DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 31–59, http://dx.doi.org/10.1515/JOS-2015-0002

Validating Sensitive Questions: A Comparison of Survey and Register Data

Antje Kirchner[1]

This article explores the randomized response technique (RRT) – to be specific, a symmetric forced-choice implementation – as a means of improving the quality of survey data collected on receipt of basic income support. Because the sampled persons in this study were selected from administrative records, the proportion of respondents who have received transfer payments for basic income support, and thus the proportion of respondents who should have reported receipt is known.

The article addresses two research questions: First, it assesses whether the proportion of socially undesirable responses (indication of receipt of transfer payments) can be increased by applying the RRT. Estimates obtained in the RRT condition are compared to those from direct questioning, as well as to the known true prevalence. Such administrative record data are rare in the literature on sensitive questions and provide a unique opportunity to evaluate the 'more-is-better' assumption. Second, using multivariate analyses, mechanisms contributing to response accuracy are analyzed for one of the subsamples.

The main results can be summarized as follows: reporting accuracy of welfare benefit receipt cannot be increased using this particular variant of the RRT. Further, there is only weak evidence that the RRT elicits more accurate information compared to direct questioning in specific subpopulations.

Key words: Randomized response technique; social desirability; validation data; welfare receipt; unemployment benefit II.

1. Introduction

Surveys that collect data on welfare and unemployment receipt often find that respondents underreport this kind of information. In German surveys the known extent of underreporting of receipt of basic income support, a form of means-tested social security payment called 'Unemployment Benefit II' (UB II), ranges between 9 percent and 17 percent (Kreuter et al. 2010, 2014). One potential motivation for underreporting might be the sensitive nature of the topic: by underreporting, respondents avoid interviewer disapproval, embarrassment, and answer in a socially desirable manner (Tourangeau and Yan 2007). The main question the following paper addresses is whether alternative questioning formats, such as the randomized response technique (Warner 1965), can be

[1] Institute for Employment Research (IAB), Regensburger Str.104, Nuremberg 90478, Germany. Email: antje.kirchner@iab.de

Acknowledgments: This study is part of a larger research project carried out with M. Trappmann, I. Krumpal and H. v. Hermanni. It has been published in part in my dissertation (Kirchner 2014). Data collection was supported by the Institute for Employment Research (IAB). I am also very grateful to S. Eckman, J. Korbmacher, F. Kreuter and the anonymous referees for their helpful comments and suggestions.

© Statistics Sweden

used to improve the response quality of data collected regarding welfare receipt in labor market surveys.

1.1. Background

While unintentional misreporting, for example due to recall error, can certainly be problematic in the reporting of social security receipt (Manzoni et al. 2010; Kreuter et al. 2014), special attention should be devoted to other causes of misreporting in interviewer-administered surveys. It can be reasonably assumed that survey respondents are more likely to conceal sensitive information in order to conform to perceived norms (Cialdini 2007). This, in turn, affects the validity of the prevalence estimates (Lee 1993): if this failure to report welfare receipt is systematically different for certain social groups, resulting parameter estimates, such as proportions, averages, as well as relationships between variables will be biased (Hausman 2001).

The level of 'threat' or 'sensitivity' of a question as perceived by the respondent can be established along three theoretical dimensions (Tourangeau and Yan 2007): intrusiveness, risk of disclosure and social desirability. Several of these apply to the receipt of basic income support: people apply for welfare benefits in Germany if they have been unemployed long-term or if they cannot sustain a living from their current job, that is, when the resulting income is below a legally defined threshold. Individuals receiving basic income support may not wish to report this information in a survey. Admitting to an interviewer that they either have not been able to find a job over a longer period, that they live in poverty or that they do not earn enough to support their families might be perceived as too embarrassing. The concept of 'injunctive social norms' (Cialdini 2007) – one's perception or expectation of what most others approve or disapprove of – plays a vital role in this context. Negative beliefs and prejudice about welfare recipients in the United States and Great Britain comprise anything from not being motivated enough to find a job, being uninterested in self-improvement and dishonesty, to laziness and dependence (Bullock 2006). The receipt of basic income support in Germany is associated with similar prejudice. It is thus considered socially undesirable in terms of the commonly perceived norm and negatively stigmatizing, causing embarrassment when admitting to such.

To avoid errors from (item) nonresponse and misreporting ('under-' as well as 'overreporting') due to the sensitive nature of a question, survey methodologists have suggested a range of guidelines with respect to the design of a questionnaire (for an extensive overview, see Lee 1993; Bradburn et al. 2004; Tourangeau and Yan 2007). Indirect surveying techniques, such as the randomized response technique (RRT), are strategies to reduce underreporting (Lensvelt-Mulders et al. 2005). The RRT method was originally developed by Warner in 1965 to reduce response bias arising from privacy concerns. Ever since its first implementation, the RRT has been refined in many different variants (Horvitz et al. 1967; Greenberg et al. 1969; Boruch 1971; Greenberg et al. 1971; Moors 1971; Kuk 1990; Mangat and Singh 1990; Mangat 1994). Warner's original design, the so-called unrelated question techniques, forced-response designs, Moor's design, as well as Kuk's or Mangat's variants are probably among the best-known RRT designs (for an overview of different RRT designs, estimators and applications, see Fox and Tracy 1986; Umesh and Peterson 1991; Lensvelt-Mulders et al. 2005; Lensvelt-Mulders

et al. 2005b, or Tourangeau and Yan 2007). More recent developments also account for the fact that respondents might still underreport sensitive attributes in the RRT and allow for an estimation of a so-called 'cheating' parameter (Clark and Desharnais 1998; Böckenholt et al. 2009; Van den Hout et al. 2010; Ostapczuk et al. 2011; De Jong et al. 2012).

1.2. The General Idea of the RRT

The main idea, common to all RRT variants, is to conceal a respondent's answer by using a randomizing device (e.g., coins, cards, dice, spinner), the outcome of which is only known to the respondent (Fox and Tracy 1986). In its original implementation (Warner 1965), survey respondents are – depending on the outcome of the randomizing device – directed to answer one of two logically opposing statements, such as: 'I am a recipient of unemployment benefits II,' or 'I am not a recipient of unemployment benefits II.' Respondents only answer 'Yes' or 'No' without revealing which statement they were directed to reply to. Due to this chance element, neither the interviewer nor the researcher can infer anything regarding the respondent's true status from the response given. Since the randomization mechanism – and thus the probability distribution of the misclassification – is known to the researcher, estimation of the population prevalence of the sensitive characteristic under study is possible (Fox and Tracy 1986), as are regression analyses analyzing randomized response dependent variables (Maddala 1983, 54ff.; for an overview of estimators, see Tourangeau and Yan 2007). Granted that respondents understand and trust the method, the RRT should then increase reporting accuracy and reduce measurement error resulting from social desirability concerns.

Lensvelt-Mulders et al. (2005) distinguish two main types of studies in order to assess the performance of the RRT compared to that of other techniques: comparative and validation studies. The first type of study is most commonly found when evaluating the RRT. It compares estimates derived by means of RRT to those obtained by means of standard direct questioning. The RRT is – or more generally indirect techniques are – then assumed to outperform direct questioning if it elicits higher prevalence estimates for questions that are assumed to be subject to underreporting. Researchers generally refer to this as the 'more-is-better' assumption (for an overview of studies relying on this assumption, see Umesh and Peterson 1991; Lensvelt-Mulders et al. 2005; Tourangeau and Yan 2007). These studies often use a split-ballot design, randomly assigning participants of a given survey to either direct questioning or RRT. From a validation perspective, studies relying on the more-is-better assumption provide the weakest form of validation (Moshagen et al. 2014). Alternatively, estimates from other sources in which the prevalence of the sensitive trait is known only for the population, or parts thereof, but not for the sample, can be used as a benchmark for comparison (Moshagen et al. 2014). The authors refer to this as an intermediate form of validation and point out that potential differences might be confounded with sampling bias.

In some rare instances, researchers have access to additional, auxiliary information on the subjects of investigation for evaluation of the RRT performance (for an overview, see Lensvelt-Mulders et al. 2005; Wolter and Preisendörfer 2013). These studies are henceforth referred to as validation studies. Validation studies provide a stronger form of

performance assessment compared to comparative studies (Lensvelt-Mulders et al. 2005). In general, two types of validation studies can be distinguished: those validating responses at the individual level and those validating responses, or rather estimates, at the aggregate level (for the same sample).

While the most powerful validation of a survey response can be achieved if a 'gold standard' or the 'true' response of a respondent is available at the individual level (Groves 1989), often this information is impossible to acquire, too costly or (legally) not accessible. However, if individual-level validation data is available, it provides a valuable resource for analyzing individual motivations that contribute to misreporting which otherwise would not be possible. The second, somewhat weaker form of validation compares RRT survey estimates to aggregate data. This information might be data that is available for certain population segments of the sample using records (such as criminal statistics) or information that is available on the sampling frame.

