4 Identification of the relationship between reluctance and motivated misreporting

From the above boosted regression models, we have estimated response propensities for all respondents and nonrespondents. To capture reluctance, we split the estimated scores for the respondents into quartiles within each study. The fourth quartile contains respondents with the highest response propensity scores, i.e. respondents who are the most likely to respond to the survey, given their observed covariates described above. The first quartile, by contrast, contain respondents with the lowest response propensity scores, i.e. those who responded, but were not likely to do so. When we compare motivated misreporting between reluctant and likely respondents, we use only respondents in the fourth and first response propensity quartiles of each dataset. Comparing the format effect between the most likely respondents (the fourth quartile) and the least likely respondents (the first quartile) allows us to study whether reluctant respondents are worse reporters. In formal terms, we define $W_j \in [0, 1]$ as an indicator of whether the filter question was answered by a reluctant respondent ($W = 1$) or not ($W = 0$).

However, it is likely that reluctant and likely respondents differ on many characteristics (recall the covariates of the response propensity models, Table 7). For example, reluctant respondents of EPBG may actually have different employment histories then likely respondents. To account for this possibility of *true* differences in the behavior measured with the filter questions between the two types of respondents, we use a difference-in-difference approach. DiD models are commonly used in causal inference settings to derive treatment effects from non-randomized designs (Angrist and Pischke, 2009, 221-247). In our case, DiD controls for any true differences, relevant to the constructs measured in the filter questions, between reluctant and non-reluctant respondents. The DiD model is simply the difference of the differences between reluctant and non-reluctant respondents in each format as shown in Equation (2).

$$\begin{aligned} DiD = & [E(Y|W=1, I=1) - E(Y|W=0, I=1)] \\ & -[E(Y|W=1, I=0) - E(Y|W=0, I=0)] \end{aligned} \tag{2}$$

Alternatively, we can rearrange terms in (2) and interpret the DiD estimate as the difference between the format effect among reluctant and non-reluctant respondents, as shown in (3).

$$\begin{aligned} & [E(Y|W=1, I=1) - E(Y|W=1, I=0)] \\ & -[E(Y|W=0, I=1) - E(Y|W=0, I=0)] \end{aligned} \tag{3}$$

If there is no dependency between respondents' reluctance and motivated misreporting, the difference between reluctant and likely respondents in the percent of filters triggered in the interleafed format ($E(Y|W=1, I=1) - E(Y|W=0, I=1)$) should be about the same as the difference in the percent of filters triggered in the grouped format ($E(Y|W=1, I=0) - E(Y|W=0, I=0)$) (equation 2). If there is a connection between respondents' reluctance and misreporting, however, we should see that the difference in the percent of filter questions triggered between reluctant and likely respondents is larger in the interleafed format than in the grouped format, due to increased misreporting among reluctant respondents in the former format:

$$\begin{aligned} & E(Y|W=1, I=1) - E(Y|W=0, I=1) \\ & > E(Y|W=1, I=0) - E(Y|W=0, I=0) \end{aligned} \tag{4}$$

Estimation of our approach is straightforward using a linear regression model, as in (4), with intercept β_0, coefficients β_1, β_2, β_3 and residual error term ν.

$$Y_j = \beta_0 + \beta_1 I_j + \beta_2 W_j + \beta_3 I_j * W_j + \nu_j \quad (5)$$

That is, two binary variables (I and W) and, of greater interest, their interaction ($I * W$) are included in the model as independent variables. The coefficient of the interaction between these two variables, β_3, is our DiD estimate. This coefficient provides the test of our hypothesis that there is a connection between response propensity and motivated misreporting. If β_3 is negative, we interpret this as evidence that reluctant respondents show more motivated misreporting. If there is no significant effect, however, we take this as lack of evidence for a connection.

6 Results

Presentation of our results proceeds in three steps. First, we present key information on the response propensity models and the estimated response propensity scores for each survey. Second, we analyze whether the data in each survey is affected by motivated misreporting. Third, we present the findings regarding the connection between response propensity and motivated misreporting. To do so, we first inspect interaction plots of the DiD model that provide a straightforward graphical interpretation of the DiD estimate. We then focus on the estimated DiD coefficient and conclude the section with several robustness tests.

6.1 Response propensity models

Table 3 shows measures of predictive performance of each response propensity model. All models are optimized based on 4-fold cross-validation to guard against overfitting

(Hastie et al., 2009, ch.7), using 75% of the data as training data and the remaining 25% as test data for performance evaluation. Using Youden's J statistic (Youden, 1950) as a probability cutoff to evaluate model performance, between 76 and 88 percent of respondents are correctly classified as respondents (sensitivity column) and between 75 and 85 percent of nonrespondents are correctly classified (specificity column). Taken together, about 75 to 86 percent of all cases are correctly classified ('Accuracy'). Moreover, the area under the receiver operating characteristic curve (AUC) indicates excellent (AUC ≥ 0.8) to outstanding (AUC ≥ 0.9) discrimination in all models, according to the rules of thumb proposed by Hosmer and Lemeshow (2000). R-squared values, that is, the percent of log-likelihood explained by each model, vary between 0.28 and 0.53. Taken together, these performance metrics indicate that the response propensity model built for the SOFT survey discriminates very well between respondents and nonrespondents, followed by good predictive performance of LISS-1, LISS-2, and EPBG.

Regarding the most influential predictors of response, sociodemographic information such as the year of birth or having a migration background dominate the response propensity model in LISS-1. In LISS-2, sociodemographic information and covariates collected in LISS-1 have the greatest influence. The response propensity models for SOFT and EPBG, both surveys with interviewer involvement, are dominated by paradata collected during contact attempts (see Table 7 in the Online Supplementary Materials for more information on the most influential predictors in each dataset).

Table 4 shows the ranges of the first and fourth quartile of the estimated response propensity scores in each dataset. Given that we built a unique response propensity model for each dataset, it is not surprising that the range of propensity scores within the quartiles varies considerably between datasets. Since the value of the response propensity score itself has no meaningful interpretation (Bethlehem et al., 2011) and we are only interested in identifying reluctant and likely respondents, differing ranges

Table 3: Performance measures of response propensity models, by survey

Dataset	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**McFadden-R^2**
LISS-1	0.82	0.79	0.81	0.89	0.44
LISS-2	0.83	0.79	0.82	0.88	0.39
SOFT	0.88	0.85	0.86	0.94	0.53
EPBG	0.76	0.75	0.75	0.84	0.28

Note: Sensitivity, specificity and accuracy at optimal probability cut-point, as determined by maximal sensitivity and specificity (Youden, 1950). Performance calculated on 25% test set.

of response propensity scores across the four studies do not interfere with our analysis.

