Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model

Adrian Hoffmann¹ · Jochen Musch¹

Published online: 17 July 2015
© Psychonomic Society, Inc. 2015

Abstract Estimates of the prevalence of sensitive attributes obtained through direct questions are prone to being distorted by untruthful responding. Indirect questioning procedures such as the Randomized Response Technique (RRT) aim to control for the influence of social desirability bias. However, even on RRT surveys, some participants may disobey the instructions in an attempt to conceal their true status. In the present study, we experimentally compared the validity of two competing indirect questioning techniques that presumably offer a solution to the problem of nonadherent respondents: the Stochastic Lie Detector and the Crosswise Model. For two sensitive attributes, both techniques met the “more is better” criterion. Their application resulted in higher, and thus presumably more valid, prevalence estimates than a direct question. Only the Crosswise Model, however, adequately estimated the known prevalence of a nonsensitive control attribute.

Keywords Randomized response technique · Stochastic lie detector · Crosswise model · Social desirability bias

When assessing the prevalence of sensitive personal attributes, the validity of prevalence estimates obtained via direct questioning (DQ) procedures is threatened by response bias. Respondents frequently choose to align their answers to sensitive questions with social norms in order to make or uphold a

socially desirable impression (Krumpal, 2013; Marquis, Marquis, & Polich, 1986; Paulhus, 1991; Paulhus & Reid, 1991; Phillips & Clancy, 1972; Rasinski, Willis, Baldwin, Yeh, & Lee, 1999; Stocké, 2007; Sudman & Bradburn, 1974; Tourangeau & Yan, 2007). Consequently, prevalence estimates of sensitive attributes may be distorted by the under-reporting of socially undesirable and the over-reporting of socially desirable attitudes and behaviors.

Warner (1965) proposed the Randomized Response Technique (RRT) to increase respondents’ willingness to cooperate on sensitive surveys. This technique improves the confidentiality of individual answers by employing a randomization procedure that removes the direct association between a respondent’s answer and his or her standing on the sensitive attribute. However, even on RRT surveys, respondents may fail to adhere to the instructions in order to conceal their true status. After providing a brief introduction to the Randomized Response Technique, we will therefore describe and evaluate two recently proposed advanced models that were designed to address the problem of nonadherence to the instructions: The Stochastic Lie Detector (SLD; Moshagen et al., 2012) and the competing Crosswise Model (CWM; Yu et al., 2008). The SLD implements an additional parameter to estimate the proportion of sensitive attribute-carriers who cheat on the survey. Arguably, this should result in a more accurate prevalence estimate than traditional RRT procedures. The competing CWM does not model cheating but is instead characterized by rather simple instructions that make it particularly easy to understand how the confidentiality of answers is protected. Like the original Warner (1965) model, the CWM is symmetrical in the sense that it does not provide a “safe” answer option that offers the opportunity to explicitly deny being a carrier of the sensitive attribute. There are, however, no studies that have compared the validity of the two approaches. Therefore, we conducted a large-scale experimental survey that

✉ Adrian Hoffmann
adrian.hoffmann@hhu.de

Jochen Musch
jochen.musch@hhu.de

¹ Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, Duesseldorf 40225, Germany

aimed to evaluate and compare the two models with regard to their convergent validity and their ability to estimate the known prevalence of a control attribute. We also tested the two models against a direct questioning control condition.

The Randomized Response Technique (RRT)

The general idea behind the RRT is to ensure the confidentiality of individual answers to sensitive questions by adding random noise to the responses. In the original Related Questions Model (RQM; Warner, 1965), respondents are simultaneously presented with two questions – A (“Are you a carrier of the sensitive attribute?”) and B (“Are you *not* a carrier of the sensitive attribute?”). Depending on the outcome of a randomization procedure, the respondents are asked to answer either of these questions. If, for example, a die is used, subjects are instructed to respond to Question A if the die shows one of the numbers 1–4 (randomization probability $p = 4/6 = .67$) and to respond to Question B if the die shows either of the numbers 5 or 6 ($1 - p = 2/6 = .33$). Because the outcome of the randomization procedure remains unknown to the questioner, the true status of an individual respondent with respect to the sensitive attribute cannot be derived from his or her answer: A “Yes” response could possibly have been given by a carrier of the sensitive attribute who was instructed to respond to Statement A or from a noncarrier who was instructed to respond to Statement B. In view of the confidentiality thus afforded, respondents are expected to answer more truthfully than when questioned directly. In spite of the confidentiality guaranteed to the individual respondent, an estimate of the prevalence π of the sensitive attribute can be obtained at the sample level. Warner (1965) showed the maximum likelihood estimate of π in the RQM to be

$$\hat{\pi} = \frac{p-1 + \frac{n'}{n}}{2p-1} \quad , \quad p \neq 1/2 \quad (1)$$

where p is the known probability that the randomization device would select Statement A, n' represents the total number of “Yes” responses, and n reflects the sample size. Compared with a conventional direct questioning procedure, the RRT has lower statistical efficiency because the randomization procedure adds unsystematic variance to the answers. The reduced efficiency, however, is supposed to be overcompensated for by an increase in the validity of the prevalence estimates resulting from the presumably higher proportion of honest respondents.

Within the last almost 50 years, a large number of RRT models have been developed with various objectives such as improving efficiency (e.g., Boruch, 1971; Dawes & Moore,

1980; Eriksson, 1973; Mangat, 1994; Mangat & Singh, 1990; Moors, 1971), including questions with multicategorical or quantitative answers (e.g., Abul-El, Greenberg, & Horvitz, 1967; Himmelfarb & Edgell, 1980; Liu & Chow, 1976; Pollock & Bek, 1976), increasing respondents’ cooperation (e.g., Greenberg, Abul-El, Simmons, & Horvitz, 1969; Daniel G. Horvitz, Shah, & Simmons, 1967; Kuk, 1990; Ostapczuk, Moshagen, Zhao, & Musch, 2009), and accounting for cheating or noncompliance with the instructions (e.g., Clark & Desharnais, 1998; Moshagen et al., 2012). The RRT has been applied in surveys covering a variety of sensitive topics such as drug use (Dietz et al., 2013; Goodstadt & Gruson, 1975), doping (James, Nepusz, Naughton, & Petroczi, 2013; Simon, Striegel, Aust, Dietz, & Ulrich, 2006; Striegel, Ulrich, & Simon, 2010), crime (IIT Research Institute and the Chicago Crime Commission, 1971; Wolter & Preisendörfer, 2013), unwed motherhood (Abul-El et al., 1967), promiscuity (Liu, Chow, & Mosley, 1975), abortion (Abernathy, Greenberg, & Horvitz, 1970; Greenberg, Kuebler, Abernathy, & Horvitz, 1971), rape (Fidler & Kleinknecht, 1977; Soeken & Damrosch, 1986), homosexuality (Clark & Desharnais, 1998), tax evasion (Edgell, Himmelfarb, & Duchan, 1982), fraud (van der Heijden, van Gils, Bouts, & Hox, 2000), academic cheating (J.-P. Fox & Meijer, 2008; Hejri, Zendehdel, Asghari, Fotouhi, & Rashidian, 2013; Ostapczuk, Moshagen, et al., 2009), xenophobia (Ostapczuk, Musch, & Moshagen, 2009), negative attitudes toward people with disabilities (Ostapczuk & Musch, 2011), dental hygiene (Moshagen, Musch, Ostapczuk, & Zhao, 2010), and domestic violence (Moshagen et al., 2012). Overviews of RRT models and their applications have been given by Greenberg, Horvitz, and Abernathy (1974), D. G. Horvitz, Greenberg, and Abernathy (1976), J. A. Fox and Tracy (1986), A. Chaudhuri and Mukerjee (1988), Umesh and Peterson (1991), Scheers (1992), Antonak and Livneh (1995), Tracy and Mangat (1996), Franklin (1998), A. Chaudhuri (2011), and Arijit Chaudhuri and Christofides (2013).