Many empirical studies have evaluated whether the RRT method is in fact better at eliciting reports of sensitive behavior than the direct questioning methods. In the most recent meta-analysis (Lensvelt-Mulders et al. 2005), a total of six individual-level RRT validation studies as opposed to 32 comparative RRT studies were investigated. In general, the RRT still produced some response error, albeit lower than a comparable standard face-to-face questioning: for the validation studies under investigation, in the RRT condition the mean response was underreported by 38 percent, while in the traditional face-to-face condition mean underreporting was 42 percent. One of these validation studies, conducted by van der Heijden and colleagues (2000; see also Lensvelt-Mulders et al. 2006), tested two different implementations of the RRT, a forced-response implementation and Kuk's method, against standard face-to-face questioning. Results suggest that both RRT versions yield significantly lower response error with respect to social security fraud. Other experimental studies without validation data (comparative studies based on the more-is-better assumption) also showed that the RRT increased the validity of the estimates by eliciting more truthful responses (e.g., Weissman et al. 1986; Lara et al. 2004; Lara et al. 2006).

In general, the RRT seems to elicit more honest answers and reduce social desirability bias, especially when dealing with more sensitive questions (Fidler and Kleinknecht 1977; Landsheer et al. 1999; Lensvelt-Mulders et al. 2005). For example, the pioneering validation study by Locander and colleagues (1976) relying on individual-level validation data for some items shows, that the response error for RRT is (significantly) lower compared to that of direct questioning in three out of five instances (voter registration, bankruptcy involvement, and drunken driving). While the trend – that is, the RRT eliciting higher prevalence estimates – is as expected in most validation studies on topics such as failing course grades, arrests per person or criminal convictions, some validation studies also find no significant difference between RRT and direct questioning or contrary evidence (Locander et al. 1976; Lamb and Stem 1978; Tracy and Fox 1981; Wolter and Preisendörfer 2013). More recently, other comparative experimental studies have been published questioning the validity of RRT estimates (Umesh and Peterson 1991; Holbrook and Krosnick 2010; Coutts and Jann 2011; Coutts et al. 2011; Höglinger et al. 2014). Those studies show that the RRT does not provide more valid prevalence estimates compared to direct questioning, and that the RRT provides impossible, out-of-range

estimates (Holbrook and Krosnick 2010; Höglinger et al. 2014), suggesting noncompliance with RRT instructions.

1.3. Research Objectives

The following article presents one of the few large-scale RRT validation studies using administrative record data. More precisely, it explores whether the RRT is successful in eliciting higher-quality responses regarding the receipt of basic income support. Drawing on survey data collected in a nationwide telephone survey in Germany in 2010, respondents were randomly assigned to one of two techniques: either randomized response technique or traditional direct questioning. Using administrative record data, the true percentage of respondents who have received transfer payments for basic income support and thus the percentage who should have reported receipt is known. This allows a validation of the reported percentage against the known true rate for the responding cases, hence assessing the bias of the estimates. Such administrative record data is quite rare in the literature on sensitive questions (Lensvelt-Mulders et al. 2005; Wolter and Preisendörfer 2013).

The study contributes to the existing RRT research and response bias in several ways: to the best of the author's knowledge, the performance of the RRT in a telephone survey has not been validated against external data (especially not with respect to the receipt of basic income support). All existing RRT validation studies implemented the RRT method in a face-to-face mode (comparing the technique with face-to-face and other modes) but never in a pure telephone setting (cf. also Lensvelt-Mulders et al. 2005; Wolter and Preisendörfer 2013). The choice of a telephone mode, however, might be perceived as more private by respondents, thus leading to more honest answers due to greater perceived social distance (Holbrook et al. 2003). While collecting data by means of the RRT has many advantages, RRT procedures also suffer from considerable disadvantages compared to direct questioning: for one, a larger sample size is needed to achieve the same statistical power (Warner 1965); second, interview duration increases due to an explanation of the application of the procedure; while third, the cognitive burden placed on respondents is higher. Validating the functioning of a telephone implementation of the RRT might prove useful, given that it is more cost efficient compared to face-to-face surveys. The study thus follows the recommendation by Lamb and Stem (1978, 617) that "each time the [RRT] method is changed or used in a different setting, further evaluation is appropriate." Furthermore, this article contributes to the literature by investigating which individual-level factors influence accurate reporting and whether these mechanisms differ across experimental conditions.

To summarize, this article addresses two research questions:

1. Can item-specific response bias in interviewer-administered telephone surveys be reduced when using the randomized response technique? This is achieved by comparing the RRT estimates with a) the true value from the administrative data and b) direct questioning (DQ) obtained from the survey data.
2. Which covariates influence response error and can the RRT contribute to diminishing response error due to perceived sensitivity?

The remainder of this article is organized as follows: Section 2 describes the experimental design, the available data, as well as the method of analysis. The results of

the experiment are found in Section 3 and the conclusions and limitations of the study in Section 4.

2. Data and Methods

The nationwide telephone survey was commissioned by the Institute for Employment Research (IAB), the research institute of the Federal Employment Agency (FEA), and was carried out by the ForschungsWerk institute from October to December 2010. This study was approved by an internal review board as part of a study investigating undeclared work (Kirchner et al. 2013). The main focus on undeclared work had design implications regarding the choice of the misclassification probabilities for the RRT design in the current study on welfare benefit receipt (see below). Due to the particular sampling design, in addition to these survey data, supplementary information is available on the sampling frame (administrative records). The combination of both data sources allows addressing the research questions stated above. The next section provides an overview of the survey data, the administrative data, the combined data, and lays out the methods of analysis.

2.1. The Survey Data

2.1.1. Sampling and Data Collection

The survey is a dual-frame survey, using two sampling frames that are maintained by the FEA. These frames consist of all registered unemployment benefit (II) recipients as well as all employed persons.

The first random sample was drawn from the FEA registers of basic income support recipients (IAB Unemployment Benefit II History (LHG) V6.03.01 and (XLHG) V01.06.00-201007). It consists of people aged 18 to 64 who were known to have received basic income support in June 2010 (henceforth referred to as UB II or benefit recipients sample). The second random sample was drawn from the register of employees that is maintained by the FEA (IAB Employment Histories (BeH) V08.04.00, Nuremberg 2010). It consists of people aged 18 to 70 who were employed in December 2009 (henceforth referred to as employee sample). For both samples the latest available registers were used.

The registers contain telephone numbers for many of the sampled individuals. Whenever there was no information available on either of the frames, an extensive telephone number research was conducted, resulting in 91.7 percent (UB II sample) and 68.2 percent (employee sample) coverage. All individuals selected into the sample received a personalized advance letter announcing the survey. During fieldwork, some of the telephone numbers turned out to be invalid. This resulted in effectively 75.8 percent (UB II sample) and 53.5 percent (employee sample) cases with working numbers. Of those cases approximately 26 percent agreed to participate in the survey. Overall 3,211 interviews were completed (UB II: 18.8 percent and employees: 16.3 percent RR1, AAPOR 2011).

2.1.2. Experimental Design and Measurement of the Dependent Variable

Individuals who were initially selected into the sample were randomized in advance into two experimental groups. To achieve approximately the same level of statistical precision

in the RRT condition as in the direct questioning condition (DQ), individuals were randomly assigned with a ratio of 2:1 (Warner 1965; Cohen 1988; Lensvelt-Mulders et al. 2005b). The unequal assignment to the experimental conditions is necessary due to the additional random noise component in the RRT.

Based on the administrative data, regression analyses were conducted for the gross sample, showing that randomization into experimental groups was successful. This approach resulted in 1,145 completes in the DQ condition and 2,066 in the RRT condition. Table 1 provides an overview of the assignments to the experimental conditions.

Of the respondents originally assigned to the RRT, 13.2 percent refused the application of the randomized response technique (DQ_RRT) and were subsequently asked to respond to the relevant survey questions directly ($n = 274$). Results from a multiple logistic regression model, not presented here, modeling refusal to comply with the randomization protocol (DQ_RRT) show that two variables in particular have a large, statistically significant effect and predictive power: poor language skills and whether a respondent refused to answer the question on household income both substantially increase the probability of a refusal. Refusal is also higher in the UB II sample, among younger and single respondents, among respondents who have never held a job before, and respondents with a lower socioeconomic status. Further analyses indicate that both splits do not differ with respect to gender, formal training, older age groups, a previous socially undesirable response, composition of social networks, various attitudes towards undeclared work – the focus of the original study – or region of residence. Given these results, all further analyses will also be conducted separately.

The survey instrument was fully standardized: All survey participants received identical instructions with respect to the voluntary nature of the survey, the survey topic, assurances of confidentiality and anonymity, definitions or further explanations regarding receipt or UB II if needed. The only differences are within the experimental splits.

Across the two samples, two different operationalizations were used: for the UB II sample – known to have received benefits in June 2010 – participants were asked to report any 'benefit receipt ever'. In the employee sample participants were asked to report receipt in 'September 2010'. While these different operationalizations guarantee that (aggregate) responses can be validated, another criterion was to keep the questions as simple as possible in order to ensure understanding and correct recall (Tourangeau et al. 2000; Groves et al. 2009; Manzoni et al. 2010). To ease recall in the employee sample (and allow validation), the question relates to a defined period of receipt just prior to data collection. Further, all question formats were kept as similar as possible to commonly used questions in labor market surveys (cf. the PASS study as described by Trappmann et al. 2010).