Table 4: Summary statistics of estimated response propensities, by survey

	Quartile						
		1st			4th		
Dataset	Min.	Mean	Max.	Min.	Mean	Max.	n[a]
LISS-1	0.08	0.55	0.70	0.92	0.95	0.99	1,883
LISS-2	0.11	0.55	0.76	0.89	0.92	0.96	1,801
SOFT	0.19	0.47	0.55	0.68	0.71	0.79	152
EPBG	0.05	0.20	0.28	0.52	0.68	0.89	600

[a]*Respondents in first and fourth response propensity quartiles only.*

6.2 Motivated misreporting

Table 5 shows results of the analysis of motivated misreporting in each dataset using *all* respondents. Motivated misreporting is taking place in all four datasets: the percent of filters triggered in the interleafed format (row one) is smaller than the percent triggered in the grouped format (row two). These results support the hypothesis that respondents learn to misreport in the interleafed format. Interestingly, when we calculate the difference in the number of filters triggered between the two formats (instead of the percent of filters triggered), the size of the effect seems to be about one filter question in every dataset (except for SOFT), i.e., misreporting patterns seem to be very consistent across these datasets (results not reported). The format effect in SOFT, significant at the 10% level, however, is only about half a filter question. The smaller effect size could be due due the fact that SOFT is a face-to-face survey, where the physical presence of an interviewer may cause respondents to report more honestly.

Table 5: Percent of filters triggered, by question format and survey

	LISS-1	LISS-2	SOFT	EPBG
Interleafed	42.9 (4.2)	43.3 (4.3)	49.5 (1.3)	42.4 (5.8)
Grouped	36.6 (3.8)	35.6 (3.7)	46.2 (1.2)	37.9 (5.6)
t-test[a]	11.22	13.48	1.84	5.65
p-value	0.000	0.000	0.066	0.000
$n_{filters}$	48,971	46,813	4,864	21,600
$n_{respondents}$	3,767	3,601	304	1,200

[a] *H_0: Percent of filters triggered interleafed = percent of filters triggered grouped.*
Note: Standard errors clustered at respondent level (in parentheses).

To sum up, we find evidence that motivated misreporting is taking place in every dataset: respondents deliberately give false or inaccurate answers to filter questions to avoid follow-up questions and reduce the burden of the survey.

6.3 Motivated misreporting among reluctant respondents

In the next step of our analysis, we reduce the analysis sample of each dataset to reluctant (lowest response propensity quartile) and likely respondents (highest response propensity quartile). We then estimate the difference-in-difference models described in Section 5.3. Before we turn to the table of results (Table 6), however, we inspect interaction plots (Figure 2). Interaction plots provide a straightforward graphical interpretation of the DiD models. If the dashed lines in Figure 2 are parallel, than there is no interaction between respondents' reluctance and motivated misreporting. If they are not parallel, however, we interpret this as evidence that there is an interaction, i.e., a connection, between reluctance and motivated misreporting. The plots for LISS-2, SOFT and EPBG (top right, bottom left and bottom right panels) suggest that there is no such connection, indicated by the nearly parallel dashed lines and the overlapping confidence intervals of the point estimates in both formats. That is, the differences in the percent of filters triggered in the grouped format is about the same as the difference in the percent of filters triggered in the interleafed format. The interaction plot of LISS-1 (top left panel), however, suggests that there may be a connection between respondents' reluctance and motivated misreporting, as the difference in the percent of filters triggered in the interleafed format is larger than the difference in the percent of filters triggered in the grouped format (indicated by the non-parallel dashed lines). That is, reluctant respondents seem to be more prone to misreporting in one of the four datasets. To inspect these results more closely, we turn to the regression estimates shown in Table 6.

Regarding LISS-1, LISS-2, and EPBG we find that the percent of filters triggered is smaller in the interleafed format than in the grouped format (first row) for likely respondents, after accounting for respondents' reluctance. The format effect in the SOFT survey, however, is no longer significant (at the 10%-level), although the negative sign and the coefficient still indicate that respondents in the interleafed format trigger fewer filters than respondents in the grouped format. This finding may be due to the very small sample size of the SOFT survey (recall that we use only half of the respondents in this analysis).

Reluctant grouped format respondents in LISS-1, LISS-2, and SOFT do not report fewer filters (indicated by the insignificant coefficients on the reluctance indicator) than likely respondents (see Section 5.4 for a discussion of the interpretation of model estimates). In EPBG, there seems to be a difference between reluctant and non-reluctant respondents: the percent of filters triggered by reluctant respondents is smaller than the percent of filters triggered by likely respondents ($\hat{\beta} = -4.01, s.e. =$