In two meta-analyses, Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) reported an overall positive effect of the RRT on the validity of self-reports. The 32 comparative studies they found generally arrived at higher prevalence estimates of sensitive attributes in the RRT condition than in the direct questioning (DQ) control condition. Applying a “more is better” criterion, these higher estimates were usually considered to be more valid. However, this validation approach can be criticized as providing only relatively weak evidence because it is possible that both direct and indirect questioning techniques will provide inaccurate prevalence estimates (e.g., Umesh & Peterson, 1991). It is therefore important that in an additional meta-analysis of six methodologically stronger validation studies in which the respondents’ true status with respect to the sensitive attribute was known to the questioner, Lensvelt-Mulders et al. (2005) also found RRT estimates to be

more valid than estimates obtained via direct questioning because the RRT estimates deviated less from the known values in the population. Interestingly, conducting surveys online does not render the application of indirect questioning techniques unnecessary. It has repeatedly been observed that even in anonymous web surveys, prevalence estimates are higher when the RRT rather than a direct question is employed (Moshagen & Musch, 2012; Ostapczuk & Musch, 2011).

Despite the apparent advantages of RRT questioning, however, not all studies have supported its alleged superiority over conventional questioning methods. In some studies, estimates obtained via the RRT did not differ from those obtained via direct questioning (e.g., Akers, Massey, Clarke, & Lauer, 1983; Locander, Sudman, & Bradburn, 1976; Wolter & Preisendörfer, 2013). In other studies, they were even lower (e.g., Holbrook & Krosnick, 2010; Kulka, Weeks, & Folsom, 1981). Furthermore, Edgell et al. (1982) showed that a substantial proportion of participants failed to follow the RRT instructions, especially on surveys addressing highly sensitive issues. In view of these diverging patterns of results, Holbrook and Krosnick (2010) called the validity of RRT surveys into question.

Respondent jeopardy and *risk of suspicion* provide potential explanations for the divergent findings because either of these response hazards may lead to a violation of the assumptions underlying RRT models (Antonak & Livneh, 1995). The influence of these response hazards can primarily be observed in – and best be described with – forced-choice RRT designs (Boruch, 1971; Dawes & Moore, 1980). In this design variant, all participants are confronted with a single sensitive question, and a randomly chosen subsample is instructed to respond “Yes” regardless of their true status. Hence, a “Yes” response can stem from either a carrier or a noncarrier of the sensitive attribute who is either responding truthfully (carrier) or has simply been told to answer in the affirmative (carriers and noncarriers). It is important to note, however, that participants can still explicitly decline being carriers of the sensitive attribute by ignoring the instructions and simply responding “No.” In this situation, *respondent jeopardy* refers to the problem that guilty respondents make themselves more vulnerable by answering a sensitive question in the affirmative because they can be identified as carriers with a higher probability after a “Yes” than after a “No” response. If carriers perceive the risk of being identified as carriers as too high, they may choose to disobey the instructions by dishonestly responding “No.” Innocent respondents, on the other hand, suffer from a *risk of suspicion* because noncarriers have a higher risk of being falsely identified as carriers if they are forced to respond “Yes” by the randomization device. For this reason, they may also be inclined to disregard the instructions and to respond “No” in spite of being told otherwise (Antonak & Livneh, 1995). Lying carriers and suspicion-avoiding noncarriers

were explicitly accounted for by the introduction of the cheating detection model.

Detection of cheating on RRT surveys

Clark and Deshamais (1998) argued that even on RRT surveys, participants may refuse to adhere to the instructions if there is an answer option that allows them to avoid being identified as a carrier. They therefore proposed the Cheating Detection Model (CDM) as an improvement over the forced-response procedure. In addition to considering carriers of the sensitive attribute who answer honestly (π) and noncarriers who answer honestly (β), it considers a third class of respondents, namely cheaters (γ) who respond “No” regardless of the outcome of the randomization procedure. Various studies have shown that the CDM provides higher and thus potentially more valid prevalence estimates of sensitive attributes than a direct question (e.g., Moshagen et al., 2010; Ostapczuk, Moshagen, et al., 2009; Ostapczuk & Musch, 2011; Ostapczuk, Musch, et al., 2009; Ostapczuk, Musch, & Moshagen, 2011; Pitsch, Emrich, & Klein, 2007). However, the CDM does not make any assumptions about the real status of cheaters; they may be either lying carriers or noncarriers who wish to avoid suspicion. Consequently, a precise estimate of the total prevalence of a sensitive attribute can be obtained only if the proportion of cheaters is zero. Whenever cheating occurs ($\gamma > 0$), the prevalence of carriers of the sensitive attribute can be located anywhere within the range of π (if no cheater is a carrier) and $\pi + \gamma$ (if all cheaters carried the sensitive attribute; Clark & Deshamais, 1998). Thus, whenever $\gamma > 0$, the CDM provides only a lower (π) and an upper bound ($\pi + \gamma$) for the proportion of carriers. Several studies using the CDM have suggested that the proportion of cheaters on surveys covering sensitive topics may often be substantial and amount to up to 50 % of the sample (e.g., Ostapczuk, Moshagen, et al., 2009; Ostapczuk & Musch, 2011; Ostapczuk, Musch, et al., 2009; Ostapczuk et al., 2011). On the one hand, this underlines the importance of a cheating detection approach to RRT surveys; on the other hand, this means that if the rate at which people cheat is substantial, the CDM allows for only a very rough estimate of the proportion of carriers in a given population. Moreover, when the CDM is applied, nothing is or can be said about the true status of respondents who have to be classified as cheaters according to the model.

Moshagen et al. (2012) recently introduced a new RRT model that is presumed to be capable of providing an estimate of the prevalence of carriers (π) and of the proportion of cheaters with a *known* status at sample level: the SLD. Based on a modification of the original RQM (Warner, 1965) by Mangat (1994), the randomization process in the SLD is restricted to the group of

noncarriers. All respondents are presented with two statements – A (the sensitive statement) and B (the negation of statement A). In the instructions, however, participants are requested to ignore the randomization procedure and respond to Statement A if they are carriers of the sensitive attribute. Noncarriers are instructed to respond to Statement A with randomization probability p_i and to statement B with complementary probability $1 - p_i$. It is important to note that the selection of the statement is carried out by the participants themselves depending on their status with respect to the sensitive attribute, and the outcome of the randomization procedure. Hence, individual answers to the sensitive statement remain completely confidential. For example, in a survey assessing the prevalence of cocaine use, instructions could be:

“In the following, you will be presented with two complementary statements.

If you have ever used cocaine, please respond to Statement A.

If you have never used cocaine, please respond to...

- Statement A if you were born in November or December,
- Statement B if you were born in any other month.”

Next, two statements A and B would have to be presented:

Statement A: “I have used cocaine.”

Statement B: “I have never used cocaine.”

Finally, the participants would be asked to indicate whether they agreed to the statement they were required to respond to.

Moshagen et al. (2012) argued that when confronted with these instructions, carriers perceiving respondent jeopardy may have an incentive to disobey the instructions by responding “No” to the sensitive statement. Noncarriers, however, should have no reason to lie about their status, as this would mean associating themselves with a socially undesirable attribute. To model potential cheating among carriers, Moshagen et al. (2012) introduced a parameter t representing the proportion of carriers responding truthfully; the remaining carriers $(1 - t)$ are assumed to be lying to conceal their true status. Figure 1 illustrates the tree diagram of the resulting SLD.

To allow for the estimation of the two parameters π and t in the SLD, two independent randomly drawn subsamples have to be assessed with different randomization probabilities $p_1 \neq p_2$ (Clark & Desharnais, 1998), with a larger difference of p_1 and p_2 resulting in a higher statistical efficiency of the model (Moshagen et al., 2012).

