

University of Bern Social Sciences Working Paper No. 9

Sensitive Questions in Online Surveys: An Experimental Evaluation of the Randomized Response Technique and the Crosswise Model

Marc Höglinger, Ben Jann, and Andreas Diekmann

Current version: October 15, 2014
First version: May 22, 2014

<http://ideas.repec.org/p/bss/wpaper/9.html>
<http://econpapers.repec.org/paper/bsswpaper/9.htm>

Sensitive Questions in Online Surveys: An Experimental Evaluation of the Randomized Response Technique and the Crosswise Model

Marc Höglinger*

ETH Zurich, Clausiusstrasse 50, 8092 Zurich, Switzerland

phone: +41 44 632 5558, fax: +41 44 632 1054, email: marc.hoeglinger@soz.gess.ethz.ch

Ben Jann

University of Bern, Institute of Sociology, Fabrikstrasse 36, 3012 Bern, Switzerland

phone: +41 31 631 4831, fax: +41 31 631 4817, email: ben.jann@soz.unibe.ch

Andreas Diekmann

ETH Zurich, Clausiusstrasse 50, 8092 Zurich, Switzerland

phone: +41 44 632 5559, fax: +41 44 632 1054, email: andreas.diekmann@soz.gess.ethz.ch

October 2014

MARC HÖGLINGER is a Ph.D. candidate at the Chair of Sociology at the ETH Zurich, Switzerland. BEN JANN is Professor of Sociology at the University of Bern, Switzerland. ANDREAS DIEKMANN is Professor of Sociology at the ETH Zurich, Switzerland. The authors thank Debra Hevenstone for her comments on an earlier draft of this article. This research was supported by the German Research Foundation (DFG) Priority Program Number 1292 on “Survey Methodology” [DI 292/5 to Andreas Diekmann]. *Address correspondence to Marc Höglinger, ETH Zurich, Clausiusstrasse 50, 8092 Zurich, Switzerland, email: marc.hoeglinger@soz.gess.ethz.ch.

Sensitive Questions in Online Surveys: An Experimental Evaluation of the Randomized Response Technique and the Crosswise Model

Abstract

Self-administered online surveys provide a higher level of privacy protection to respondents than surveys administered by an interviewer. Yet, studies indicate that asking sensitive questions is problematic also in self-administered surveys. Because respondents might not be willing to reveal the truth and provide answers that are subject to social desirability bias, the validity of prevalence estimates of sensitive behaviors from online surveys can be challenged. A well-known method to overcome these problems is the Randomized Response Technique (RRT). However, convincing evidence that the RRT provides more valid estimates than direct questioning in online surveys is still lacking. A new variant of the RRT called the Crosswise Model has recently been proposed to overcome some of the deficiencies of existing RRT designs. We therefore conducted an experimental study in which different implementations of the RRT, including two implementations of the crosswise model, were tested and compared to direct questioning. Our study is a large-scale online survey ($N = 6,037$) on sensitive behaviors by students such as cheating in exams and plagiarism. Results indicate that the crosswise-model RRT—unlike the other variants of RRT we evaluated—yields higher prevalence estimates of sensitive behaviors than direct questioning. Whether higher estimates are a sufficient condition for more valid results, however, remains questionable.

1 Introduction

Obtaining accurate answers to sensitive questions is a persistent challenge to survey research. Respondents might misreport on sensitive questions and, hence, introduce systematic measurement error into survey data. Results from validation studies, that is, studies in which the researchers know the true answers, illustrate that the proportion of respondents who do not answer truthfully to questions on norm violations and deviant behavior can be substantial. For example, in a validation study by Preisendörfer and Wolter (2014), 42 percent (face-to-face interviews) and 33 percent (mail survey) of respondents did not admit that they were convicted in court. Likewise, 75 percent of respondents who committed welfare or unemployment benefit fraud denied having done so in face-to-face interviews by van der Heijden et al. (2000). As a consequence of such misreporting, the prevalence of sensitive behaviors is likely to be underestimated by population surveys and estimated correlations between sensitive characteristics and other variables might be biased.

The Randomized Response Technique

A well-known strategy to elicit truthful answers to sensitive questions is the Randomized Response Technique (RRT), introduced by Warner (1965). The idea behind the RRT is to protect the privacy of respondents by introducing random noise into their answers. Respondents who appreciate the anonymity induced by the procedure, it is assumed, are more inclined to provide truthful answers. A widely used RRT variant is the forced-response design proposed by Boruch (1971) and Greenberg et al. (1969, footnote p. 532), in which respondents employ a randomizing

device (e.g., dice, coins) to determine whether they should answer the sensitive questions (“yes” or “no”) or simply give an automatic “yes” or “no” response, irrespective of the true answer to the sensitive question. The result of the randomizing device is known only to the respondent, not to the researchers. Nonetheless, given the properties of the randomizing device, it is possible to infer the population prevalence of the sensitive behavior in question. A meta-analysis of 32 studies on the RRT in face-to-face or paper-and-pencil mode revealed that, on average, the RRT was successful in eliciting more valid estimates of sensitive behaviors and attitudes than direct questioning (Lensvelt-Mulders et al. 2005). Other studies, however, cast doubt on the validity of the RRT (e.g., Holbrook and Krosnick 2010; Wolter and Preisendörfer 2013). Furthermore, for self-administrated online mode, empirical evidence on the performance of the RRT is still scarce and inconclusive.