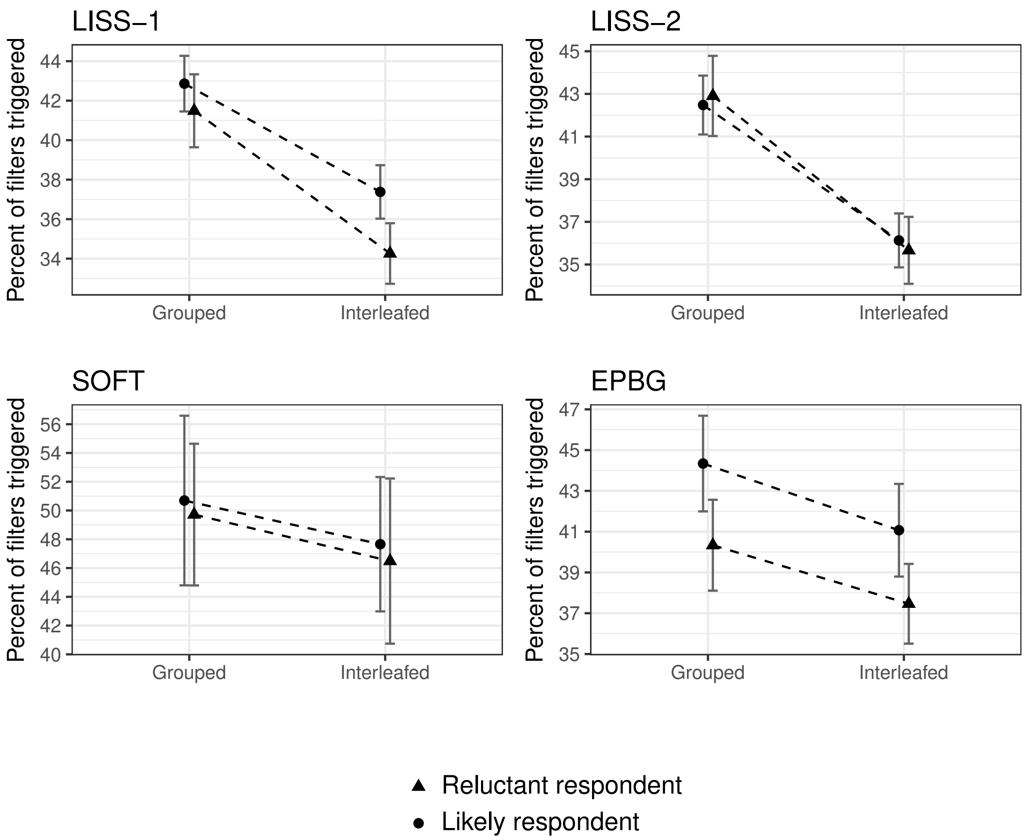

Figure 2: Interaction plots of respondents' reluctance and percent of filter questions triggered, by survey. Point estimates with 95% confidence intervals. Dashed lines added for ease of interpretation. Respondents in the first and fourth response propensity quartiles only.

1.65). This result is likely due to true differences in purchasing behavior among reluctant and non-reluctant respondents.

Table 6: OLS difference-in-difference estimates of the influence of response propensity on motivated misreporting, by survey.

	LISS-1	**LISS-2**	**SOFT**	**EPBG**
Interleafed	-5.48***	-6.35***	-3.04	-3.27*
(ref. grouped)	(1.00)	(0.95)	(3.81)	(1.66)
Reluctant respondent	-1.37	0.43	-0.98	-4.01*
(ref. likely resp.)	(1.18)	(1.19)	(3.89)	(1.65)
Interleafed*Reluctant	-1.75	-0.89	-0.19	0.40
respondent	(1.58)	(1.57)	(5.40)	(2.25)
$n_{\text{filters}}{}^{a}$	24,479	23,413	2,432	10,800
$n_{\text{respondents}}{}^{a}$	1,883	1,801	152	600

*Note: *** p<0.001,** p<0.01, * p<0.05. Standard errors clustered at respondent level (in parentheses).*
[a]Respondents in the first and fourth response propensity quartiles only.

To answer our research question, we check whether the format effect is different for reluctant respondents (recall the DiD model described in Section 5.4). The interaction effect (third row), our main coefficient of interest, is non-significant in all four models. That is, the difference in the percent of filters triggered by the interleafed and the grouped format is the same for reluctant as for likely respondents. Thus, we do not find evidence that motivated misreporting is stronger among reluctant respondents. However, looking at the effect size of the DiD estimate, we replicate the finding from Figure 2 that there is a tendency among reluctant respondents in the interleafed format of LISS-1 to report fewer filters than likely respondents, after accounting for true differences in behavior. Thus, there seems to be only small evidence for a connection between respondents' reluctance and motivated misreporting.

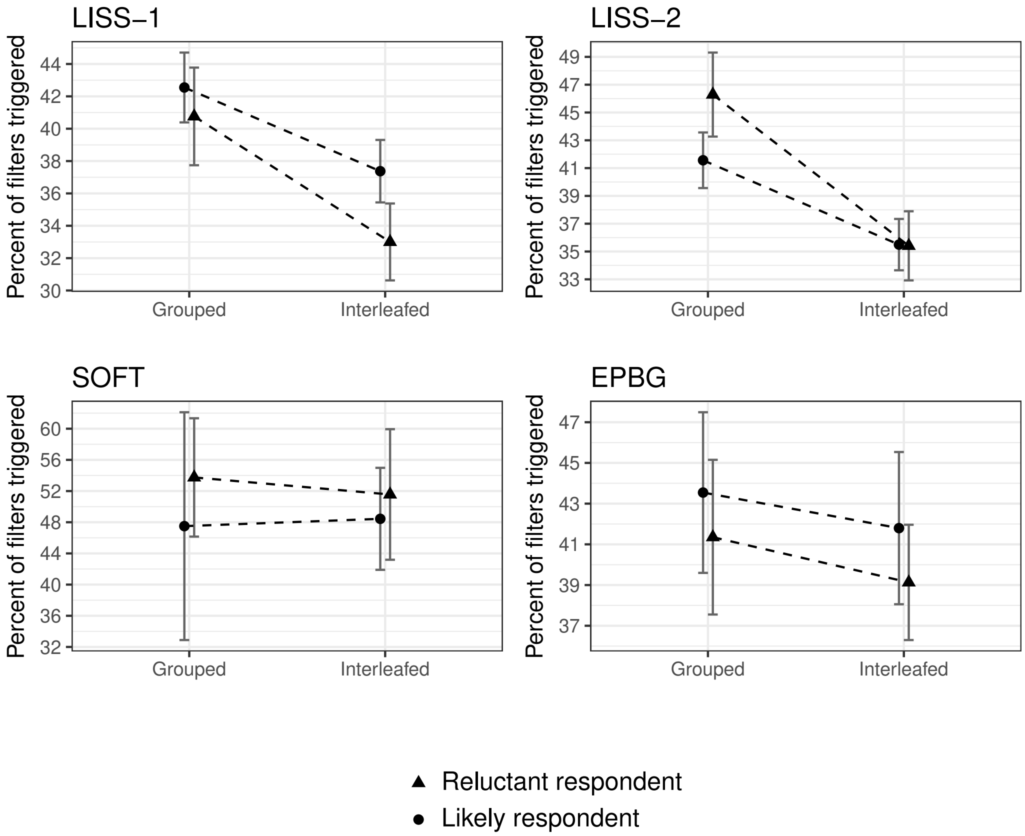

Figure 3: Interaction plots of respondents' reluctance and percent of filter questions triggered, by survey. Point estimates with 95% confidence intervals. Dashed lines added for ease of interpretation. Respondents in the first and tenth response propensity deciles only.