Moshagen et al. (2012) showed the maximum likelihood estimates of π and t to be

$$\hat{\pi} = \frac{\left(\frac{n'_2}{n_2} - \frac{n'_1}{n_1}\right) + (p_2 - p_1)}{(p_2 - p_1)} \quad (2)$$

And

$$\hat{t} = \frac{\left[\frac{n'_2}{n_2}(1 - p_1)\right] - \left[\frac{n'_1}{n_1}(1 - p_2)\right]}{\left(\frac{n'_2}{n_2} - \frac{n'_1}{n_1}\right) + (p_2 - p_1)} \quad (3)$$

where n_1 and n_2 denote the sample sizes of the two samples tested with different randomization probabilities p_1 and p_2 , and n'_1 and n'_2 represent the absolute frequencies of “Yes” responses in these groups. Equations deriving the variances of π and t were also provided by Moshagen et al. (2012).

The SLD was first applied in two pilot studies by Moshagen et al. (2012): In an experimental survey assessing the prevalence of domestic violence, the SLD yielded a prevalence estimate that was about four times higher than with direct questioning and more than two times higher than with the Mangat (1994) model. In addition, the estimated proportion of carriers responding truthfully (t) differed significantly from 100 %, which indicated that a substantial number of carriers had decided to “play it safe” by choosing an answer option that would not make them look suspiciously like carriers (Moshagen et al., 2012). In a second experiment, estimates of the prevalence of nonvoting in the 2009 German federal elections obtained via DQ, the SLD, and the Mangat (1994) model were compared with the known true proportion of nonvoters in the general population obtained by official statistics. Again, the SLD provided an estimate of the proportion of nonvoters that was higher than the ones provided by direct questioning and by applying the Mangat (1994) model. Moreover, only the SLD estimate concurred almost exactly with the known true proportion of nonvoters (Moshagen et al., 2012).

The most compelling evidence supporting the validity of the SLD was provided in a recent validation study by Moshagen, Hilbig, Erdfelder, and Moritz (2014). In an adaptation of the “die-under-the-cup” paradigm (cf. Hilbig & Hessler, 2013), participants were instructed to secretly roll a die and to report the outcome to the experimenter. Some of the outcomes were associated with a monetary reward. As the outcome of the individual die rolls was unknown to the questioner, the participants’ actual behavior remained confidential. Thus, participants were given an opportunity to misrepresent the outcome of their die rolls in order to maximize their financial benefit. As the distribution of die roll outcomes was known to the experimenters, Moshagen et al. (2014) could

Fig. 1 Tree diagram of the Stochastic Lie Detector (Moshagen et al., 2012)

determine that of the alleged “winners,” about 53 % seemed to have cheated on the task. This known prevalence could then be used as an external criterion for the validation of the prevalence estimate obtained with the SLD and a DQ procedure. Moshagen et al. (2014) showed that a conventional DQ procedure substantially underestimated the known prevalence of cheaters (36 %), whereas the application of the SLD resulted in an estimate of 48 %, which did not differ significantly from the ground truth. In light of these results, Moshagen et al. (2014) considered the SLD to be a promising candidate within the class of advanced RRT models. It is important to note, however, that the SLD offers a “safe” answer category because a “No” response can stem only from a noncarrier. If noncarriers are attracted to this answer to avoid the risk of suspicion, the model assumptions are violated, and distorted prevalence estimates are to be expected. We therefore felt it necessary to conduct a further validation of the SLD and to compare it with the competing Crosswise Model (Yu et al., 2008), which claims to counteract both respondent jeopardy and risk of suspicion.

The Crosswise Model (CWM)

Within the last couple of years, a new class of so-called “nonrandomized response models” has been proposed (for an overview, see Tian & Tang, 2014). The goal of these models is to question the respondents indirectly without having to employ an external randomization procedure such as the rolling of a die. With the CWM as a member of this class, Yu et al. (2008) introduced a questioning technique that is arguably easier for the respondents to understand than other models. Moreover, the CWM holds the particular advantage of response symmetry because none of the answer options provides a “safe” alternative that clearly dispels the possibility of the respondent being a carrier of the sensitive attribute. In the CWM, participants are simultaneously presented with two

statements: one statement referring to a sensitive attribute with unknown prevalence π and another statement referring to a nonsensitive attribute with known prevalence p (e.g., a statement about the month of the respondent’s birth). Respondents are then asked to indicate whether “both statements are true or both statements are false” or whether “exactly one and only one of the two statements is true.” Neither of these answer options directly indicates whether the respondent is a carrier of the sensitive attribute, and neither of them clearly marks the respondent as a noncarrier. Respondent jeopardy and risk of suspicion are thus thoroughly circumvented. Yu et al. (2008) argued that the clear and easy-to-understand rationale and the convincing protection offered to the respondents by the symmetric CWM “would presumably not only make [them] willing to participate in the survey, but also persuade them to provide truthful responses” (p. 254). Response symmetry has, in fact, been shown to increase compliance with the instructions in other RRT models (e.g., Ostapczuk, Moshagen, et al., 2009). If response symmetry makes cheating-detection mechanisms such as the ones implemented in the CDM and the SLD dispensable, the CWM may be the model of choice for the assessment of sensitive attributes.

Because the CWM is mathematically equivalent to the model by Warner (1965), the maximum likelihood estimator for π is given by

$$\hat{\pi} = \frac{p-1 + \frac{n'}{n}}{2p-1}, \quad p \neq 1/2 \quad (4)$$

where p is the known prevalence of the nonsensitive statement, n' represents the total number of “both true or both false” responses, and n reflects the sample size. Equations deriving the variance of π are provided in Yu et al. (2008). Figure 2 illustrates the CWM as a tree diagram.

So far, a small number of publications have presented data from applications of the CWM. In two recently published studies, the CWM was applied without a direct questioning

Fig. 2 Tree diagram of the Crosswise Model (Yu et al., 2008)

control group (Eslami et al., 2013; Vakilian, Mousavi, & Keramat, 2014). More relevant to the present research are studies comparing the CWM and a direct questioning procedure. In two such studies, the CWM yielded a higher and therefore arguably more valid prevalence estimate for plagiarism in student papers than direct questions (Coutts, Jann, Krumpal, & Näher, 2011; Jann, Jerke, & Krumpal, 2012). When assessing the incidence of tax evasion in small and medium Serbian firms, Kundt, Misch, and Nerré (2013) also obtained significantly higher prevalence estimates when using the CWM than direct self-reports. The lifetime prevalence of anabolic steroid use in athletes was estimated as being more than two times higher when using the CWM rather than a direct question in a study by Nakhaee, Pakravan, and Nakhaee (2013). Jann et al. (2012) therefore evaluated the existing body of research as showing that “the [CWM] is successful in decreasing the social-desirability bias” (p. 13). It is important to note, however, that none of the existing studies provided a strong validation and direct evidence for the validity of the CWM because estimates obtained with the model were never compared with a known prevalence of carriers or non-carriers. If the CWM or the SLD does not provide correct estimates for the known prevalence of control attributes as well, the validity of the respective model will be called into question. We therefore decided to investigate whether the CWM and the SLD can correctly recover the known prevalence of a nonsensitive control attribute.

In contrast to models implementing a cheating detection device, the application of the CWM does not allow the user to test whether participants adhered to the instructions. Hence, it seemed worthwhile to compare its performance with the SLD as an alternative model that is not symmetrical but is rather based on a cheating detection procedure. To investigate the extent to which either model would succeed in motivating respondents to provide truthful answers to questions

addressing a sensitive topic, we also included a direct questioning (DQ) control condition.

Xenophobia, islamophobia, and the influence of the social desirability bias

We used a repeated measures design to compare the three questioning procedures (SLD, CWM, and DQ). To assess the ability of the different questioning techniques to control for social desirability, we included two questions pertaining to sensitive issues and a control question pertaining to an issue that was nonsensitive in nature but for which the true prevalence was known from official statistics. The two sensitive questions referred to xenophobia and islamophobia, respectively.