Table 1. Experimental conditions

Assigned condition	N	Realized condition	N
DQ	1,145	DQ	1,145
RRT_assigned	2,066	RRT	1,792
		DQ_RRT	274

2.1.3. The RRT Implementation

Lensvelt-Mulders et al. (2005b) compared the efficiency of various RRT designs. The authors demonstrate that one variant, the so-called forced-choice RRT variant (Boruch 1971), was shown to be among the statistically most efficient RRT designs, is usually well understood (Landsheer et al. 1999; De Schrijver 2012) and shows higher rates of rule compliance compared to other RRT designs (Böckenholt et al. 2009).

In the symmetric forced-choice design, respondents are instructed to reply according to a set of rules: the randomization device determines whether the respondent is forced to answer 'Yes' (with probability p_1) – irrespective of their true status –, 'No' (with probability p_2) – irrespective of their true status –, or whether the sensitive question is to be answered truthfully, that is 'Yes' or 'No' (with probability p_3). The survey was designed to minimize two respondent hazards: neither a positive nor a negative answer should risk suspicion. The advantage of this so-called 'symmetric' variant of the forced-choice RRT is that it is shown to reduce respondents incentive to cheat in the RRT condition (i.e., provide a negative response when they should say 'Yes') and leads to greater rule compliance compared to an asymmetric variant that protects only singular responses (Ostapczuk et al. 2009). Regarding statistical efficiency, Lensvelt-Mulders et al. (2005b) recommend that the probability of providing a forced 'Yes' should be approximately the same as the expected prevalence of the sensitive item under investigation (Clark and Desharnais 1998), while the probability to tell the truth should be between 0.7 and 0.8.

Assuming that the probability distribution of the randomization procedure is known, the population prevalence as well as standard errors (s.e.) and confidence intervals for the forced-choice RRT can be estimated as follows (Fox and Tracy 1986): the observed sampling distribution of 'Yes' responses $\hat{\Phi}$ is used as an estimator for the unknown population parameter Φ. The overall proportion of positive responses (Φ) is the sum of the proportion of 'forced Yes' responses (p_1), and the product of the (unknown) population parameter π multiplied by the probability of having to respond truthfully (p_3): $\Phi = p_1 + p_3 * \pi$. The prevalence estimate of the sensitive characteristic $\hat{\pi}_{RRT}$ is then given as (Lensvelt-Mulders et al. 2005b):

$$\hat{\pi}_{RRT} = \frac{\hat{\Phi} - p_1}{p_3} \tag{1}$$

An estimate of the sampling variance of $\hat{\pi}_{RRT}$ is given as:

$$Var(\hat{\pi}_{RRT}) = \frac{\hat{\Phi} * (1 - \hat{\Phi})}{n * (p_3)^2} \tag{2}$$

where n is the sample size.

Regarding the administration of the forced-choice RRT over the telephone, the RRT design as developed by Krumpal (2012) was implemented and refined based on results of pretest interviews (cognitive pretest $n = 31$; pretest with the fully programmed instrument $n = 63$). Krumpal (2012) demonstrates that those instructions are well understood by respondents and elicit more undesirable responses yielding higher prevalence estimates of xenophobia and anti-Semitism in Germany.

More precisely, in the survey respondents in the RRT condition were asked first to gather three coins, paper and pencil in order to note down the rules. Respondents were then asked to flip the three coins prior to each question in the RRT section. Should a respondent accidentally reveal the outcome of the coin flip, interviewers were trained to ask respondents to flip the coin again without revealing the outcome. The exact rules implemented to provide an answer were the following (for the entire instructions see Appendix, translated from German):

> *". . . please, answer as follows: 3 tails, please always respond with 'Yes;' 3 heads, please always respond with 'No;' a mixture, that is a combination of heads and tails, such as 2 heads and 1 tail, please respond truthfully."*

(Note to the reader: Interviewers were trained to leave enough time 1) for respondents to note down the rules and 2) for respondents to toss the coins and possibly to consult their notes.)

It follows from this that $p_1 = 0.125$ 'forced Yes,' $p_2 = 0.125$ 'forced No,' and $p_3 = 0.75$ truthful response. The main interest of the original study was 'undeclared work' (see Section 2), with an assumed prevalence of about 10 percent to 12 percent in Germany. The probabilities of a forced 'Yes'/'No' and 'the truth' were chosen accordingly. Regarding the above mentioned recommendations, this design is not optimal with respect to the investigation of UB II receipt.

To ensure respondent understanding of the technique as stressed in Landsheer et al. (1999), a minimum of one 'training' example – in which the true answer had been reported by the respondent earlier in the questionnaire – was provided to everyone in this experimental condition so as to familiarize the respondents with the RRT (for the implementation of the training example, see Appendix). If this 'training example' was answered incorrectly, or the interviewer was under the impression that the technique had not been fully understood, another standardized example was provided to the respondent. Only when full understanding of the rules had been assured, did the main RRT section begin.

2.1.4. Independent Variables and Operationalizations

A range of indicators explaining underreporting of UB II will be analyzed in the scope of the second research question. Existing empirical evidence shows that underreporting of UB II is more frequent among males, among people aged 25 and younger as well as employed people (Kreuter et al. 2014). The authors also find a significant effect of recall period and household size. Those respondents with a longer recall period and those living in a larger household underreported more frequently. Household size in this particular instance is not to be taken literally: rather it is an indicator capturing a higher propensity to conduct the interview with someone less knowledgeable about the receipt of UB II, and thus response error should be larger. Kreuter et al. (2010, 2014) also show that respondents who are more reluctant to participate in a survey are slightly more likely to underreport benefit receipt. The authors attribute this effect to a lower motivation of these respondents while controlling for sample composition and recall error due to a longer recall period. Both studies mentioned above only applied direct questioning techniques.

Drawing on main insights of these studies, as well as on behavioral theories and the response process (Tourangeau and Rasinski 1988), variables that capture subjective costs, risks and utilities that are associated with accurate reporting of UB II will be included in the models. It can be reasonably assumed that significant (negative) effects regarding reporting accuracy in the model of the direct questioning split are observed for those characteristics that are associated with higher subjective reporting costs. These are higher if receipt of UB II is perceived as particularly sensitive, for example, when a respondent is employed.

Table 2 presents an overview of all independent variables. Factors contributing to perceived item sensitivity and hence associated reporting costs, comprise: employment status, occupational status, and a respondent's willingness to provide socially undesirable answers. Further, the reluctance of the respondents to answer sensitive questions is operationalized with an indicator variable, measuring item nonresponse for the item household income. Equally important is a measure of how common the receipt of UB II is in a respondent's environment: admitting to receiving UB II could then be perceived as less of a norm violation and reported more accurately. Ideally this indicator would be measured at the neighborhood level, which is not possible in this particular case due to data privacy issues. Thus, the recipient rate at the more aggregate municipal level is included in all models.

According to the work of Böckenholt and van der Heijden (2007), the RRT works especially well if the RRT instructions are clearly understood and the cognitive burden is kept as low as possible. A second set of indicators thus relates to the survey process and to the application of the RRT by the respondents. The first indicator captures whether a respondent was reluctant to cooperate in the RRT condition (DQ_RRT) and was then surveyed in the direct questioning mode. In order to capture understanding of the RRT, two proxy indicators are used (Landsheer et al. 1999): first, interviewers were asked to rate the language skills (German) of a respondent immediately following the telephone interview. A second indicator pertaining to the understanding of the RRT instructions is educational attainment (formal training). Response latency, that is, the speed at which a respondent answers, is used as a measure for response quality.

All models control for gender (0 male, 1 female), age (below 25, 25–40, 41–57, 58 and above), which region of Germany a respondent resides in (0 West, 1 East) and single-person household (0 multi-member household, 1 single-person household). Including these controls seems appropriate given respondents refusing to stay in the assigned RRT condition and the assumed differential underlying mechanisms in both experimental groups.

2.2. *Register Data*

The analysis uses supplementary register data based on social security reports and reports from the FEA itself as gold standard. Information relating to basic income receipt is a by-product of the FEA activities, that is, process data generated from information provided by the applicants during the application process. This information, such as household composition or income, is used to evaluate entitlement to receive UB II. These de-identified basic income receipt records are accessible to researchers at IAB.

Table 2. Description of variables used in the multivariate analyses

Indicator	Description	
Factors contributing to perceived reporting costs and item sensitivity		
Employment status	At the time of survey	
	0	Not employed (unemployed, parental leave, student etc.)
	1	Marginally employed with income up to 400€
	2	Employed with labor income >400€
Occupational status	International socioeconomic index of occupational status (ISEI) (Ganzeboom et al. 1992). Coded based on ISCO88 of present or last job (Hendrickx 2002)	
	0	No ISEI available, that is, never held a job before (score =.)
	1	Low or medium ISEI of present or last job (score 16–43)
	2	High ISEI of present or last job (score >43)
Socially undesirable response	Socially undesirable response regarding tax honesty. Tax honesty is:	
	0	Absolutely worthwhile, worthwhile
	1	Not worthwhile, absolutely not worthwhile
Reluctance	Item nonresponse for household income	
	0	Substantive response
	1	Missing response
Recipient rate	Share of UB II in municipality	
Survey process and application of RRT		
RRT refusal (DQ_RRT)	0	RRT condition
	1	DQ_RRT condition
Language skills	Scale from 1= very good to 6= nonexistent (recoded 0,1)	
	0	Good (<3)
	1	Poor (>=3)
Formal training	0	Secondary degree and below
	1	Tertiary degree
Response latency	Standardized response time in experimental section (recoded according to quartiles)	
	0	Slow response ($<Q_{25}$)
	1	Mean response ($Q_{25}-Q_{75}$)
	2	Fast response ($>Q_{75}$)

For the analyses, only one indicator in these records is of relevance: whether an individual received UB II. As a general rule, all data relevant to payments and claims (taxes, pensions, unemployment benefits etc.), that is, the primary use of the social security system, are known to be of very good data quality (Jacobebbinghaus and Seth 2007). The analyses thus rest on the crucial assumption that the true value of the respondents can be

captured with these data. The UB II receipt indicator is known to be both accurate and complete and can serve as gold standard.