6.4 Robustness checks

To assess the robustness of our results, we specify several alternative models, which we briefly discuss below. As a first check, we modify the reluctance indicator to assess the robustness of the results presented in Table 6 to the specification of reluctant and likely respondents. Instead of comparing misreporting between respondents of the first and fourth response propensity quartile, we analyze misreporting of respondents in the first and tenth response propensity *decile*. That is, our definition of reluctance covers only 20% of all respondents (instead of 50%) - the most reluctant 10% and the most likely 10%.

The results of these robustness checks (Figure 3 and Table 9 in the Online Supplementary Materials) are generally in line with the findings presented in the previous section. In LISS-1 and LISS-2, the most reluctant decile of respondents are worse reporters than the most likely decile of respondents. That is, in LISS-1, the plot (top left panel of Figure 3) suggests that reluctant respondents trigger fewer filters than likely respondents, after accounting for true differences in behavior. In LISS-2 (top right panel), we see that reluctant respondents trigger *more* filters than likely respondents in the grouped format. In the interleafed format, however, this difference disappears. Based on the assumptions given in Section (misreporting is only possible in the interleafed format and differences between reluctant and likely respondents should be the same in the two formats), we interpret this finding as evidence that reluctant respondents are worse reporters than likely respondents. These findings are supported by the regression estimates (Table 9 in the Online Supplementary Materials). In SOFT and EPBG, however, there is no evidence that motivated misreporting to filter questions is worse among reluctant respondents, a finding also supported by the small and nonsignificant interaction effects in Table 10 in the Online Supplementary Materials.

As additional robustness checks, we modify the specification of the models described in Section 5.4. First, instead of clustering variances at the respondent level, we specify random intercept models to account for the correlation of filters within respondents (Murray et al., 2004; see also Section 5.3). Results of these models and their interpretation regarding a connection between response propensity and motivated misreporting, however, do not differ from the results presented in Section 6. Second, we include socio-demographic control variables (e.g., age, education, gender) in the models specified in Section 5.4 to reduce some variation in the dependent variables and thereby increase the precision of our estimates. Including those control variables, however, does not lead to substantial changes in coefficients of our central indicators (neither regarding their magnitude, nor regarding their significance; results not shown). Third, to increase the power of our analyses to detect interactions between format and reluctance, we combined three of our four data sets and ran one analysis. The case base for the model is all cases in the first and fourth response propensity quartiles from the LISS-1, LISS-2, and SOFT data sets. The dependent variable in the model is the yes/no filter question response (as in all other models). The independent variables are the indicators I, W, and $I * W$, as well as an indicator of the data set. The final model contained 50,324 filters (nested in 3,836 respondents). This model also does not detect a significant interaction between I and W. Including respondents' age and sex as additional independent variables does not meaningfully change the results.

To sum up, the results reported in Section 6 and the results of the robustness checks discussed above, provide mixed support to our hypothesis of a connection between response propensity and motivated misreporting to filter questions. Contrary to our expectations, we find some evidence for the hypothesized connection in the LISS datasets, but not in EPBG and SOFT.

7 Discussion

Are reluctant respondents more likely to introduce measurement error, specifically motivated misreporting? Using data from four surveys conducted in different modes and countries, we analyzed the connection between response propensity and motivated misreporting to filter questions, a form of measurement error, to answer this question. We estimated response propensities using a data mining algorithm that allows us to sidestep the challenge of having to pre-specify a set of predictors of response from all available covariates and their correct functional form. While we did find evidence for a connection between response propensity and motivated misreporting in two of our datasets (two waves of the Dutch LISS panel survey), we did not find evidence in the other two surveys.

The nonresponse-measurement error model offers a theoretical explanation of why reluctant respondents may report less accurate data than likely respondents. This model states that the survey reports are a function of the true value and an error term that is determined by the response propensity. Our results, at least for two out of four datasets, do not support this model. We do not believe that this model is wrong, rather, there may be additional factors that determine whether this model holds or not. Interestingly, data for the two surveys where we found some evidence for a connection between nonresponse and measurement error were both conducted on a self-administered basis *without* interviewer involvement (web surveys). In the other two surveys, by contrast, interviewers were involved in the data collection process (CAPI and CATI). A likely explanation for the lack of a connection between response propensity and motivated misreporting is that the presence of an interviewer guards against excessive misreporting among the most reluctant respondents (see also the discussion of the results of the SOFT survey in Section 6.2). We do not believe that cross-country differences explain the differing findings as the phenomenon of motivated misreporting seems to be consistent across countries. Furthermore, it seems unlikely that the content of the filter questions explains the differences in our findings because all surveys contain similar questions on purchasing behavior and the findings regarding the level of motivated misreporting (see Section 6.2) are consistent across three of the four surveys.

Another possible explanation for the absence of a connection between nonresponse and motivated misreporting is that once a sampled person decides to participate in the survey, her motivation or interest in the survey is high enough to give answers as correct as any other respondents (at least, in SOFT and EPBG). The desire to reduce survey burden may be a good explanation for motivated misreporting, but sampled units with a strong desire to reduce survey burden may simply decide to not participate in the survey at all. In other words, the lowest response propensity cases are in fact nonrespondents, and we are not able to explore the patterns of measurement error among nonrespondents.

For a better understanding of the nexus between nonresponse and measurement error, we would like to see our results replicated with other forms of motivated misreporting. Looping questions, for example, another form of eligibility questions, have also been shown to be prone to motivated misreporting (Eckman and Kreuter, 2018) and similar findings have been reported for screening questions (Tourangeau et al., 2012). However, we did not include these types of questions in our study as there exist only two studies (mentioned above) regarding misreporting to looping and screener questions so far and these two studies unfortunately do not come with the kind of information necessary to estimate accurate prediction models of response propensity. In addition, future studies should also follow up on the research mentioned in Section 3 and explore the connection between nonresponse and other forms of measurement error.

The finding that reluctant respondents do not misreport more to filter questions than likely respondents in all cases is good news for researchers who put extra effort into achieving high response rates. While high response rates do not necessarily decrease bias due to nonresponse (see the literature reviewed in Section 1), we find only limited evidence that the extra effort introduces additional measurement error in terms of increased levels of motivated misreporting.

References

AAPOR (2016). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys (9th ed.). Oakbrook Terrace, IL: American Association for Public Opinion Research.

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? NBER Working Paper Series 24003.

Angrist, J. and J.-S. Pischke (2009). Mostly Harmless Econometrics. An Empiricist's Companion. Princeton, NY: Princeton University Press.