Xenophobia (i.e., a negative attitude toward people with an immigration background) has been shown to be an attitude that is rather widespread but that is usually met with social disapproval in Germany. Whereas self-reports using direct questioning procedures have shown only a moderate level of xenophobia in Germany that was comparable to the level observed in other Western European countries (Zick, Küpper, & Hövermann, 2011), Klink and Wagner (1999) demonstrated a heavy discrimination against ethnic minorities in a series of field experiments in which they manipulated the names, accents, and appearances of confederates presumably seeking help in everyday situations. Confederates mimicking a Turkish immigration background were far less likely to obtain a viewing appointment for a vacant house or to receive help in several situations in which they needed support. Regarding the frequently observed deviance of self-reports and actual behavior, Hjerm (1998) commented:

“The problem of social desirability is obviously important in studies that deal with such issues as xenophobia. It is possible that although they respond anonymously, people give socially desirable answers so as not to appear xenophobic. This might lead to an underestimation of the actual prevalence of xenophobia in a society.” (p. 338)

Krumpal’s (2012) results support this conjecture. Using a forced-response variant of the RRT (Boruch, 1971; Dawes & Moore, 1980) in a German telephone survey, he found that the RRT produced higher estimates of xenophobia than conventional DQ methods. Similarly, Ostapczuk, Musch, et al. (2009) showed that in a German sample, the proportion of xenophobes was substantially higher under the truth-eliciting CDM questioning procedure (Clark & Deshamais, 1998) than under a direct questioning procedure. They therefore concluded that the participants seemed to be “less unprejudiced than their answers to a direct question had suggested” (Ostapczuk,

Musch, et al., 2009, p. 928). The question we used to assess the prevalence of xenophobia read: “I would mind if my daughter had a relationship with a Turkish man.” This question was modeled after Bogardus (1933) and had been used before with other ethnic minorities by Silbermann and Hüßers (1995), Jimenez (1999), and Ostapczuk, Musch, et al. (2009).

As a second sensitive attribute, we assessed islamophobia, that is, a negative attitude toward, or even a fear of, people of the Muslim religion. Islamophobia is widespread in European countries (e.g., EUMC - European Monitoring Center on Racism and Xenophobia, 2006; Savelkoul, Scheepers, van der Veld, & Hagendoorn, 2012; Sheridan, 2006; Zick et al., 2011) and has been argued to be one of the most important political issues in modern Europe, possibly even “[m]uch more pressing” than anti-Semitism (Bunzl, 2005, p. 506). Even though the German constitution guarantees religious freedom, Germany is one of the highest ranked European countries in anti-Muslim attitudes (Zick et al., 2011). A strong connection between islamophobia and negative attitudes toward the construction of Muslim religious buildings was recently reported by Imhoff and Recker (2012). Individual scores of German participants on an Islamoprejudice subscale proved highly predictive of negative attitudes toward the construction of a great new mosque in the city of Cologne. We therefore decided to use an item that asked for negative attitudes toward the construction of minarets in Germany. This item was chosen because, in a recent popular vote, the citizens of Switzerland had voted in favor of a constitutional addendum that prohibited any further construction of minarets. The result of this popular vote had not been predicted by representative polls (gfs.bern, 2009a, b; reformiert, 2009), arguably because voters refrained from revealing what had been stigmatized as an attitude that tends to be met with social disapproval in the debate preceding the poll (fög, 2010).

Umesh and Peterson (1991) have argued that “[s]tudies that compared the RR[T] with other forms of questioning [...] are not validation studies” and that a “true validation study must compare the randomized response estimate and the actual value” (p. 127). Two such “strong” validation studies have been conducted for the SLD (Moshagen et al., 2014; Moshagen et al., 2012), but such studies have yet to be reported for the CWM. Therefore, one goal of the present study was to investigate whether the SLD and the CWM would be capable of recovering the known prevalence of an attribute. Unfortunately, however, the ground truth for sensitive attributes is usually unknown and difficult to obtain, as reflected in the relatively small number of only six “strong” validation studies that compared Randomized Response estimates with a known prevalence as reported in Lensvelt-Mulders et al.’s (2005) meta-analysis. In one of these studies that reported on social security fraud, the assessment of a sample that had a true prevalence of carriers of 100 % was possible only because of the public availability of databases containing the addresses of people

who had previously been convicted of such crimes in The Netherlands (van der Heijden et al., 2000). No such databases are available in Germany, however. Because there was no way to know the true value of a sensitive attribute in our student sample, we included a nonsensitive control question that pertained to the first letter of the respondents’ surname, for which the incidence in the general population could be determined. This allowed us to go beyond the usual “more is better” validation approach and to detect method-specific biases in the assessment of the prevalence of an attribute. Official statistics from the German *Statistisches Bundesamt* (Federal Office of Statistics) show that the proportion of citizens in Germany with a surname that begins with one of the relatively frequent letters K, L, M, R, S, or T is about 43 % (Reinders, 1996). This proportion was cross-checked with the student office of the University of Düsseldorf to rule out the possibility that the proportion was different in our student sample; however, the two proportions were almost identical, as 43 % of the 15,658 students carried a surname starting with one of the letters mentioned above. If the SLD and the CWM are capable of obtaining valid prevalence estimates of sensitive attributes, they should also perform well when applied to a nonsensitive control attribute.

To summarize, the present experiment addressed the following two questions: (a) Are the SLD and the CWM capable of controlling for social desirability? To the extent to which they are, the two indirect questioning techniques were expected to provide higher prevalence estimates of the two sensitive attributes than a direct question. (b) Are the SLD and the CWM prone to a method-specific bias that results in systematic over- or underestimates? If so, the two indirect questioning techniques should provide estimates that are at odds with official statistics with regard to the prevalence of surnames that begin with certain letters.

Method

A total of 1,312 subjects volunteered to participate in our survey. The sample (56 % female, mean age = 21.21 years, SD = 3.14) consisted of students from three German universities (Düsseldorf 81 %, Duisburg 10 %, and Bochum 9 %) who were recruited and assessed in groups in lecture halls before classes began.

Survey design

Respondents filled out a one-page questionnaire consisting of a short introduction, the three (sensitive and nonsensitive) experimental questions, and two demographic questions asking for the respondents’ age and gender. The questioning technique was varied as an independent within-subjects variable and consisted of the SLD (randomization device: mother’s

month of birth; subdivided into two groups with low vs. high randomization probabilities of $p_1 = .158$ vs. $p_2 = .842$, respectively), the CWM (nonsensitive statement: father's month of birth; known prevalence $p = .158$), and the DQ format. The question format was determined randomly for each question with the constraint that all three questioning techniques should be applied; thus, every participant responded to all three questions, but each question was presented in a different format. Two questions referred to sensitive attributes (xenophobia/negative attitudes toward Turkish immigrants; islamophobia/negative attitudes toward the construction of minarets in Germany) with unknown prevalences π_{s1} and π_{s2} , respectively. The third question referred to the first letter of the respondents' surname as a nonsensitive control attribute. The prevalence of this nonsensitive attribute was known ($\pi_{ns} = .43$) because it could be obtained from official statistics for the set of letters that was used for this question (first letter K, L, M, R, S, or T; Reinders, 1996). Examples of the three questioning formats are given below.

SLD format For the question referring to xenophobia, the SLD format (with a low randomization probability of $p_1 = .158$) read as follows:

“Assume that you have a 20-year-old daughter: Would you mind if she had a relationship with a Turkish man?
If yes, please respond to Statement A.
If not, ...

- please respond to Statement A if your mother was born in November or December,
- please respond to Statement B if your mother was born in any other month.”

Next, two statements A and B were presented:

Statement A: “I would mind if my daughter had a relationship with a Turkish man.”

Statement B: “I would not mind if my daughter had a relationship with a Turkish man.”

Finally, the participants were asked to indicate whether they agreed with the statement they were required to respond to. For the two other topics, the SLD questioning format was adapted accordingly. **CWM format** For the question referring to islamophobia, the CWM question (with a prevalence of the nonsensitive statement of $p = .158$) was presented as follows:

“Please read the following two statements:

Statement A: The construction of minarets should be prohibited in Germany.

Statement B: My father was born in November or December.”

Subsequently, the respondents were asked to indicate whether “both statements are true or both statements are false,” or whether “exactly one statement is true (regardless of which one).” For the two other topics, the CWM format was adapted accordingly.

DQ format For the nonsensitive control question with known prevalence ($\pi_{ns} = .43$), the direct question was presented as follows:

Statement: “My surname begins with one of the following letters: K, L, M, R, S, or T.”