2.3. Combined Data

Since respondents were not asked for consent to link their survey data to the administrative data, the two data sources cannot be merged at an individual level.

However, due to sampling on the dependent variable (known as reverse record check studies; Groves 1989), each individual in the UB II sample should by default respond with 'Yes' to the 'benefit receipt ever' question. Overreporting is not possible by definition. With the true aggregate prevalence being 100 percent, an indicator variable can be created on the individual person level that captures whether an individual reported accurately without linkage of the two data sources. This measure of reporting accuracy is a binary variable that takes on the value 1 if the survey report matches the true value in the administrative records, and 0 if the survey report is 'No,' that is, a mismatch between the survey data and the administrative records. Item nonresponse is equally spread across all experimental conditions (three out of 1,598 respondents). Those cases are excluded from the analyses.

For the employee sample, the missing linkage consent question only allows an assessment of the first research question. Since it is not possible to link the survey data to the respective administrative records, it is impossible to construct a variable indicating reporting accuracy at an individual level. However, it is possible to derive and compare aggregate measures for respondents. According to the administrative data, the true aggregate prevalence of 'benefit receipt in September 2010' for respondents of the employee sample is 3.0 percent in the DQ condition and 4.2 percent for the RRT_assigned condition. Only the original assignment (DQ or RRT_assigned) and the response indicator can be used to obtain these true values. This is why RRT and DQ_RRT cannot be separated and have identical values. In the employee subsample, overreporting could theoretically be an issue. However, it seems unreasonable to assume that respondents, aside from overreporting due to satisficing or acquiescence (Krosnick 1991), would (consciously) overreport UB II receipt. Item nonresponse occurred once in the DQ condition (out of 1,613 respondents).

Due to the above mentioned limitations in the employee sample, the second research question can only be addressed using the UB II sample.

2.4. Statistical Analyses

The response bias is used to assess the impact of measurement error from the two alternative techniques of data collection. The bias of a statistic is simply the difference between the statistic's expectation and the true population value. The estimator of the response bias (B_j) in the respective experimental condition j is thus (adapted from Biemer 2010, 49):

$$B_j = \bar{y}_{j,svy} - \bar{y}_{j,adm} \tag{3}$$

which is the difference of the means of accurate reporting in the sample survey measurements ($\bar{y}_{j,svy}$) and the gold standard measurements ($\bar{y}_{j,adm}$). This approach will then

allow for a comparison of the overall response bias of the RRT and the DQ in both subsamples using one-sided unpaired t-test assuming unequal variances.

Subsequent to analyzing the overall bias for both samples (research question 1), logistic regression models will be used to model accurate reporting by experimental condition as a function of covariates for the UB II sample (research question 2). Again, the dependent variable Y_{ij} represents an individual's (i) response behavior (0 underreporting, 1 accurate reporting) in the experimental condition j. If the assumptions of privacy protection in the RRT condition hold, predictors related to perceived item sensitivity in the DQ model should be more positively related to accurate reporting. While for the direct condition a logistic regression model is appropriate, the RRT requires a logistic regression with an adapted likelihood function that accounts for the additional noise introduced by the RRT procedure, such as rrlogit (Jann 2011).

3. Empirical Results

Table 3 shows the prevalence estimates in percent for all experimental groups across both subsamples ($\bar{y}_{j,svy}$), the resulting response bias estimates (B_j in %pts) as well as the difference in biases ($B_{DQ} - B_j$ in %pts). Estimates presented in column 'RRT_assigned,' are based on the logic of 'intention-to-treat' analysis (Angrist et al. 1996). They provide a more conservative estimate of the average treatment effect of assignment. The last two columns take into account whether respondents actually received treatment – that is the RRT – or refused its application: 18 percent of the UB II respondents and 9 percent of the employee sample did not follow the randomization protocol. Since the exact questions asked in the survey differ across the two subsamples, response bias estimates are not comparable across subsamples and should be interpreted individually. The estimated response bias pointing in the expected direction is boldfaced, indicating a statistically significant amount of underreporting.

Replicating results from prior studies (Kreuter et al. 2010, 2014), receipt of benefit is underreported in both DQ conditions: for the UB II sample benefit receipt is underreported by 13.0 percentage points. While receipt of benefits is also underreported by 0.9 percentage points in the employed sample, this result is statistically nonsignificant. In absolute terms, the bias is larger in the UB II sample; in relative terms, standardized on the value of true prevalence, it is much larger in the employed sample (29.3% compared to 13.0%). However, these differences could be confounded by the fact that the question of receipt 'ever,' in the UB II sample, as opposed to 'September,' in the employee sample, might be perceived as less difficult or less sensitive by the respondents.

3.1. Reduction of Response Bias by Means of RRT?

Assuming that bias is solely due to item sensitivity and that the RRT can alleviate this bias, the RRT survey data estimates in Table 3 – granted that the RRT is understood and trusted – should not diverge significantly from the gold standard.

Contrary to the initial expectations, the response bias in the RRT_assigned condition differs significantly from zero. In the UB II sample, receipt of welfare benefits is underreported by 12.7 percentage points and, in the employee sample, by 1.9 percentage points. As for the DQ condition, the relative bias is larger in the employee sample (45.7%)

Table 3. (Estimated) proportions ($\bar{y}_j$), absolute response bias (B_j) and differential response bias ($B_{DQ} - B_j$)

Sample type	Estimate/statistic	DQ	RRT_assigned	RRT_assigned	
				RRT	DQ_RRT
UB II	$\bar{y}_{j,adm}$	1.000	1.000	1.000	1.000
	$\bar{y}_{j,svy}$	0.870	0.873	0.854	0.906
	B_j (s.e.)	**−0.130** (0.014)	**−0.127** (0.014)	**−0.146** (0.020)	**−0.094** (0.022)
	t-statistic	−9.274	−9.045	−7.465	−4.321
	$B_{DQ} - B_j$ (s.e.)		−0.003 (0.020)	0.016 (0.024)	−0.035 (0.026)
	t-statistic		−0.140	0.683	−1.353
	sample size (n)	579	1,016	836	180
	effective n	579	650	470	180
Employee	$\bar{y}_{j,adm}$	0.030	0.042	0.042	0.042
	$\bar{y}_{j,svy}$	0.021	0.023	0.004	0.043
	B_j (s.e.)	−0.009 (0.006)	−0.019 (0.014)	**−0.038** (0.014)	0.001 (0.021)
	t-statistic	−1.449	−1.393	−2.659	0.049
	$B_{DQ} - B_j$ (s.e.)		0.010 (0.015)	0.030 (0.016)	−0.010 (0.022)
	t-statistic		0.689	1.887	−0.447
	sample size (n)	564	1,048	955	93
	effective n	564	630	537	93

compared to the UB II sample (12.7%). Conducting separate analyses for those respondents who complied with the randomization protocol and those who did not, response bias for the RRT is larger for the former group in both samples (UB II sample: 14.6%pts vs. 9.4%pts; employee sample: 3.8%pts vs. 0.1%pts). Respondents who refused to apply the RRT are the ones who show the lowest levels of underreporting in both subsamples across all experimental conditions and thus seem to be the more accurate respondents (also in relative terms: 9.4% and 2.3%).

Furthermore, the RRT estimates should be less biased compared to those in the DQ condition ($B_{DQ} - B_j$ in %pts), resulting in a negative difference. The difference in response bias estimates in the UB II sample is statistically nonsignificant across all conditions: the response bias is 0.97 times smaller in the RRT_assigned condition compared to the DQ condition, 1.13 times higher for RRT and 0.73 times smaller for DQ_RRT. In the employee sample, the differences are nonsignificant as well: the response bias is 2.12 times higher in the RRT_assigned condition compared to direct questioning and 0.12 times smaller for DQ_RRT. Contrary to the expectations, it is significantly larger in the RRT condition (4.35, $p = 0.03$).

To summarize some of the results for the initial research question: 1) the particular forced-choice telephone implementation of the RRT cannot reduce bias in the estimated prevalence of basic income support in Germany, while 2) the RRT performs significantly worse if the item under investigation is of a low prevalence rate, as in the case of the employee sample. Furthermore, due to the random noise in the RRT condition, variance estimates are inflated by a factor of $1.\bar{7}$ or, put differently, the effective sample size is reduced accordingly. All other things being equal, this leads to an increased mean squared error (MSE) in the RRT condition. The MSE estimate in the UB II DQ condition is 0.13 and 0.02 in the employee sample. Assuming identical sample sizes in both conditions, namely those of the respective DQ split, the MSE in the UB II RRT condition would then be 0.28 and 0.19 in the employee sample. Since the actual sample size in the RRT splits is larger, MSE estimates are 0.20 in the UB II sample and 0.11 in the employee sample.