Bach, R. L. and S. Eckman (2018). Motivated misreporting in web panels. Journal of Survey Statistics and Methodology 6(3), 418–430.

Bethlehem, J., F. Cobben, and B. Schouten (2011). Handbook of Nonresponse in Household Surveys. Hoboken, NJ: John Wiley & Sons, Inc.

Biemer, P. P. (2001). Nonresponse bias and measurement bias in a comparison of face-to-face and telephone interviewing. Journal of Official Statistics 17(2), 295–320.

Bollinger, C. R. and M. H. David (2001). Estimation with response error and non-response: Food-stamp participation in the sipp. Journal of Business & Economic Statistics 19(2), 129–141.

Buskirk, T. D. and S. Kolenikov (2010). Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification. Public Opinion Quarterly 74(3), 413–432.

Cannell, C. F. and F. J. Fowler (1963). Comparison of a self-enumerative procedure and a personal interview: A validity study. Public Opinion Quarterly 27(2), 250–264.

Curtin, R., S. Presser, and E. Singer (2000). The effects of response rate changes on the index of consumer sentiment. Public Opinion Quarterly 64(4), 413–428.

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

Duan, N., M. Alegria, G. Canino, T. McGuire, and D. Takeuchi (2007). Survey conditioning in self-reported mental health service use: Randomized comparison of alternative instrument formats. Health Research and Educational Trust 42(2), 890–907.

Eckman, S. and F. Kreuter (2018). Misreporting to looping questions in surveys: Recall, motivation and burden. Survey Research Methods 12(1), 59–74.

Eckman, S., F. Kreuter, A. Kirchner, A. Jäckle, R. Tourangeau, and S. Presser (2014). Assessing the Mechanisms of Misreporting to Filter Questions in Surveys. Public Opinion Quarterly 78(3), 721–733.

Fricker, S. and R. Tourangeau (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. Public Opinion Quarterly 74(5), 934–955.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5), 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4), 367–378.

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

Groves, R. M., M. P. Couper, and S. Dipko (2004). The role of topic interest in survey participation decisions. Public Opinion Quarterly 68(1), 2–31.

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

Hastie, T., R. Tibshirani, and J. Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second ed.). Berlin: Springer.

Hosmer, D. and S. Lemeshow (2000). Applied Logistic Regression. New York: Wiley.

Hox, J. J., E. de Leeuw, and H.-T. Chang (2012). Nonresponse versus measurement error: Are reluctant reporters worth pursuing? Bulletin de Méthodologie Sociologique 113, 5–19.

Keeter, S., A. Kohut, C. Miller, R. Groves, and S. Presser (2000). Consequences of reducing nonresponse in a large national telephone survey. Public Opinion Quarterly 64(2), 125–148.

Kessler, R. C., H.-U. Wittchen, J. M. Abelson, K. McGonagle, N. Schwarz, K. S. Kendler, B. Knäuper, and S. Zhao (1998). Methodological Studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS). International Journal of Methods in Psychiatric Research 7(1), 33–55.

Kreuter, F., S. McCulloch, S. Presser, and R. Tourangeau (2011). The Effects of Asking Filter Questions in Interleafed Versus Grouped Format. Sociological Methods & Research 40(1), 88–104.

Kreuter, F., G. Müller, and M. Trappmann (2010). Nonresponse and measurement error in employment research. making use of administrative data. Public Opinion Quarterly 74(5), 880–906.

Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving propensity score weighting using machine learning. Statistics in Medicine 29(3), 237–262.

Martin, C. L. (1994). The impact of topic interest on mail survey response behaviour. Journal of the Market Research Society 36(4), 327–338.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods 9(4), 403–425.

Merkle, D. and M. Edelman (2002). Nonresponse in exit polls: A comprehensive analysis. In R. Groves, D. Dillman, J. Eltinge, and R. Little (Eds.), Survey Nonresponse, pp. 243–258. New York: John Wiley and Sons.

Murray, D. M., S. P. Varnell, and J. L. Blitstein (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. American Journal of Public Health 3(94), 423–432.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. Public Opinion Quarterly 70(5), 737–758.

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(1), 129–145.

Peytchev, A., E. Peytcheva, and R. M. Groves (2010). Measurement error, unit nonresponse, and self-reports of abortion experiences. Public Opinion Quarterly 74(2), 319–327.

Phipps, P. and D. Toth (2012). Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data. Annals of Applied Statistics 6, 772–794.

R Core Team (2018). A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Ridgeway, G. (2017). gbm: Generalized boosted regression models. R package version 2.1.3. `https://cran.r-project.org/package=gbm`.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70(1), 41–55.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: How the liss panel was built and how it can be used. BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique 109, 56–61.

Sinibaldi, J. and S. Eckman (2015). Using call-level interviewer observations to improve response propensity models. Public Opinion Quarterly 79(4), 976–993.

Tourangeau, R., R. M. Groves, and C. D. Redline (2010). Sensitive topics and reluctant respondents. demonstrating a link between nonresponse bias and measurement error. Public Opinion Quarterly 74(3), 413–432.

Tourangeau, R., F. Kreuter, and S. Eckman (2012). Motivated underreporting in screening interviews. Public Opinion Quarterly 76(3), 453–469.

Tourangeau, R., F. Kreuter, and S. Eckman (2015). Motivated misreporting: Shaping answers to reduce survey burden. In U. Engel (Ed.), Survey Measurements. Techniques, Data Quality and Sources of Error, pp. 24–41. Frankfurt/New York: Campus.

Triplett, T., J. Blair, T. Hamilton, and Y. C. Kang (1996). Initial cooperators vs. converted refusers: Are there response behavior differences? Proceedings of the American Statistical Association, Survey Research Methods Section, 1038–1041.

Willimack, D. K., H. Schuman, B. E. Pennell, and J. M. Lepkowski (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. Public Opinion Quarterly, 78–92.

Yan, T., R. Tourangeau, and Z. Arens (2004). When less is more: Are reluctant respondents poor reporters? Proceedings of the American Statistical Association, Survey Research Methods Section, 4632–4651.

Youden, W. J. (1950). Index for rating diagnostic tests. Cancer 3, 32–35.