The respondents were then asked to indicate whether this statement was true or false. For the two sensitive questions, the DQ format was adapted accordingly.

Statistical analysis

In the CWM and SLD conditions, prevalence estimates can be obtained using Eqs. 2 through 4. In the DQ condition, the proportion of respondents answering “true” to the direct question provides a direct prevalence estimate. Following the procedure detailed in Moshagen, Hilbig, and Musch (2011), Moshagen and Musch (2012), Moshagen et al. (2012), Moshagen et al. (2010), Ostapczuk, Moshagen, et al. (2009), Ostapczuk and Musch (2011), Ostapczuk, Musch, et al. (2009), and Ostapczuk et al. (2011), however, we formulated multinomial processing tree models (MPT; Batchelder, 1998; Batchelder & Riefer, 1999) for all three questioning techniques. This approach offers more flexibility in parameter estimation and convenient statistical tests of parameter restrictions (Moshagen et al., 2012). Within the multinomial modeling framework and using the procedures detailed in Hu and Batchelder (1994), it was possible to estimate the prevalence parameters for each questioning technique and to conduct the necessary statistical tests of our hypotheses. On the basis of the empirically observed answer frequencies in the different experimental conditions, we computed maximum likelihood estimates for all parameters using the expectation-maximization algorithm (EM; Dempster, Laird, & Rubin, 1977; Hu & Batchelder, 1994) implemented in the software multiTree (Moshagen, 2010). The model fit was tested via the asymptotically χ^2 -distributed log-likelihood statistic G^2 . The MPT models for all three questioning techniques were saturated with $df = 0$ and $G^2 = 0$ as the number of independent answer categories was just sufficient to estimate all parameters in the three questioning technique conditions: The two proportions of “Yes” responses in the conditions with a low versus high randomization probability allowed us to estimate the two parameters π and t in the SLD condition; the proportion of “Both true or both false” responses allowed us to estimate π in the CWM condition; and the proportion of “Yes” responses allowed us to estimate π in the DQ condition. Comparisons between parameter estimates and comparisons between the

parameters and a constant were conducted by assessing the significance of the difference in model fit (ΔG^2) between an unrestricted baseline model and an alternative model in which either the two parameters under question were restricted to be equal or one parameter was set to a constant value (e.g., $\pi_{ns} = .43$). A tree representation of the multinomial model and the observed answering frequencies for all conditions are given in Appendices A and B.

Results

Table 1 shows the prevalence estimates for the two sensitive attributes, and the nonsensitive control attribute obtained via DQ, the SLD, and the CWM.

Xenophobia (sensitive attribute 1)

To investigate whether different questioning techniques would result in different parameter estimates, pairwise comparisons between DQ, SLD, and CWM conditions were conducted. These revealed that in comparison with the DQ condition (26.98 %), respondents were more likely to answer truthfully in both the SLD (53.38 %) and the CWM conditions (48.67 %), ΔG^2 ($df = 1$) = 16.80, $p < .001$ and ΔG^2 ($df = 1$) = 28.20, $p < .001$, respectively. This pattern suggests that the prevalence of xenophobia was presumably underestimated in the DQ condition. A comparison of the two indirect questioning techniques revealed no significant difference in the prevalence estimates between the SLD and the CWM conditions, ΔG^2 ($df = 1$) = 0.43, $p = .51$. The SLD estimated the proportion of carriers of the sensitive attribute answering honestly at $t = .79$, a value significantly below 1.0, ΔG^2 ($df = 1$) = 14.56, $p < .001$. This finding suggests that according to the SLD, a substantial proportion of 21 % of the carriers seems to have disobeyed the instructions, possibly to conceal their true status.

Islamophobia (sensitive attribute 2)

As for the xenophobia item, the pattern of results suggests an underestimation of the prevalence in the DQ condition. In both the SLD (76.93 %) and CWM (51.64 %) conditions, the proportion of respondents with negative attitudes was estimated as higher than in the DQ condition (43.33 %), ΔG^2 ($df = 1$) = 23.97, $p < .001$ and ΔG^2 ($df = 1$) = 3.89, $p < .05$, respectively. However, unlike for the xenophobia item, the two indirect questioning techniques showed diverging results: In the SLD condition, the estimated proportion of carriers was significantly higher than in the CWM condition, ΔG^2 ($df = 1$) = 11.80, $p < .001$. The SLD estimated the proportion of carriers of the sensitive attribute answering honestly at $t = .68$, a value significantly below 1.0, ΔG^2 ($df = 1$)

= 65.07, $p < .001$. This finding suggests that a substantial proportion of 32 % of the carriers seems to have disobeyed the instructions.

Nonsensitive control attribute with known prevalence: First letter of surname

As expected for a nonsensitive attribute, there was no significant difference between the prevalence estimates obtained via DQ (40.99 %) and the CWM (46.57 %), ΔG^2 ($df = 1$) = 1.73, $p = .19$. Unexpectedly, however, the SLD estimate (62.72 %) was significantly higher than both the DQ and CWM estimates, ΔG^2 ($df = 1$) = 11.00, $p < .001$ and ΔG^2 ($df = 1$) = 5.15, $p < .05$, respectively. The DQ and CWM estimates deviated only slightly and nonsignificantly from the known prevalence of $\pi_{ns} = .43$, ΔG^2 ($df = 1$) = 0.73, $p = .39$ and ΔG^2 ($df = 1$) = 1.02, $p = .31$, respectively. By contrast, the SLD significantly overestimated the known prevalence, ΔG^2 ($df = 1$) = 10.42, $p < .01$. The SLD estimated the proportion of carriers of the sensitive attribute answering honestly at $t = .78$, a value significantly below 1.0, ΔG^2 ($df = 1$) = 23.16, $p < .001$, indicating that approximately 22 % of the carriers of the nonsensitive attribute seemed to have disobeyed the instructions.

Discussion

Social desirability bias may lead to the under-reporting of socially undesirable attributes. The present study investigated the validity of two competing indirect questioning techniques, the Stochastic Lie Detector (SLD; Moshagen et al., 2012) and the Crosswise Model (CWM; Yu et al., 2008), both of which aim to experimentally address the problem of social desirability bias. According to the “more is better” criterion, higher estimates of socially undesirable attributes can be considered more valid as they presumably suffer less from distortion. Using a large-scale survey, we therefore assessed whether the application of the two indirect questioning techniques would result in higher prevalence estimates than a conventional direct questioning (DQ) approach for two sensitive statements. Because the “more is better” criterion fails if a questioning technique provides estimates that surpass the known prevalence of a criterion, we also tested whether the application of the SLD or the CWM would result in undistorted estimates of a third nonsensitive control attribute. To the extent to which estimates provided by an indirect questioning technique are higher than the actual known prevalence of a control attribute, the validity of this indirect questioning technique is called into question.

With regard to the prevalence of xenophobia, both indirect questioning techniques yielded prevalence estimates that were

Table 1 Prevalence estimates (standard errors in parentheses) by questioning technique for the two sensitive attributes, and the nonsensitive control attribute

Parameter	Direct questioning	Stochastic Lie Detector	Crosswise Model
Sensitive attribute 1: Xenophobia			
π_{s1}	26.98 % (2.11)	53.38 % (6.31)	48.67 % (3.48)
t	–	79.43 % (4.59)	–
Sensitive attribute 2: Islamophobia			
π_{s2}	43.33 % (2.40)	76.93 % (6.62)	51.64 % (3.46)
t	–	67.94 % (3.19)	–
Nonsensitive control attribute: First letter of surname (known prevalence: $\pi_{ns} = .43$)			
π_{ns}	40.99 % (2.33)	62.72 % (6.25)	46.57 % (3.54)
t	–	78.23 % (3.92)	–

approximately twice as high and thus presumably more valid than the estimate from the direct question. The SLD estimated the prevalence of carriers responding truthfully to be substantially lower than 100 %. A quite similar pattern of results was observed for the islamophobia item. The CWM estimated the true prevalence of islamophobia to be significantly higher than estimated by a direct question. The SLD estimate even surpassed the CWM estimate, and the proportion of carriers answering truthfully was, again, estimated substantially lower than 100 %. These results add to the evidence that suggests that the self-report of both xenophobic and islamophobic attitudes may be distorted by a social desirability bias and that indirect questioning techniques may be capable of yielding more valid prevalence estimates by granting respondents full confidentiality of their answers. It has to be kept in mind, however, that our results are based on a convenience sample. Thus, the prevalence estimates we obtained might not be generalizable to the German population at large.