One can only speculate about the reasons for the poor performance of the RRT in this particular study. One reason might be that the initial assumption – that unemployment benefit receipt is sensitive – is false. In that case, one would not expect to see the RRT producing estimates closer to the truth compared to direct questioning. The second argument might be that respondents do not apply the randomization procedure correctly, that is, that either they do not flip coins at all or they do not adhere to the RRT instructions (Clark and Desharnais 1998). In the first instance this could mean that a face-to-face implementation, with an interviewer supervising the randomization procedure, could perform better. The second issue is trust in the method: despite understanding the method, it is also crucial that respondents trust the privacy protection provided by the RRT (Holbrook and Krosnick 2010; Coutts and Jann 2011). While it can be reasonably assumed that unintentional noncompliance with the rules, that is, respondents accidentally providing a wrong answer, should not occur if the method is understood, nevertheless trust is essential. Respondents might consciously decide to edit their answers and ignore the RRT instructions if they lack trust: they might respond 'No' even if the randomization device prompted them to answer 'Yes' (cheating). Or, if prompted to answer truthfully, respondents might edit their answer and report a 'No' (even if the truth is 'Yes'), resulting

in underreporting. These so-called 'cheaters' and 'under reporters' lead to the fact that the RRT estimates are biased (see also Boeije and Lensvelt-Mulders 2002; Böckenholt et al. 2009; Coutts and Jann 2011; Ostapczuk et al. 2011). Specific to our study, there is a unique, indirect method of assessing the amount of cheating and underreporting relying on a few assumptions: 1) overreporting (incl. false positives in the employee sample) does not occur, 2) both effects are homogeneous across samples. Equation $\Phi_{sample} = p_1*(1-c) + p_3*(1-u)*\pi$ introducing a cheating as well as an underreporting parameter is then identifiable. Estimates of cheating in the RRT condition of this study amount to 18.4 percent and underreporting of 11.5 percent. These results also underline the utility and necessity of designs that allow for an estimation of and correction for cheating (Clark and Desharnais 1998; Böckenholt et al. 2009; Van den Hout et al. 2010; Ostapczuk et al. 2011; De Jong et al. 2012). These designs typically allow the identification and estimation of a cheating parameter by assigning two different misclassification probabilities to different RRT subsamples. The particular design of this study was chosen due to a successful prior implementation in the study conducted by Krumpal (2012), considerations of a loss in statistical efficiency and the proposed indirect estimation strategy. A third reason for the poor performance of the RRT could be the mode of data collection via telephone itself. Respondents might find it easier to 'cheat' on the phone than in a face-to-face mode (De Leeuw and van der Zouwen 1988; Aquilino 1994).

This result is particularly relevant for future studies due to the cost implications: the increased costs in the RRT condition are due to – all other things being equal – a larger sample size, longer interview durations (the RRT section was on average six minutes longer than DQ; see also Wolter and Preisendörfer 2013), statistically more complex analyses, more intensive interviewer training and, most important, a higher respondent burden. Given the empirical evidence, the additional costs of a forced-choice RRT data collection for welfare receipt are not justified. Thus, in terms of bias versus efficiency, these results clearly favor direct questioning to collect data on welfare benefit receipt in Germany.

3.2. *Is Response Bias Subgroup Specific?*

Contrary to the expectations in both experimental conditions, the results for research question 1 indicate a tremendous amount of misreporting.

The following section will analyze response error between subgroups while controlling for a differential sample composition across both experimental conditions. Since individual-level data is available only for the UB II sample, further analyses are limited to this sample and inferences can only be drawn with respect to this specific population. The dependent variable, 'accurate reporting,' will be modeled separately as a function of several individual characteristics for respondents in the UB II sample for each experimental split. In order to account for potential nonlinear relationships, all variables enter the regression equation categorically.

Table 4 displays the average marginal effects (AME) from logistic regression models (Stata version 12.1, rrlogit, Jann 2011), modeling accurate reporting as a function of the covariates mentioned above, as well as the difference in AMEs ($DQ - RRT\_assigned$). The AME is the average of discrete or partial changes over all observations. It yields a

Table 4. Logistic regression models analyzing accurate reporting of receipt of UB II (average marginal effects and 95% confidence intervals)

Y: Accurate reporting	Model 1: DQ	Model 2: RRT_as.	Difference: DQ − RRT_as.
Factors contributing to perceived reporting costs and item sensitivity			
No employment (ref. employed >400€)	**0.118*** **	**0.095** **	0.023
	[0.055;0.180]	[0.025;0.165]	[−0.071;0.117]
Marginally employed	0.017	0.013	0.004
	[−0.050;0.083]	[−0.056;0.082]	[−0.092;0.100]
Low/Med. ISEI (ref. n/a (never employed))	**0.069**$^{+}$	**0.063**$^{+}$	0.006
	[−0.000;0.137]	[−0.010;0.136]	[−0.094;0.106]
High ISEI	0.037	−0.003	0.040
	[−0.046;0.120]	[−0.098;0.092]	[−0.086;0.167]
Socially undesirable response (tax honesty)	**−0.045**$^{+}$	**0.158** **	**−0.203** **
	[−0.097;0.007]	[0.039;0.276]	[−0.333;−0.074]
Reluctance (item NR)	**−0.113** **	**−0.148*** **	0.036
	[−0.186;−0.039]	[−0.227;−0.069]	[−0.073;0.144]
Recipient rate	0.050	0.069	−0.018
	[−0.221;0.322]	[−0.241;0.379]	[−0.043;0.394]
Survey process and application of RRT			
DQ_RRT	–	0.049	–
	–	[−0.020;0.117]	–
Language skills (poor)	**−0.057**$^{+}$	−0.058	0.000
	[−0.120;0.005]	[−0.128;0.013]	[−0.094;0.095]
Tertiary degree	0.028	0.075	0.047
	[−0.072;0.128]	[−0.042;0.192]	[−0.201;0.107]
Fast response (ref. mean response)	0.038	−0.004	0.042

Table 4. Continued

Y: Accurate reporting	Model 1: DQ	Model 2: RRT_as.	Difference: DQ – RRT_as.
	[−0.030;0.106]	[−0.070;0.061]	[−0.052;0.137]
Slow response	0.011	0.042	−0.031
	[−0.049;0.071]	[−0.030;0.114]	[−0.125;0.062]
Controls			
Female	−0.001	**0.049**$^{+}$	**−0.050**$^{+}$
	[−0.051;0.049]	[−0.005;0.103]	[−0.124;0.023]
Age <25 (ref. 25 to 40)	**−0.137*****	**−0.168*****	0.031
	[−0.199;−0.076]	[−0.240;−0.097]	[−0.063;0.125]
Age 41 to 57	0.007	−0.028	0.035
	[−0.055;0.069]	[−0.109;0.053]	[−0.067;0.138]
Age >57	−0.008	−0.063	0.054
	[−0.122;0.105]	[−0.164;0.039]	[−0.098;0.207]
East Germany	0.009	0.032	−0.023
	[−0.050;0.069]	[−0.037;0.102]	[−0.114;0.068]
Single-person household	**0.092***	**0.071***	0.021
	[0.020;0.163]	[0.001;0.140]	[−0.079;0.121]
Model fit			
N	579	1,016	
LR Chi2 (df)	119.98 (17)	95.89 (18)	
Pseudo R2	0.27	0.09	
AIC	360.43	966.08	
BIC	434.58	1054.70	

95% confidence intervals in brackets; $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

straightforward interpretation of estimation results and effect sizes, and allows a comparison between models (Bartus 2005; Mood 2010). Subsequently, only AMEs will be reported.

Two models are presented in Table 4: Model 1 analyzes accurate reporting in the direct questioning condition and serves as a baseline for examining reporting accuracy. Model 2 replicates the same model in the RRT condition. I expect to see more accurate reporting in the RRT condition, especially for those variables related to perceived item sensitivity. Thus all (negative) effects related to item sensitivity that are found in the direct split should become more positively (or nonsignificantly) related to accurate reporting. This second model also presents insights regarding the question of which variables related to the survey process contribute to more accurate reporting.

Turning to the DQ Model 1, those variables related to perceived item sensitivity are of particular interest. Unconditional on other covariates, as expected, respondents with no current *employment* are on average 11.8 percentage points more likely than respondents with an income of 400 Euro and above to report receipt of UB II. Marginally employed respondents do not differ systematically from the reference category. Regarding *occupational status* respondents with a high (present or past) status are expected to report receipt of UB II less often than the other categories. Contrary to the initial expectations, respondents with a high ISEI have a slight tendency to report more accurately compared to the reference category (no job), while respondents with a low or medium status report receipt significantly more accurately (6.9%pts) than those who have never held a job before. Regarding the difference between respondents with a high ISEI and those with a low or medium ISEI, no significant difference is observed. The item '*socially undesirable response*' regarding tax honesty significantly explains accurate reporting, but in a surprising way: respondents with an honest, but more socially undesirable attitude towards tax dishonesty are on average 4.5 percentage points more likely to underreport the receipt of basic income support than those respondents displaying a more desirable attitude towards tax honesty. At first, this finding seems counterintuitive: responding in a socially undesirable manner in one instance would result in a higher propensity to admit another undesirable characteristic. One potential explanation could be that, given that 'tax dishonesty' is acceptable, misreporting on other characteristics is considered acceptable as well. *Reluctance* contributes significantly to the explanation of underreporting of UB II (11.3%pts). Regarding the *share of UB II* recipients at the municipality level, there is no significant effect, supporting the hypothesis regarding the wrong level of measurement.