8 Online Supplementary Materials

8.1 Additional results

Table 7: Relative influence[a] of most influential predictors of response, by dataset

LISS-1	LISS-2	SOFT	EPBG
Gross household income 6.5 %	Number of filters triggered in wave one 30.6 %	Screening interview started 27.1 %	Appointment made 10.7 %
Year of birth 6.3 %	Gross household income 4.9 %	Date of second contact attempt 9.3 %	Income full-time job 8.4 %
Age at interview 5.7 %	Net household income 4.2 %	Time of first contact attempt 5.9 %	Year of birth 8.1 %
Age of household head 4.7 %	Wave one FQ format 3.9 %	Date of first contact attempt 5.8 %	Calls per case 7.5 %
Net household income 4.5 %	Age of household head 3.0 %	Time of second contact attempt 5.6 %	Income part-time job 6.7 %
Number of predictors			
116	120	47	42

[a] *Percentage of log likelihood explained by predictor relative to the total log likelihood explained by the model.*

Table 8: Percent of filters triggered, by respondents' reluctance, question format, and survey

		LISS-1	LISS-2	SOFT	EPBG
Interleafed	Reluctant respondents	34.3 (0.78)	35.7 (0.80)	46.5 (2.91)	37.5 (1.00)
	Likely respondents	37.4 (0.69)	36.1 (0.65)	47.7 (2.36)	41.1 (1.16)
Grouped	Reluctant respondents	41.5 (0.94)	42.9 (0.96)	49.7 (2.49)	40.3 (1.14)
	Likely respondents	42.9 (0.72)	42.5 (0.70)	50.7 (2.98)	44.3 (1.20)
n_{filters} [a]		24,479	23,413	2,432	10,800
$n_{\text{respondents}}$ [a]		1,883	1,801	152	600

[a] *Respondents in the first and fourth response propensity quartiles only.*

Note: Standard errors clustered at respondent level (in parentheses).

Table 9: Summary statistics of estimated response propensities, by survey

	Decile						
		1st			10th		
Dataset	Min.	Mean	Max.	Min.	Mean	Max.	n[a]
LISS-1	0.08	0.43	0.56	0.95	0.96	0.99	1,883
LISS-2	0.11	0.32	0.62	0.92	0.93	0.96	1,801
SOFT	0.33	0.40	0.46	0.67	0.72	0.78	152
EPBG	0.05	0.14	0.20	0.70	0.78	0.89	600

[a] *Respondents in first and tenth response propensity deciles only.*

Table 10: Difference-in-difference estimates of the influence of response propensity on motivated misreporting, by survey (lowest and highest decile)

	LISS-1	**LISS-2**	**SOFT**	**EPBG**
Interleafed	-5.17***	-6.07***	-3.04	-1.75
(ref. grouped)	(1.47)	(1.39)	(3.81)	(2.76)
Reluctant respondent	-1.79	4.72*	-0.98	-2.20
	(1.89)	(1.85)	(3.89)	(2.78)
Interleafed*Reluctant	-2.59	-4.81*	-0.19	-0.47
respondent	(2.45)	(2.43)	(5.40)	(3.66)
n_{filters} [a]	9,789	9,373	976	4,320
$n_{\text{respondents}}$ [a]	753	721	61	240

*Note: *** $p<0.001$, ** $p<0.01$, * $p<0.05$. Standard errors clustered at respondent level (in parentheses).*

[a] *Respondents in the first and tenth response propensity deciles only.*

8.2 Text of filter questions — LISS-1 and LISS-2

- In the past month, have you purchased coffee for consumption at home?
- In the past month, have you purchased beer or wine for consumption at home?
- In the past month, have you purchased tobacco?
- In the past month, have you purchased children's clothing or shoes?
- In the past month, have you purchased clothing or shoes for yourself?
- In the past month, have you purchased chocolate?
- In the past month, have you purchased medication?
- In the past month, have you purchased flowers?
- In the past month, have you purchased pet supplies?
- In the past month, have you purchased movies on DVD or VHS?
- In the past month, have you purchased music on CD or as MP3s (or other digital formats)?
- In the past month, have you purchased a ticket for a concert, theater performance or a movie?
- In the past month, have you purchased any cleaning supplies for your home?

Follow-up questions

For each yes answer to the above filter questions:

Thinking about your most recent purchase of (fill: item)

- How much did it cost?

  - (Open ended response in Euros)
  - a. Don't know
  - b. Refused

- For whom was it purchased?
  - a. self
  - b. another household member
  - c. someone else
  - d. Don't know
  - e. Refused

8.3 Text of filter questions — SOFT

Sports section filter questions

- Do you follow professional hockey?
- Do you follow professional basketball?
- Do you follow professional soccer?
- Do you follow professional football?

Sports section follow-up questions

For each yes answer to the above filter questions:

- Do you have a favorite team you support, or do you have no particular favorite team?

  - a. I have a favorite team
  - b. I do not have a favorite team
  - c. cannot say

- When you watch (or attend) (fill: item) games, do you usually do so alone, or with family members or friends?
  - a. alone
  - b. with family members
  - c. with friends
  - d. with both family members and friends
  - e. it depends

- In the last 12 months, how much money have you spent purchasing (fill: item)-team merchandise ?
  - (Open ended response in Dollars)

- When (fill: item) is in season, how many hours per week, on average, do you spend watching (fill: item)?
  - (Open ended response in hours)

- Do you also follow college or high school (fill: item), or only professional (fill: item)?
  - a. college
  - b. high school
  - c. mentioned another league

- d. only professional

- Were you also a fan as a child or teenager, or did you only become a fan since getting older?
  - a. also fan as child
  - b. fan only since getting older

Clothing section filter questions

- In the last 12 months, have you purchased shoes?
- In the last 12 months, have you purchased jeans or pants?
- In the last 12 months, have you purchased a shirt or sweater?
- In the last 12 months, have you purchased sport clothing?
- In the last 12 months, have you purchased a coat or jacket?
- In the last 12 months, have you purchased a business suit?

Clothing section follow-up questions

For each yes answer to the above filter questions:

- Thinking of the most recent (fill: item) you purchased, was/were it/those for yourself, or for someone else?
  - a. self
  - b. someone else

- How much did the/those (fill: item) cost?