Despite this limitation, it is interesting to note that our results would predict diverging outcomes for a hypothetical popular vote on this issue, at least within the population our sample was drawn from. On the basis of the results of the direct question condition, one would have to predict that a majority would vote against the introduction of a law prohibiting the construction of minarets; according to the results obtained in the SLD and CWM conditions, however, one would have to predict that the proposal of such a law would pass a referendum. The latter result was in fact the outcome of a popular vote conducted in Switzerland in 2009, a result that was generally considered surprising because a poll had predicted the opposite outcome just prior to the vote. This poll, however, had been based on a direct question. Future studies based on probability samples could clarify whether the use of indirect questioning techniques might, indeed, increase the predictive validity of voting polls.

In summary, the results we obtained for the two sensitive questions attest to the validity of the indirect questioning

techniques with regard to the “more is better” criterion. The application of both indirect questioning techniques resulted in higher and therefore presumably more valid prevalence estimates for the two sensitive topics under investigation.

To determine whether a method bias that would result in a general tendency to over- or underestimate the prevalence of any attribute is inherent to either the SLD or the CWM, we included a control question that referred to a nonsensitive attribute with known prevalence. In accordance with the assumption of no bias, the CWM yielded a prevalence estimate (47 %) that was fairly close to and not significantly different from the known true prevalence of 43 %. Supporting the validity of this estimate, the CWM estimate did not differ significantly from the estimate yielded by the direct question (41 %). This result was to be expected considering that the first letter of a person’s surname is not a sensitive attribute, and corresponding self-reports should therefore not be distorted by social desirability bias. Thus, the validity of the CWM was confirmed with regard to both better control over social desirability bias as compared with a direct question and the lack of a method bias resulting in a systematic tendency to over- or underestimate.

Unlike the CWM, however, the SLD substantially overestimated the known prevalence of the control attribute (SLD: 63 % vs. true: 43 %). The SLD estimate also differed significantly from the estimate yielded by the direct question, which closely mirrored the known true prevalence of the nonsensitive control attribute (DQ: 41 % vs. true: 43 %). The proportion of carriers responding truthfully on the SLD was estimated at 78 %, which is significantly lower than the 100 % that would have to be expected if all participants had completely complied with the instructions.

Several alternative explanations for this unexpected outcome seem possible. First, the SLD may have a harmful tendency to overestimate the prevalence of any given attribute. Holbrook and Krosnick (2010) called the validity of the RRT method into question when obtaining an estimate for the prevalence of a socially desirable attribute that was unexpectedly

higher than the corresponding estimate obtained with a direct question and even reached “impossible levels” (Holbrook & Krosnick, 2010, p. 336) of over 100 %. Wolter and Preisendörfer (2013), however, argued that to draw general conclusions regarding the validity of the RRT might be premature. When assessing the validity of the SLD, it has to be taken into account that the technique performed well in two studies by Moshagen et al. (2012) and Moshagen et al. (2014), both of which found that the SLD provided estimates in accordance with the known prevalence of a sensitive attribute. Moreover, the SLD performed well for the xenophobia item in the current study, providing an estimate close to the estimate obtained using the CWM, which in turn provided presumably valid estimates for all questions in the present investigation.

An alternative explanation for why the SLD did not yield valid results for all questions in the present study may be found in its specific implementation. Although the SLD was designed to address one particular type of nonadherence to instructions, namely, untruthful responding by carriers of a sensitive attribute, its assumptions are clearly violated if (a) noncarriers falsely claim to carry the attribute, (b) carriers strategically use the randomization procedure to appear as though they are noncarriers, (c) response behavior varies for different randomization probabilities, or (d) respondents generally fail to understand and follow the instructions (cf. Moshagen et al., 2012). Any of the above problems can lead to distorted prevalence estimates. However, given that the surname control item was nonsensitive in nature, the three potential violations described in (a), (b), and (c) would be unlikely causes of the observed distortion. Moreover, as pointed out by Moshagen et al. (2012), violations of the assumptions according to (b) and (c) should have led to an under- rather than an overestimation of the attribute’s prevalence. A general failure to understand and follow the instructions, however, might offer a potential explanation for the present findings. Various researchers have pointed out that RRT questions may generally be difficult for some participants to understand (e.g., Landsheer, van der Heijden, & van Gils, 1999; Locander et al., 1976; van der Heijden, van Gils, Bouts, & Hox, 1998). The validity of an RRT estimate, however, strongly depends on the participants’ comprehension of the instructions (e.g., Abul-Ela et al., 1967; Holbrook & Krosnick, 2010; Soeken & Macready, 1982). In a recent survey using the unrelated question variant of the RRT, James et al. (2013) surmised that a misunderstanding of the instructions might have led to the inflated estimates they obtained for the use of performance-enhancing drugs. Comprehension problems might be specifically prevalent in related question RRT designs. These designs implement positively and negatively worded statements, and require some participants to use a double negative as their response, which has been shown to be potentially confusing (e.g., Johnson, Bristow, & Schneider, 2011). The instructions of the SLD, however, have been modified in a way that

generally excludes the possibility of respondents having to solve a double negative: All carriers of the attribute in question are required to respond to the positively worded Statement A. Even though some of the noncarriers are instructed to respond to the negatively worded Statement B, their response to this statement should always be “true.” Hence, the inflated parameter estimates for the nonsensitive control attribute observed in the SLD condition might be attributable to a general failure to understand the instructions, but can hardly be explained by difficulties in solving a double negative.

Another potential reason for the performance problems we observed for the SLD might be that, unlike the CWM, the SLD does not offer response symmetry to the respondents. Using the SLD, it is possible to respond in a way that allows the respondent to avoid looking suspicious of being a carrier of the sensitive attribute. This possibility to “play it safe” may lead to distorted response behavior. However, the distortion we observed occurred for a question that was nonsensitive in nature. Thus, a tendency to “play it safe” can hardly explain why we obtained an overestimate for a nonsensitive control item using the SLD. However, it is conceivable that the application of the SLD to a nonthreatening control question may have seemed odd to some of the respondents. This may have led to some confusion or even to a rejection of the method, but unfortunately, no post hoc test of this explanation was possible with the data we collected. It should however be noted that any response behavior that deviated from the instructions – including random responses – that equally extends to both the low and high randomization probability conditions can be shown to necessarily lead to an overestimation when using the SLD whenever $\pi < 1.00$. Therefore, it seems necessary to provide a more systematic investigation of the comprehensibility of RRT questions and compliance with the instructions in future research. Judging from the present results, it would appear that both comprehensibility and compliance with the instructions might be better for the CWM than for the SLD.

In conclusion, our results suggest that the CWM offers a valid and useful means for achieving the experimental control of social desirability. While previous studies have found evidence for the validity of the SLD (e.g., Moshagen et al., 2014), our results tentatively suggest that the CWM might be superior to the SLD with regard to applicability and validity. Even though both models met the “more is better” criterion in the assessment of two sensitive attributes, only the CWM succeeded in estimating the known prevalence of a nonsensitive control attribute. This finding further supports the notion of Umesh and Peterson (1991) that studies validating indirect questioning techniques have to go beyond the “more is better” criterion, and should best apply an external validation criterion. On the basis of our results, it seems justifiable to recommend the use of the CWM in future studies investigating sensitive issues.