Those characteristics relating to the survey process contribute less to the explanation of accurate reporting. Poor *language skills* are the only significant predictor contributing to underreporting of UB II (5.7%pts). With respect to the controls, younger respondents, aged 24 and below, significantly underreport receipt (13.7%pts). In line with expectations, the indicator 'single-person household' significantly improves reporting accuracy (9.2%pts). Both results support the argument that proxy reports with less knowledgeable persons on receipt of UB II are less accurate, since younger respondents are more likely to still live with their parents who apply for UB II for the entire household.

Turning to Model 2—the RRT model—the results are strikingly similar, both in direction and magnitude. Contrary to the expectations, variables related to perceived item

sensitivity exert approximately the same influence as in the DQ model with one exception: *socially undesirable response*. Respondents stating that tax honesty is (absolutely) not worthwhile report on average 15.8 percent more accurately in the RRT condition. This difference between both models is statistically significant, indicating that the RRT reduces social desirability concerns for those respondents ($p < 0.01$). Given this evidence, the above explanation for this finding seems implausible. A different explanation might help to solve the puzzle: in Germany, tax dishonesty is largely associated with undeclared work/income. Receipt of UB II is based on accurate reporting of all forms of income and misreporting of income to the authorities is heavily pursued. Stating that tax honesty is (absolutely) not worthwhile in the direct questioning condition might be considered indirect evidence for potential concealing of income when applying for UB II and is thus a highly sensitive question itself when confirming receipt of UB II. This would explain the negative relationship. This same question is potentially perceived as less intrusive in the RRT condition and hence respondents more openly state their opinion. The positive relationship in Model 2 is thus internally consistent.

To summarize, contrary to expectation, the RRT does not elicit more accurate reports for respondents for whom reports of UB II can be assumed to be particularly sensitive, with one exception. This indicates that the same misreporting mechanisms are at work in both experimental conditions.

Similar to Model 1, those characteristics relating to the survey process and the application of the RRT overall contribute less to the explanation of accurate reporting. Respondents who *refused the application of the RRT* report more accurately than those respondents in the RRT condition (4.9%pts). Anecdotal evidence from interviewer observations suggests that those respondents either distrust the RRT or claim that they 'have nothing to hide' and want to be questioned directly. The effect size of lack of *language skills* is negative and roughly the same as in Model 1; however, it just fails to be statistically significant ($p = 0.101$). It can be assumed that respondents who do not accurately understand what is asked of them in either condition (particularly so in the RRT) will not trust the method and therefore report (a 'self-protective' or 'nonincriminating') 'No' (Böckenholt et al. 2009; Coutts and Jann 2011). Thus the result is as expected for both models. Remember that while a *tertiary degree* contributes to accurate reporting (2.8%pts) in Model 1, in Model 2 this effect is larger in comparison to Model 1, but not compared to the reference category (7.5%pts). Due to the small number of people holding a tertiary degree, confidence intervals are rather large for this estimate. Further regression analyses were conducted but are not presented here: they account for the fact that if language skills are poor, neither educational degree will make a difference in the reporting accuracy. Assuming good language skills (essentially modeling an interaction), the results show a larger effect of university degree in Model 2. This suggests that the RRT reduces underreporting for these respondents: however, it remains unclear whether this effect is due to a better understanding of the RRT compared to the reference category (Poor German Skills and No Tertiary Degree) or the RRT guaranteeing anonymity and reducing item sensitivity for the more highly-educated group. *Response latency*, that is, the speed at which a respondent answers, is used as a measure for response quality. Surveying in the RRT condition by definition takes longer than a comparable direct question, since

respondents have to follow the RRT protocol. In theory, irrespective of the experimental condition, a longer answering process could indicate more editing of the true response and thus a poorer data quality (Holtgraves 2004). On the other hand, it could also be associated with higher-quality information and processing in the RRT condition (Wolter 2012). Results for response latency exhibit no clear pattern across models and are nonsignificant: in Model 2, a slower response indicates on average greater accuracy (4.2%pts; 0.4%pts more underreporting for fast respondents; this difference is statistically nonsignificant), while in Model 1, both fast and slow reporting is associated with greater accuracy compared to the reference category (3.8%pts and 1.1%pts).

With respect to the controls, effects are similar to those of Model 1, with the exception of women on average reporting more accurately in Model 2 (4.9%pts). The difference between both models is statistically significant ($p < 0.10$).

To summarize the results, results from previous studies (Kreuter et al. 2014) can be replicated in Model 1, that is, especially for characteristics relating to item sensitivity (employment status, occupational status, socially undesirable response, reluctance) and structural characteristics (age, single-person household). Contrary to the initial expectations, the RRT cannot resolve social desirability concerns for these items; as expected, structural influences persist. The hypotheses relating to the survey process and the application of the RRT cannot be confirmed with these results.

Analyzing DQ, RRT and DQ_RRT in one joint model while controlling for covariates shows that while RRT and DQ_RRT result in more accurate responses, these effects are statistically nonsignificant. A fully interacted model (all covariates and the RRT indicator) yields the following significant interaction effects: more accurate reporting by respondents with a socially undesirable response and those with a tertiary degree, as well as respondents taking longer to respond under RRT.

4. Discussion and Conclusion

The initial research question addressed the performance of a forced-choice telephone implementation of the RRT for the estimation of welfare receipt compared to direct questioning. The results show that this particular RRT design does not reduce underreporting in the data collection on welfare benefit receipt in a telephone survey. The RRT performs worse in the employee sample, where the overall prevalence is close to zero.

Insights into who underreports receipt of UB II were the main focus of the second research question. Inferences are limited to the population of UB II recipients in Germany. Reporting accuracy is significantly higher in both methods for respondents who perceive reporting of UB II as less of a norm violation, that is, respondents who are not employed. Respondents who admit to tax dishonesty report more accurately in the RRT model, but less accurately in the DQ model, as do respondents who are unwilling to provide information on other items such as income. Thus, there is a tendency for underreporting whenever receipt of welfare benefits is perceived as more sensitive in both models. If the RRT were to resolve the concerns of social desirability, differential effects would have been observed across both methods for those items capturing sensitivity. The results do not

support this argument: differences between models are statistically significant only for those respondents having given another socially undesirable response. Furthermore, it was expected that those items fostering understanding of the RRT would contribute to a higher reporting accuracy. While most effects point in the expected direction, they are statistically nonsignificant.

One can only speculate about the potential reasons for the failure of the RRT in this study. One argument discussed above relates to the potential lack of sensitivity of the item under study. If underreporting were not caused by perceived sensitivity, then the RRT would not be expected to decrease bias. Studies regarding the perception of welfare receipt would not support this argument (Bullock 2006). Other arguments explaining the poor performance of the RRT relate to 'cheating' and 'noncompliance' with the instructions of the RRT (Clark and Desharnais 1998; Böckenholt et al. 2009; De Jong et al. 2012). For one, it remains unclear whether respondents are really implementing the randomization procedure while on the telephone (Holbrook and Krosnick 2010). In that instance, a face-to-face mode might seem more appropriate. A second concern – which is more in line with the results – is that respondents 'forced' by the randomization device to provide a (false) positive answer might decide not to comply with the RRT rules (and reply 'No' instead of 'Yes') or underreport if asked to provide a truthful response (Böckenholt and van der Heijden 2007; Coutts and Jann 2011). This concern cannot be ruled out even in the face-to-face mode. However, it highlights the importance of RRT designs allowing for an estimation of underreporting and cheating, as prevalence estimates can then be corrected. The last argument pertains to the telephone mode itself: if the benefits of noncompliance are large and social control is weak, persons are less willing to comply (Böckenholt and van der Heijden 2007).

Overall, the finding that the (forced-choice variant of the) RRT still contains response bias has been confirmed by other recent studies (Holbrook and Krosnick 2010; Coutts and Jann 2011; Wolter and Preisendörfer 2013; Höglinger et al. 2014), but it is worth reiterating that the RRT does not outperform direct questioning (Lensvelt-Mulders et al. 2005). On the contrary, yielding approximately the same bias, mean squared error increased due to an inflated variance. Furthermore, there is a tremendous amount of visible refusal to follow the randomization protocol in the RRT condition as well as a large share of covert misreporting. Using this implementation of the RRT, the main implication is that the additional burden imposed on respondents in combination with additional surveying costs, for example in terms of sample size and duration, are not justified. Given that respondent burden is associated with a decreased probability of future survey participation and an increase in breakoffs, these results are particularly important. Overall, 95 out of 229 respondents broke off the interview during the RRT introduction or first item within the experimental section, while most of the 46 breakoffs in the DQ condition occurred either before or after the experimental condition, and none while asking about welfare benefits or undeclared work.