- (Open ended response in Dollars)

- Was this an impulse purchase or a planned purchase?
  - a. planned
  - b. impulse

- At what kind of store did you buy the/these (fill: item)?
  - a. department or big box store
  - b. local or boutique store
  - c. online
  - d. other

- How comfortable would you feel purchasing such items online?
  - a. very comfortable
  - b. somewhat comfortable
  - c. somewhat uncomfortable
  - d. not at all comfortable

TV section filter questions

- Do you regularly watch soap operas on television?
- Do you regularly watch reality television shows?
- Do you regularly watch late night talk shows?
- Do you regularly watch evening drama shows?

- Do you regularly watch evening scripted comedy shows?
- Do you regularly watch evening news programs?

TV section follow-up questions

For each yes answer to the above filter questions:

- What is the name of your favorite show in this category?
  - (open ended response)
- How many hours per week, on average, do you spend watching such shows?
  - (open ended response in hours)
- Do you believe such programs have educational value?
  - a. yes
  - b. no
- Approximately how many different shows in this category do you watch regularly?
  - (open ended response)

8.4 Text of filter questions — EPBG

Clothing section filter questions

- This year, that is in 2011, have you bought a coat or jacket for yourself or for someone else?

- This year, that is in 2011, have you bought a shirt or a blouse for yourself or for someone else?
- This year, that is in 2011, have you bought trousers for yourself or for someone else?
- This year, that is in 2011, have you bought shoes for yourself or for someone else?
- This year, that is in 2011, have you bought sportswear for yourself or for someone else?
- This year, that is in 2011, have you bought swimwear for yourself or for someone else?

Clothing section follow-up questions

- For whom did you purchase this/those (fill: item)? For yourself, a family member, or someone else?
- In what month did you purchase this/those (fill: item)?
- How much did this/those (fill: item) cost?
- How satisfied are you with this/those (fill: item)? Are you very satisfied, somewhat satisfied, somewhat dissatisfied, or very dissatisfied?

Employment section filter questions

- Have you ever held a full-time job? (Note: We explicitly instructed respondents not to include self-employment.)
- Have you ever held a part-time job? (Note: We explicitly instructed respondents not to include self-employment or Mini-Jobs.)

- Have you ever held a so-called Mini-Job, with a payment of 400 Euros a month or less?
- Have you ever received professional training?
- Have you ever received paid practical training?
- Have you ever been self-employed?

Employment section follow-up questions

- From when and until when did you hold your most recent (fill: item)?
- How many hours per week did/do you work in your most recent (fill: item)?
- In what industry was/is your most recent (fill: item)?
- What was your last monthly income at your most recent (fill: item)?

Income section filter questions

- In the year 2010: Did you or another person in your household have income from interest or investment income, e.g., savings, shares, equity funds, or fixed-interest securities?
- In the year 2010: Did you or another person in your household have income from rental property, including leases and subleases?
- In the year 2010: Did you or another person in your household receive a child benefit?
- In the year 2010: Did you or another person in your household receive parental money or a maternity benefit?

- In the year 2010: Did you or another person in your household receive income support?
- In the year 2010: Did you or another person in your household receive unemployment insurance?

Income section follow-up questions

- Which person in your household has received income from (fill: item)? You yourself or another member of your household?
- How often (with what regularity) did your household receive (fill: item)?
- How large was the last amount of income from (fill: item) that your household received in 2010?
- In what month in 2010 did your household first receive income from (fill: item)?

8.5 Predictor variables, by survey

Survey	Variables
LISS-1	Gender
	Position in household
	Year of birth
	Age in CBS categories
	Age of household head
	Number of household members
	Number of children in household
	Household Head lives together with a partner
	Civil Status
	Domestic situation
	Type of household's dwelling
	Urban character of place of residence

Primary occupation

Personal gross monthly income (in Euros)

Personal gross monthly income (in Euros), imputed

Personal net monthly income (in Euros) (incl. nettocat)

Personal net monthly income (in Euros)

Personal net monthly income (in Euros), imputed

Personal gross monthly income in categories

Personal net monthly income in categories

Gross household income (in Euros)

Net household income (in Euros)

Highest level of education (without diploma)

Highest level of education (with diploma)

Level of education in CBS categoriess

Household member participates in the panel

Recruitment wave

Ethnic group

Does the household have a simPC?

Count of responses in previous waves

Interviewer ID recruitment interview

Instrument (through which respondent was originally contacted)

Instrument datafile (through which respondent was finally contacted)

Usable address

Contact successful

Minimally posed central question

Complete recruitment interview

Willing to participate in the panel

Registered as panel member

Number of CATI contacts

Number of CAPI contacts

Total number of contact attempts

Interviewer used designated arguments

Life satisfaction (in general)

Statement: Feel good about myself

Confidence in abilities

Follow news on TV or radio

Follow news on the internet

Follow news through free daily paper

Follow news through national or regional newspaper

Follow news hardly or never

Follow news: do not know

Interested in news

Been to cinema in past 12 months

Visited museum in the Netherlands in the past 12 months

Doing voluntary work

Actively taken part in activities of one or more associations or organisations

Doing sports? If yes how many hours a week (on average)

Grade of health at present

Suffer from one or more long-term diseases afflictions or handicaps

Frequency of contact to friends, close acquaintances or family members

Statement: Enough people to fall back on in event of misfortune

Statement: Miss having people around

Statement: Life is meaningless without religion

Interested in political topics

Voted in 2006 election

Which of the two best decribes own view

Household has computer with internet connection

Type of internet connection: Cable connection

Type of internet connection: ADSL

Type of internet connection: Dial-up connection 1

Type of internet connection: Dial-up connection 2

Type of internet connection: Mobile internet

Type of internet connection: other fast, broadband

Type of internet connection: do not know 1

Type of internet connection: do not know 2

Type of internet connection: Dial-up connection used 1

Type of internet connection: Dial-up connection used 2

Type of internet connection: Cable connection

Type of internet connection: ADSL connection, DSL

Type of internet connection: Mobile internet

Type of internet connection: other fast, broadband

Type of internet connection: do not know

Household has computer without internet connection

Computer has: Windows Vista

Computer has: Windows XP

Computer has: Windows 2000

Computer has: Windows 95 or 98

Computer has: Linux variant

Computer has: Mac OS (Apple)

Computer has: other operating system

Computer has: do not know

Gender

Date of birth: day

Date of birth: month

Date of birth: year

Age

Composition of household

Household composition

Age of youngest child in household

Which of following descriptions best applies to respondent

Does respondent work

How many hours a week working in total (in a normal week)