Appendix A

Tree diagram representation for the model established for parameter estimation in multiTree (Moshagen, 2010):

Sensitive attribute 1: Xenophobia

Sensitive attribute 2: Islamophobia

Nonsensitive control attribute: First letter of surname

Appendix B

Table 2 Empirically observed frequencies of answers used in parameter estimation in multiTree (Moshagen, 2010)

Parameter	SLDp1	SLDp2	CWM	DQ
Sensitive attribute 1: Xenophobia				
“True” / “Both true / both false”	178	106	224	119
“False / One True”	40	107	216	322
Sensitive attribute 2: Islamophobia				
“True” / “Both true / both false”	157	123	218	185
“False / One True”	62	97	228	242
Nonsensitive control attribute: First letter of surname				
“True” / “Both true / both false”	177	122	223	182
“False / One True”	43	100	203	262

SLD = Stochastic Lie Detector, CWM = Crosswise Model, DQ = direct questioning

References

- Abernathy, J. R., Greenberg, B. G., & Horvitz, D. G. (1970). Estimates of induced abortion in urban North-Carolina. *Demography*, 7(1), 19–29.
- Abul-El, A.-L. A., Greenberg, B. G., & Horvitz, D. G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association*, 62, 990–1008.
- Akers, R. L., Massey, J., Clarke, W., & Lauer, R. M. (1983). Are self-reports of adolescent deviance valid? Biochemical measures, randomized-response, and the bogus pipeline in smoking-behavior. *Social Forces*, 62, 234–251.
- Antonak, R. F., & Livneh, H. (1995). Randomized-response technique - a review and proposed extension to disability attitude research. *Genetic, Social, and General Psychology Monographs*, 121, 97–145.
- Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, 10, 331–344. doi:10.1037/1040-3590.10.4.331
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57–86. doi:10.3758/BF03210812
- Bogardus, E. S. (1933). A social distance scale. *Sociology & Social Research*, 17, 265–271.
- Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *American Sociologist*, 6, 308–311.
- Bunzl, M. (2005). Between anti-Semitism and Islamophobia: Some thoughts on the new Europe. *American Ethnologist*, 32, 499–508. doi:10.1525/ae.2005.32.4.499
- Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton, Florida: Chapman & Hall, CRC Press, Taylor & Francis Group.
- Chaudhuri, A., & Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys*. Berlin, Heidelberg: Springer.
- Chaudhuri, A., & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.
- Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods*, 3, 160–168.
- Coutts, E., Jann, B., Krumpal, I., & Näher, A.-F. (2011). Plagiarism in student papers: Prevalence estimates using special techniques for sensitive questions. *Jahrbücher für Nationalökonomie Und Statistik*, 231, 749–760.
- Dawes, R. M., & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen [Guttman scaling of orthodox and randomized reactions]. In F. Petermann (Ed.), *Einstellungsmessung, Einstellungsforschung [Attitude measurement, attitude research]* (pp. 117–133). Göttingen: Hogrefe.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1–38.
- Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., & Ulrich, R. (2013). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy*, 33, 44–50.
- Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of forced responses in a randomized-response model. *Sociological Methods & Research*, 11, 89–100. doi:10.1177/0049124182011001005
- Eriksson, S. A. (1973). A new model for randomized response. *International Statistical Review*, 41, 101–113.
- Eslami, M., Yazdanpanah, M., Taheripana, R., Andalib, P., Rahimi, A., & Nakhaee, N. (2013). Importance of pre-pregnancy counseling in Iran: Results from the High Risk Pregnancy Survey 2012. *International Journal of Health Policy and Management*, 1, 213–218.
- EUMC - European Monitoring Center on Racism and Xenophobia. (2006). *Muslims in the European Union: Discrimination and Islamophobia*. Vienna, Austria: FRA.
- Fidler, D. S., & Kleinknecht, R. E. (1977). Randomized response versus direct questioning - 2 data-collection methods for sensitive information. *Psychological Bulletin*, 84, 1045–1049.
- fög. (2010). Berichterstattung zur Volksinitiative 'Gegen den Bau von Minaretten'. Retrieved June 8th, 2010, from <http://www.foeg.unizh.ch>
- Fox, J.-P., & Meijer, R. R. (2008). Using item response theory to obtain individual information from randomized response data: an application using cheating data. *Applied Psychological Measurement*, 32, 595–610. doi:10.1177/0146621607312277
- Fox, J. A., & Tracy, P. E. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.
- Franklin, L. (1998). Randomized response techniques. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (Vol. 5, pp. 3696–3703). New York: Wiley.
- gfs.bern. (2009a). 'Minarett-Initiative': Das Nein überwiegt – SVP-Wählerschaft dafür. Retrieved June 6th, 2010, from www.gfsbern.ch
- gfs.bern. (2009b). 'Minarett-Initiative': Ja nimmt zu - Nein unverändert stärker. Retrieved June 8th, 2010, from www.gfsbern.ch
- Goodstadt, M. S., & Gruson, V. (1975). Randomized response technique - test on drug-use. *Journal of the American Statistical Association*, 70, 814–818.
- Greenberg, B. G., Abul-El, A. L. A., Simmons, W. R., & Horvitz, D. G. (1969). Unrelated question randomized response model - theoretical framework. *Journal of the American Statistical Association*, 64, 520–539.
- Greenberg, B. G., Horvitz, D. G., & Abernathy, J. R. (1974). A comparison of randomized response designs. In F. Proschan & R. J. Serfling (Eds.), *Reliability and biometry, statistical analysis of life length* (pp. 787–815). Philadelphia: SIAM.
- Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., & Horvitz, D. G. (1971). Application of randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243–250.