The evidence in this study also supports the notion that this particular RRT design performs slightly better in certain populations: those respondents with good language skills, those more highly educated, and those who take enough time to respond in the RRT condition, that is, the correct application of the randomization process being observed in some way. Furthermore, language skills and respondent reluctance are significant

predictors of whether respondents comply with the randomization protocol. When the research focus is on populations with a lower educational background, the results may thus be very different. The results and the tremendous amount of underreporting do not support the use of this implementation of the RRT in large-scale population surveys. Other techniques, such as the crosswise or triangular technique, a different variant of the RRT (Yu et al. 2008), might be a preferable method. These methods do not require a randomization device, are less of a cognitive burden for respondents, are easier to implement over the telephone, provide less incentives to misreport and might thus be a viable alternative to direct questioning (Jann et al. 2012; Korndörfer et al. 2014; Höglinger et al. 2014).

Appendix: RRT Introduction and Training Example

"I will now introduce you to a technique, that will allow you to keep your personal experiences anonymous by means of a coin flip. Even if this might sound strange to you, I kindly ask you to help us to try this new method. This method is scientifically approved and is fun. Would you please get a paper, a pencil, and three coins?

You will be able to answer all of the following questions either with 'Yes' or 'No.' Before answering each question, I would kindly ask you to flip the three coins. Please do not tell me the outcome of this coin flip. According to the outcome, please answer as follows:

- 3 tails; please always respond with 'Yes'
- 3 heads; please always respond with 'No'
- a mixture; that is, a combination of heads and tails, such as 2 heads and 1 tail, please respond truthfully

As you can see chance decides whether you actually respond to the question or provide a surrogate answer. Thus, your privacy is always protected. I, as the interviewer, will never know the result of your coin toss. Thus, I can never know, why you respond with 'Yes' or 'No.' Do you have any further questions regarding the technique?

Let us walk through one example together.

If you flip 3x heads, and I ask you if you are 18 years or older, what would you reply? *(Int: Pause; let the respondent reply first. 'No,' according to the rule)*

If you flip 3x tails, and I ask you if you are 18 years or older, what would you reply? *(Int: Pause; let the respondent reply first. 'Yes,' according to the rule)*

If you have a mixed result, for example, flip 2x heads and 1x tail, and I ask you if you are 18 years or older, what would you reply? *(Int: Pause; let the respondent reply first. The response has to be 'Yes' as part of the requirements of the sampling design)*

Do you have any further questions?"

(Note to the reader: If there were further questions, the rules were repeated and a new example provided before asking one question on UB II receipt followed by two questions on undeclared work.) (Translated from German)

5. References

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

Angrist, J.D., G.W. Imbens, and D.B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444–455. DOI: http://dx.doi.org/10.1080/01621459.1996.10476902.

Aquilino, W.S. 1994. "Interview Mode Effects in Surveys of Drug and Alcohol Use: A Field Experiment." *Public Opinion Quarterly* 58: 210–240. DOI: http://dx.doi.org/10.1086/269419.

Bartus, T. 2005. "Estimation of Marginal Effects Using Margeff." *The Stata Journal* 5: 309–329.

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

Boeije, H. and G.J.L.M. Lensvelt-Mulders. 2002. "Honest by Chance: A Qualitative Interview Study to Clarify Respondents' (Non-)compliance with Computer-Assisted-Randomized Response." *Bulletin de Methodologie Sociologique* 75: 24–39.

Boruch, R.F. 1971. "Assuring Confidentiality of Responses in Social Research: A Note on Strategies." *The American Sociologist* 6: 308–311.

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

Bullock, H.E. 2006. "Attributions for Poverty: A Comparison of Middle-Class and Welfare Recipient Attitudes." *Journal of Applied Social Psychology* 29: 2059–2082. DOI: http://dx.doi.org/10.1111/j.1559-1816.1999.tb02295.x.

Böckenholt, U., S. Barlas, and P.G.M. van der Heijden. 2009. "Do Randomized-Response Designs Eliminate Response Biases? An Empirical Study of Non-Compliance Behavior." *Journal of Applied Econometrics* 24: 377–392. DOI: http://dx.doi.org/10.1002/jae.1052.

Böckenholt, U. and P.G.M. van der Heijden. 2007. "Item Randomized-Response Models for Measuring Noncompliance: Risk-Return Perceptions, Social Influences, and Self-Protective Responses." *Psychometrika* 72: 245–262. DOI: http://dx.doi.org/10.1007/s11336-005-1495-y.

Cialdini, R.B. 2007. "Descriptive Social Norms as Underappreciated Sources of Social Control." *Psychometrika* 72: 263–268. DOI: http://dx.doi.org/10.1007/s11336-006-1560-6.

Clark, S.J. and R.A. Desharnais. 1998. "Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model." *Psychological Methods* 3: 160–168. DOI: http://dx.doi.org/10.1037/1082-989X.3.2.160.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Vol. 2. Hillshale, NJ: Erlbaum.

Coutts, E. and B. Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count

Technique (UCT)." *Sociological Methods & Research* 40: 169–193. DOI: http://dx.doi.org/10.1177/0049124110390768.

Coutts, E., B. Jann, I. Krumpal, and A.-F. Näher. 2011. "Plagiarism in Student Papers: Prevalence Estimates Using Special Techniques for Sensitive Questions." *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)* 231: 749–760.

De Jong, M.G., R. Pieters, and S. Stremersch. 2012. "Analysis of Sensitive Questions Across Cultures: An Application of Multigroup Item Randomized Response Theory to Sexual Attitudes and Behavior." *Journal of Personality and Social Psychology* 19: 153–176. DOI: http://dx.doi.org/10.1037/a0029394.

De Leeuw, E.D. and J. van der Zouwen. 1988. "Data Quality in Telephone and Face to Face Surveys: A Comparative Metaanalysis." In *Telephone Survey Methodology*, edited by R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg, 283–299. New York: John Wiley & Sons, Inc.

De Schrijver, A. 2012. "Sample Survey on Sensitive Topics: Investigating Respondents' Understanding and Trust in Alternative Versions of the Randomized Response Technique." *Journal of Research Practice* 8: 1–17.

Fidler, D.S. and R.E. Kleinknecht. 1977. "Randomized Response versus Direct Questioning: Two Data-Collection Methods for Sensitive Information." *Psychological Bulletin* 84: 1045–1049. DOI: http://dx.doi.org/10.1037/0033-2909.84.5.1045.

Fox, J.A. and P.E. Tracy. 1986. *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills: Sage Publications.

Ganzeboom, H.B.G., P.M. De Graaf, and D.J. Treiman. 1992. "A Standard International Socio-Economic Index of Occupational Status." *Social Science Research* 21: 1–56. DOI: http://dx.doi.org/10.1016/0049-089X(92)90017-B.

Greenberg, B.G., A.L.A. Abul-Ela, W.R. Simmons, and D.G. Horvitz. 1969. "The Unrelated Question Randomized Response Model: Theoretical Framework." *Journal of the American Statistical Association* 64: 520–539. DOI: http://dx.doi.org/10.1080/01621459.1969.10500991.

Greenberg, B.G., R.R. Kuebler Jr., J.R. Abernathy, and D.G.G. Horvitz. 1971. "Application of the Randomized Response Technique in Obtaining Quantitative Data." *Journal of the American Statistical Association* 66: 243–250. DOI: http://dx.doi.org/10.1080/01621459.1971.10482248.

Groves, R.M. 2004 [1989]. *Survey Error and Survey Costs*. Hoboken: Wiley & Sons.

Groves, R.M., F.J. Fowler, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken: Wiley & Sons.

Hausman, J. 2001. "Mismeasured Variables in Econometric Analysis: Problems From the Right and Problems from the Left." *The Journal of Economic Perspectives* 15: 57–67.

Hendrickx, J. 2002. "ISKO: Stata Module to Recode 4 Digit ISCO-88 Occupational Codes, Statistical Software Components s425802." Boston College Department of Economics. revised 20 Oct 2004. Available at: https://ideas.repec.org/c/boc/bocode/s425802.html (accessed February 14, 2015).

Holbrook, A.L., M.C. Green, and J.A. Krosnick. 2003. "Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires. Comparisons

of Respondent Satisficing and Social Desirability Response Bias." *Public Opinion Quarterly* 67: 79–125. DOI: http://dx.doi.org/10.1086/346010.

Holbrook, A.L. and J.A. Krosnick. 2010. "Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling into Question the Method's Validity." *Public Opinion Quarterly* 74: 328–343. DOI: http://dx.doi.org/10.1093/poq/nfq012.

Holtgraves, T. 2004. "Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding." *Personality and Social Psychology Bulletin* 30: 161–172. DOI: http://dx.doi.org/10.1177/0146167203259930.

Horvitz, D.G., B.V. Shah, and W.R. Simmons. 1967. "The Unrelated Question Randomized Response Model." In Proceedings of the Social Statistics Section. American Statistical Association, 65–72.