Highest form of completed education

How well can respondent make ends meet on household's income

Estimated gross income

Born in the Netherlands

In which country born otherwise

Farther born in Netherlands

In which country was farther born otherwise

Mother born in Netherlands

	In which country was mother born otherwise
	Usually speaking Dutch at home or other language
	Statement: to acquire material posse is one of most important things in life
	Statement: Tolerance better than intolerance
	Statement: Survey research is important for society
LISS-2	Gender
	Position in household
	Year of birth
	Age in CBS categories
	Age of household head
	Number of household members
	Number of children in household
	Household Head lives together with a partner
	Civil Status
	Domestic situation
	Type of household's dwelling
	Urban character of place of residence
	Primary occupation
	Personal gross monthly income (in Euros)
	Personal gross monthly income (in Euros), imputed
	Personal net monthly income (in Euros) (incl. nettocat)
	Personal net monthly income (in Euros)
	Personal net monthly income (in Euros), imputed
	Personal gross monthly income in categories
	Personal net monthly income in categories
	Gross household income (in Euros)
	Net household income (in Euros)
	Highest level of education (without diploma)
	Highest level of education (with diploma)
	Level of education in CBS categoriess
	Household member participates in the panel
	Recruitment wave
	Ethnic group

Does the household have a simPC?

Count of responses in previous waves

Interviewer ID recruitment interview

Order of answers

Instrument (through which respondent was originally contacted)

Instrument datafile (through which respondent was finally contacted)

Usable address

Contact successful

Minimally posed central question

Complete recruitment interview

Willing to participate in the panel

Registered as panel member

Number of CATI contacts

Number of CAPI contacts

Total number of contact attempts

Interviewer used designated arguments

Life satisfaction (in general)

Statement: Feel good about myself

Confidence in abilities

Follow news on TV or radio

Follow news on the internet

Follow news through free daily paper

Follow news through national or regional newspaper

Follow news hardly or never

Follow news: do not know

Interested in news

Been to cinema in past 12 months

Visited museum in the Netherlands in the past 12 months

Doing voluntary work

Actively taken part in activities of one or more associations or organisations

Doing sports? If yes how many hours a week (on average)

Grade of health at present

Suffer from one or more long-term diseases, afflictions or handicaps

Frequency of contact to friends, close acquaintances or family members

Statement: Enough people to fall back on in event of misfortune

Statement: Miss having people around

Statement: Life is meaningless without religion

Interested in political topics

Voted in 2006 election

Which of the two best decribes own view

Household has computer with internet connection

Type of internet connection: Cable connection

Type of internet connection: ADSL

Type of internet connection: Dial-up connection 1

Type of internet connection: Dial-up connection 2

Type of internet connection: Mobile internet

Type of internet connection: other fast, broadband

Type of internet connection: do not know 1

Type of internet connection: do not know 2

Type of internet connection: Dial-up connection used 1

Type of internet connection: Dial-up connection used 2

Type of internet connection: Cable connection

Type of internet connection: ADSL connection, DSL

Type of internet connection: Mobile internet

Type of internet connection: other fast, broadband

Type of internet connection: do not know

Household has computer without internet connection

Computer has: Windows Vista

Computer has: Windows XP

Computer has: Windows 2000

Computer has: Windows 95 or 98

Computer has: Linux variant

Computer has: Mac OS (Apple)

Computer has: other operating system

Computer has: do not know

Gender

	Date of birth: day
	Date of birth: month
	Date of birth: year
	Age
	Composition of household
	Household composition
	Age of youngest child in household
	Which of following descriptions best applies to respondent
	Does respondent work
	How many hours a week working in total (in a normal week)
	Highest form of completed education
	How well can respondent make ends meet on household's income
	Estimated gross income
	Born in the Netherlands
	In which country born otherwise
	Farther born in Netherlands
	In which country was farther born otherwise
	Mother born in Netherlands
	In which country was mother born otherwise
	Usually speaking Dutch at home or other language
	Statement: to acquire material posse is one of most important things in life
	Statement: Tolerance better than intolerance
	Statement: Survey research is important for society
	Filter question format in LISS 1
	Number of filters triggered in LISS 1
	Case responded in LISS 1
SOFT	Screening interview started
	Call First
	Condition: How-many
	Condition: Grouped
	Condition: Paired
	Primary sampling unit
	Segment

Transfer case

Screener completed

Screener started

Start disposition code

Number of contacts total

Time of 1st contact

Time of 2nd contact

Time of 3rd contact

Time of 4th contact

Time of 5th contact

Time of 6th contact

Time of 7th contact

Time of 8th contact

Time of 9th contact

Time of 10th contact

Interviewer ID

Contact mode of 1st contact

Contact mode of 2nd contact

Contact mode of 3rd contact

Contact mode of 4th contact

Contact mode of 5th contact

Contact mode of 6th contact

Contact mode of 7th contact

Contact mode of 8th contact

Contact mode of 9th contact

Contact mode of 10th contact

Person ever refused interview

Person refused two times

Appointment scheduled

Appointment broken

Date of 1st contact

Date of 2nd contact

Date of 3rd contact

	Date of 4th contact
	Date of 5th contact
	Date of 6th contact
	Date of 7th contact
	Date of 8th contact
	Date of 9th contact
	Date of 10th contact
EPBG	Appointment scheduled
	Average likelihood to cooperate
	Minimum likelihood to cooperate
	Maximum likelihood to cooperate
	SD of likelihood rating
	Calls per case with likelihood rating
	Share of attempts on weekends
	Share of attempts before 10am
	Share of attempts between 10am and 5pm
	Share of attempts between 5pm and 8pm
	Share of attempts after 8pm
	Stratum
	Always same interviewer
	Share of different interviewers
	Number of calls
	At least one contact without realization
	Gender
	Birthday
	Call accepted
	Call not accepted, busy
	Call not accepted, not connected
	Call not accepted, timeout or no answer
	Call not accepted, port is not reached
	German nationality
	Education
	Case needed phone research after delivery to LINK

Ever did vocational training
Vocational training started before 1999
Ever did a minijob
Ever did an internship
Ever worked full time
Ever worked part time
Recent employment not full time
Recent employment not part time
Received unemployment benefit II in 2010
Received unemployment benefit I in 2010
Mean income in last full time spell (in Euros)
Mean income in part time spell (in Euros)
Mean income in internship spell (in Euros)
Mean income in vocational training spell (in Euros)
Mean income in minijob spell (in Euros)
Mean income in other spell (in Euros)