- Hejri, S. M., Zendehele, K., Asghari, F., Fotouhi, A., & Rashidian, A. (2013). Academic disintegrity among medical students: a randomised response technique study. *Medical Education*, 47, 144–153. doi:10.1111/Medu.12085
- Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology*, 49, 263–266. doi:10.1016/j.jesp.2012.11.010
- Himmelfarb, S., & Edgell, S. E. (1980). Additive constants model - a randomized-response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin*, 87, 525–530. doi:10.1037//0033-2909.87.3.525
- Hjem, M. (1998). National identities, national pride and xenophobia: A comparison of four Western countries. *Acta Sociologica*, 41, 335–347.
- Holbrook, A. L., & Krosnick, J. A. (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:10.1093/Poq/Nfq012
- Horvitz, D. G., Greenberg, B. G., & Abernathy, J. R. (1976). Randomized response - data-gathering device for sensitive questions. *International Statistical Review*, 44, 181–196.
- Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association*.
- Hu, X., & Batchelder, W. H. (1994). The statistical-analysis of general processing tree models with the em algorithm. *Psychometrika*, 59, 21–47. doi:10.1007/Bf02294263
- IIT Research Institute and the Chicago Crime Commission. (1971). A study of organized crime in Chicago. Chicago: IITRI Project No. H-6031, Report prepared for the Illinois Enforcement Commission.
- Imhoff, R., & Recker, J. (2012). Differentiating Islamophobia: Introducing a new scale to measure Islamoprejudice and secular Islam critique. *Political Psychology*, 33, 811–824. doi:10.1111/j.1467-9221.2012.00911.x
- James, R. A., Nepusz, T., Naughton, D. P., & Petroczi, A. (2013). A potential inflating effect in estimation models: Cautionary evidence from comparing performance enhancing drug and herbal hormonal supplement use estimates. *Psychology of Sport and Exercise*, 14, 84–96. doi:10.1016/j.psychsport.2012.08.003
- Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise method. *Public Opinion Quarterly*, 76, 32–49. doi:10.1093/Poq/Nfr036
- Jimenez, P. (1999). Weder Opfer noch Täter - die alltäglichen Einstellungen 'unbeteiligter' Personen gegenüber Ausländern [Neither victim nor offender—the common attitudes of 'non-involved' persons towards foreigners]. In R. Dollase, T. Kliche, & H. Moser (Eds.), *Politische Psychologie der Fremdenfeindlichkeit. Opfer - Täter - Mittäter* (pp. 293–306). Weinheim: Juventa.
- Johnson, J. M., Bristow, D. N., & Schneider, K. C. (2011). Did you not understand the question or not? An investigation of negatively worded questions in survey research. *Journal Of Applied Business Research*, 20, 75–86.
- Klink, A., & Wagner, U. (1999). Discrimination against ethnic minorities in Germany: Going back to the field. *Journal of Applied Social Psychology*, 29, 402–423. doi:10.1111/j.1559-1816.1999.tb01394.x
- 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:10.1016/j.ssresearch.2012.05.015
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47, 2025–2047. doi:10.1007/s11135-011-9640-9
- Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436–438. doi:10.1093/biomet/77.2.436
- Kulka, R. A., Weeks, M. F., & Folsom, R. E. (1981). *A comparison of the randomized response approach and direct questioning approach to asking sensitive survey questions*. Working paper. NC: Research Triangle Institute.
- Kundt, T. C., Misch, F., & Nerré, B. (2013). Re-assessing the merits of measuring tax evasions through surveys: Evidence from Serbian firms. ZEW Discussion Papers, No. 13-047. Retrieved Dec 12th, 2013, from <http://hdl.handle.net/10419/78625>
- Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (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:10.1023/A:1004361819974
- Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: thirty-five years of validation. *Sociological Methods & Research*, 33, 319–348. doi:10.1177/0049124104268664
- Liu, P. T., & Chow, L. P. (1976). A new discrete quantitative randomized response model. *Journal of the American Statistical Association*, 71, 72–73. doi:10.2307/2285733
- Liu, P. T., Chow, L. P., & Mosley, W. H. (1975). Use of randomized response technique with a new randomizing device. *Journal of the American Statistical Association*, 70, 329–332.
- Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269–275. doi:10.2307/2285297
- Mangat, N. S. (1994). An improved randomized-response strategy. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 56, 93–95.
- Mangat, N. S., & Singh, R. (1990). An alternative randomized-response procedure. *Biometrika*, 77, 439–442. doi:10.1093/biomet/77.2.439
- Marquis, K. H., Marquis, M. S., & Polich, J. M. (1986). Response bias and reliability in sensitive topic surveys. *Journal of the American Statistical Association*, 81, 381–389. doi:10.2307/2289227
- Moors, J. J. A. (1971). Optimization of unrelated question randomized response model. *Journal of the American Statistical Association*, 66, 627–629.
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42, 42–54.
- Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology*, 61, 48–54. doi:10.1027/1618-3169/a000226
- Moshagen, M., Hilbig, B. E., & Musch, J. (2011). Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *European Journal of Social Psychology*, 41, 638–644. doi:10.1002/Ejsp.793
- Moshagen, M., & Musch, J. (2012). Surveying multiple sensitive attributes using an extension of the randomized-response technique. *International Journal of Public Opinion Research*, 24, 508–523.
- Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2010). Reducing socially desirable responses in epidemiologic surveys. An extension of the randomized-response technique. *Epidemiology*, 21, 379–382. doi:10.1097/Ede.0b013e3181d61dbc
- Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, 44, 222–231. doi:10.3758/s13428-011-0144-2 21858604
- Nakhaee, M. R., Pakravan, F., & Nakhaee, N. (2013). Prevalence of use of anabolic steroids by bodybuilders using three methods in a city of Iran. *Addict Health*, 5(3–4), 1–6.
- Ostapczuk, M., Moshagen, M., Zhao, Z., & Musch, J. (2009a). Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. *Journal of*

- Educational and Behavioral Statistics*, 34, 267–287. doi:10.3102/1076998609332747
- Ostapczuk, M., & Musch, J. (2011). Estimating the prevalence of negative attitudes towards people with disability: A comparison of direct questioning, projective questioning and randomised response. *Disability and Rehabilitation*, 33, 1–13. doi:10.3109/09638288.2010.492067
- Ostapczuk, M., Musch, J., & Moshagen, M. (2009b). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, 39, 920–931. doi:10.1002/ejsp.588
- Ostapczuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. *Statistical Methods in Medical Research*, 20, 489–503. doi:10.1177/0962280210372843
- Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17–59). San Diego, CA: Academic Press.
- Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, 60, 307–317. doi:10.1037/0022-3514.60.2.307
- Phillips, D. L., & Clancy, K. J. (1972). Some effects of social desirability in survey studies. *American Journal of Sociology*, 77, 921–940. doi:10.1086/225231
- Pitsch, W., Emrich, E., & Klein, M. (2007). Doping in elite sports in Germany: results of a www survey. *European Journal of Sport and Society*, 4, 89–102.
- Pollock, K. H., & Bek, Y. (1976). A comparison of 3 randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884–886. doi:10.2307/2286855
- Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W. C., & Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology*, 13, 465–484. doi:10.1002/(Sici)1099-0720(199910)13:5<465::Aid-Acp609>3.0.Co;2-Y
- reformiert. (2009). Mehrheit ist gegen ein Minarettverbot. Retrieved June 8th, 2010, from www.ref.ch
- Reinders, M. (1996). Häufigkeit von Namensanfängen. *Statistische Rundschau Nordrhein-Westfalen*, 11, 651–660.
- Savelkoul, M., Scheepers, P., van der Veld, W., & Hagendoorn, L. (2012). Comparing levels of anti-Muslim attitudes across Western countries. *Quality & Quantity*, 46, 1617–1624. doi:10.1007/s1135-011-9470-9
- Scheers, N. J. (1992). A review of randomized-response techniques. *Measurement and Evaluation in Counseling and Development*, 25, 27–41.
- Sheridan, L. P. (2006). Islamophobia pre- and post-September 11th, 2001. *Journal of Interpersonal Violence*, 21, 317–336. doi:10.1177/0886260505282885
- Silbermann, A., & Hüßers, F. (1995). *Der 'normale' Haß auf die Fremden. Eine sozialwissenschaftliche Studie zu Ausmaß und Hintergründen von Fremdenfeindlichkeit in Deutschland [The 'normal' xenophobia. A socio-scientific study on the extent and determinants of xenophobia in Germany]*. München: Quintessenz.
- Simon, P., Striegel, H., Aust, F., Dietz, K., & Ulrich, R. (2006). Doping in fitness sports: estimated number of unreported cases and individual probability of doping. *Addiction*, 101, 1640–1644. doi:10.1111/j.1360-0443.2006.01568.x
- Soeken, K. L., & Damrosch, S. P. (1986). Randomized-response technique - applications to research on rape. *Psychology of Women Quarterly*, 10, 119–125. doi:10.1111/j.1471-6402.1986.tb00740.x
- Soeken, K. L., & Macready, G. B. (1982). Respondents perceived protection when using randomized-response. *Psychological Bulletin*, 92, 487–489.
- Stocké, V. (2007). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 3, 125–138.
- Striegel, H., Ulrich, R., & Simon, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence*, 106, 230–232. doi:10.1016/j.drugalcdep.2009.07.026
- Sudman, S., & Bradburn, N. (1974). *Response effects in surveys*. Chicago: Aldine.
- Tian, G.-L., & Tang, M.-L. (2014). *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys*. Boca Raton, FL: CRC Press, Taylor & Francis Group.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 859–883. doi:10.1037/0033-2909.133.5.859
- Tracy, D. S., & Mangat, N. S. (1996). Some development in randomized response sampling during the last decade - a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4, 147–158.
- Umesh, U. N., & Peterson, R. A. (1991). A critical evaluation of the randomized-response method - applications, validation, and research agenda. *Sociological Methods & Research*, 20, 104–138.
- Vakilian, K., Mousavi, S. A., & Keramat, A. (2014). Estimation of sexual behavior in the 18-to-24-years-old Iranian youth based on a cross-wise model study. *BMC Research Notes*, 7(28), 1–4.
- van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (1998). A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of social security fraud. *Kwantitatieve Methoden*, 19, 15–34.
- van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (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.
- Warner, S. L. (1965). Randomized-response - a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63–69.
- Wolter, F., & Preisendörfer, P. (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:10.1177/0049124113500474
- Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika*, 67, 251–263. doi:10.1007/s00184-007-0131-x
- Zick, A., Küpper, B., & Hövermann, A. (2011). Intolerance, prejudice and discrimination: A European report. In N. Langenbacher (Ed.). Berlin: Friedrich-Ebert-Stiftung.