Höglinger, M., B. Jann, and A. Diekmann. 2014. *Sensitive Questions in Online Surveys: An Experimental Evaluation of the Randomized Response Technique and the Crosswise Model*. University of Bern Social Science Working Paper No. 9, 1–51. Available at: ftp://repec.sowi.unibe.ch/files/wp9/hoeglinger-jann-diekmann-2014.pdf (accessed September 17, 2014).

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

Jann, B. 2011. "Rrlogit: Stata module to estimate logistic regression for randomized response data." Statistical Software Components, Boston College Department of Economics. Available at: https://ideas.repec.org/c/boc/bocode/s456203.html (accessed February 14, 2015).

Jann, B., J. Jerke and I. Krumpal. 2012. "Asking Sensitive Questions Using the Crosswise Model. An Experimental Survey Measuring Plagiarism." *Public Opinion Quarterly* 71: 32–49. DOI: http://dx.doi.org/10.1093/poq/nfr036.

Kirchner, A. 2014. Techniques for Asking Sensitive Question in Labor Market Surveys. IAB-Bibliothek Dissertationen, 348. Bielefeld: Bertelsmann. Available at: http://edoc.ub.uni-muenchen.de/17192/1/Kirchner_Antje.pdf (accessed February 14, 2015).

Kirchner, A., I. Krumpal, M. Trappmann, and H. von Hermanni. 2013. "Messung und Erklärung von Schwarzarbeit in Deutschland – Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit." *Zeitschrift für Soziologie* 42: 291–314.

Korndörfer, M., I. Krumpal, and S.C. Schmukle. 2014. "Measuring and Explaining Tax Evasion: Improving Self-Reports Using the Crosswise Model." *Journal of Economic Psychology* 45: 18–32. DOI: http://dx.doi.org/10.1016/j.joep.2014.08.001.

Kreuter, F., G. Müller, and M. Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74: 880–906. DOI: http://dx.doi.org/10.1093/poq/nfq060.

Kreuter, F., G. Müller, and M. Trappmann. 2014. "A Note on Mechanisms Leading to Lower Data Quality of Late or Reluctant Respondents." Sociological Methods and Research 43: 452-464. DOI: http://dx.doi.org/10.1177/0049124113508094.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. DOI: http://dx.doi.org/10.1002/acp.2350050305.

Krumpal, I. 2012. "Estimating the Prevalence of Xenophobia and Anti-Semitism in Germany: A Comparison of Randomized Response and Direct Questioning." *Social Science Research* 41: 1387–1403. DOI: http://dx.doi.org/10.1016/j.ssresearch.2012.05.015.

Kuk, A.Y.C. 1990. "Asking Sensitive Questions Indirectly." *Biometrika* 77: 436–438. DOI: http://dx.doi.org/10.1093/biomet/77.2.436.

Lamb, C.W. and D.E. Stem. 1978. "An Empirical Validation of the Randomized Response Technique." *Journal of Marketing Research* 15: 616–621.

Landsheer, J.A., P.G.M. van der Heijden, and G. van Gils. 1999. "Trust and Understanding. Two Psychological Aspects of Randomized Response. A Study of a Method for Improving the Estimate of Social Security Fraud." *Quality & Quantity* 33: 1–12. DOI: http://dx.doi.org/10.1023/A:1004361819974.

Lara, D., S.G. García, C. Ellertson, C. Camlin, and J. Suárez. 2006. "The Measure of Induced Abortion Levels in Mexico Using Random Response Technique." *Sociological Methods & Research* 35: 279–301. DOI: http://dx.doi.org/10.1177/0049124106290442.

Lara, D., J. Strickler, C.D. Olavarrieta, and C. Ellertson. 2004. "Measuring Induced Abortion in Mexico." *Sociological Methods & Research* 32: 529–558. DOI: http://dx.doi.org/10.1177/0049124103262685.

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

Lensvelt-Mulders, G.J.L.M., J.J. Hox, P.G.M. van der Heijden, and C.J.M. Maas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods & Research* 33: 319–348. DOI: http://dx.doi.org/10.1177/0049124104268664.

Lensvelt-Mulders, G.J.L.M., J.J. Hox, and P.G.M. Van der Hejden. 2005b. "How to Improve the Efficiency of Randomized Response Designs." *Quality & Quantity* 39: 253–265. DOI: http://dx.doi.org/10.1007/s11135-004-0432-3.

Lensvelt-Mulders, G.J.L.M., P.G.M. Van der Heijden, O. Laudy, and G. van Gils. 2006. "A Validation of Computer-Assisted Randomized Response Survey to Estimate the Prevalence of Undeclared Work in Social Security." *Journal of the Royal Statistical Society (Series A)* 169: 305–318.

Locander, W., S. Sudman, and N. Bradburn. 1976. "An Investigation of Interview Method. Threat and Response Distortion." *Journal of the American Statistical Association* 71: 269–275. DOI: http://dx.doi.org/10.1080/01621459.1976.10480332.

Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

Mangat, N.S. 1994. "An Improved Randomized Response Strategy." *Journal of the Royal Statistical Society (Series B)* 56: 93–95.

Mangat, N.S. and R. Singh. 1990. "An Alternative Randomized Response Procedure." *Biometrika* 77: 439–442. DOI: http://dx.doi.org/10.1093/biomet/77.2.439.

Manzoni, A., J.K. Vermunt, R. Luijkx, and R. Muffels. 2010. "Memory Bias in Retrospectively Collected Employment Careers: A Model-Based Approach to Correct for Measurement Error." *Sociological Methodology* 40: 39–73. DOI: http://dx.doi.org/10.1111/j.1467-9531.2010.01230.x.

Mood, C. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About it." *European Sociological Review* 26: 67–82. DOI: http://dx.doi.org/10.1093/esr/jcp006.

Moors, J.J.A. 1971. "Optimization of the Unrelated Randomized Response Model." *Journal of the American Statistical Association* 66: 627–629. DOI: http://dx.doi.org/10.1080/01621459.1971.10482320.

Moshagen, M., E.B. Hilbig, E. Erdfelder, and A. Moritz. 2014. "An Experimental Validation Method for Questioning Techniques That Assess Sensitive Issues." *Experimental Psychology* 61: 48–54. DOI: http://dx.doi.org/10.1027/1618-3169/a000226.

Ostapczuk, M., M. Moshagen, Z. Zhao, and J. Musch. 2009. "Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry." *Journal of Educational and Behavioral Statistics* 43: 267–287. DOI: http://dx.doi.org/10.3102/1076998609332747.

Ostapczuk, M., J. Musch, and M. Moshagen. 2011. "Improving Self-Report Measures of Medication Non-Adherence Using a Cheating Detection Extension of the Randomized-Response Technique." *Statistical Methods in Medical Research* 20: 489–503. DOI: http://dx.doi.org/10.1177/0962280210372843.

Tourangeau, R. and K.A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103: 299–314. DOI: http://dx.doi.org/10.1037/0033-2909.103.3.299.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Tourangeau, R. and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133: 859–883. DOI: http://dx.doi.org/10.1037/0033-2909.133.5.859.

Tracy, P.E. and J.A. Fox. 1981. "The Validity of Randomized Response for Sensitive Measurements." *American Sociological Review* 46: 187–200.

Trappmann, M., S. Gundert, C. Wenzig, and D. Gebhardt. 2010. "PASS: A Household Panel Survey for Research on Unemployment and Poverty." *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften* 130: 609–622.

Umesh, U.N. and R.A. Peterson. 1991. "A Critical Evaluation of the Randomized Response Method: Applications, Validation, and Research Agenda." *Sociological Methods & Research* 20: 104–138. DOI: http://dx.doi.org/10.1177/0049124191020001004.

Van den Hout, A., U. Böckenholt, and P.G.M. van der Heijden. 2010. "Estimating the Prevalence of Sensitive Behavior and Cheating with Dual Design for Direct Questioning and Randomized Response." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59: 723–736. DOI: http://dx.doi.org/10.1111/j.1467-9876.2010.00720.x.

Van der Heijden, P.G.M., G. van Gils, J. Bouts, and J.J. Hox. 2000. "A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning: Eliciting Sensitive Information in the Context of Welfare and Unemployment Benefit." *Sociological Methods & Research* 28: 505–537. DOI: http://dx.doi.org/10.1177/0049124100028004005.

Warner, S.L. 1965. "Randomized-Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60: 63–69. DOI: http://dx.doi.org/10.1080/01621459.1965.10480775.

Weissman, A.N., R.A. Steer, and D.S. Lipton. 1986. "Estimating Illicit Drug Use Through Telephone Interviews and the Randomized Response Technique." *Drug and Alcohol Dependence* 18: 225–233. DOI: http://dx.doi.org/10.1016/0376-8716(86)90054-2.

Wolter, F. 2012. Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik. Springer VS.

Wolter, F. and P. Preisendörfer. 2013. "Asking Sensitive Questions: An Evaluation of the Randomized Response Technique versus Direct Questioning Using Individual Validation Data." *Sociological Methods & Research* 42: 321–353. DOI: http://dx.doi.org/10.1177/0049124113500474.

Yu, J.-W., G.L. Tian, and M.L. Tang. 2008. "Two New Models for Survey Sampling With Sensitive Characteristic: Design and Analysis." *Metrika* 67: 251–263. DOI: http://dx.doi.org/10.1007/s00184-007-0131-x.

Received August 2013
Revised October 2014
Accepted October 2014