Sensitive-question techniques in online surveys

Online surveys, as well as other self-administered surveys such as paper-and-pencil interviews or interactive voice recognition (IVR), offer respondents more anonymity and privacy than interviewer-administered surveys. Therefore, effects of social desirability and perceived intrusiveness (Tourangeau, Rips, and Rasinski 2000), two main causes of potential misreporting, might be moderated. Conforming to that expectation, Kreuter, Presser, and Tourangeau (2008) found lower misreporting for several sensitive items in a validation study with university alumni for online mode compared to CATI. However, misreporting remained substantial also in online mode, indicating that the application of sensitive-question techniques such as the RRT could be valuable. Unfortunately, results from the few methodological studies comparing RRT to direct

questioning in online mode are not very promising. Holbrook and Krosnick (2010) found unrealistically high voter turnout estimates using the RRT in online mode. Likewise, Coutts and Jann (2011), Peeters (2005), and Snijders and Weesie (2008) found no improvement or even worse results with the RRT compared to direct questioning. We are aware of only one online survey in which an RRT variant actually outperformed direct questioning (de Jong, Pieters, and Fox 2010).

Reasons for the failure of the RRT

There are several reasons why implementations of the RRT might fail in online surveys. First, respondents' comprehension of the underlying principle, protection through randomization, is far from universal in most samples but seems crucial to elicit truthful answers (Landsheer, van der Heijden, and Gils 1999). In contrast to interviewer-administered surveys, it is difficult in online mode to provide respondents with additional assistance and tailored information about the sensitive-question procedure if required. But if respondents do not comprehend the RRT and, as a consequence, do not trust it, they might prefer to behave in a self-protective way and answer „no,“ irrespective of instructions. Second, in the RRT forced-response design, respondents might be reluctant to provide a “yes” answer if they did not engage in the sensitive behavior, as this might be perceived as giving a wrong answer or being forced to lie, or because they fear being falsely accused of something they did not do (Edgell, Himmelfarb, and Duchan 1982; Lensvelt-Mulders and Boeije 2007). Third, it is difficult to find a suitable randomizing device for online mode that is at respondents' immediate disposition, imposes no mode shift, and is perceived as trustworthy. Conventional devices such as dice or coins (Holbrook and Krosnick 2010; Coutts

and Jann 2011; de Jong, Pieters, and Fox 2010) are problematic because they require respondents to leave the computer and pause with the survey. This might induce respondents to refrain from applying the randomizing device or break off the interview. Furthermore, electronic devices such as virtual dice, virtual coins or a virtual random wheel (Snijders and Weesie 2008; Coutts and Jann 2011; Peeters 2005) can be manipulated or tracked by experimenters, and thus might not be judged trustworthy by the respondents. Because the randomizing devices employed in the existing studies did not solve these problems, it remains unclear whether the poor performance of the RRT in online mode is simply due to the lack of a suitable randomizing device.

The crosswise-model RRT

Yu, Tian, and Tang (2008) introduced the crosswise-model RRT as a promising alternative to conventional RRT schemes. In the crosswise-model RRT respondents are presented two questions at the same time: a sensitive question and an unrelated non-sensitive question. Respondents then have to indicate whether their answers to the two questions are the same (i.e. both „yes“ or both „no“) or different (i.e. one „yes“, one „no“). As long as the answer to the unrelated question is unknown, the respondent's answer to the sensitive question remains private. Again, however, prevalence estimation is feasible if the probability distribution of the non-sensitive question is known. That the crosswise-model RRT protects privacy is easy to understand by respondents since the possible answers, “the same” or “different”, are obviously ambiguous. Furthermore, there is no obvious self-protective answering strategy and no one is forced to give a “false” answer. Note that the crosswise-model RRT is formally equivalent to the

original RRT scheme by Warner (1965). However, it follows a different logic than the Warner scheme and appears qualitatively different from the perspective of the respondents.

A first empirical application of the crosswise-model RRT in a paper-and-pencil survey on paper plagiarism among students yielded significantly higher prevalence estimates compared to direct questioning (Jann, Jerke, and Krumpal 2012). Promising results are also reported by Shamsipour et al. (2014). However, the crosswise-model RRT has not yet been tested in online mode.

Our study

In our study we compare different variants of the RRT, including the crosswise model, to direct questioning in a large online survey on student misbehavior such as cheating in exams and plagiarism. One of the first empirical studies of student misconduct was carried out in the early 1960s at the Bureau of Applied Social Research in Columbia (Bowers 1964) and a series of similar studies followed (for reviews see: McCabe, Trevino, and Butterfield 2001; Crown and Spiller 1998). In continental Europe, concerns about student cheating and, in particular plagiarism, received increased attention as the Internet has provided growing opportunities for plagiarism—and, at the same time, new sophisticated tools for detecting plagiarism. Survey questions on exam cheating and paper plagiarism may thus raise social desirability concerns as well as worries about consequences in case of disclosure. The items in our survey cover different aspects of sensitivity (Tourangeau, Rips, and Rasinski 2000; Tourangeau and Yan 2007) and we expect substantial underreporting if the questions are asked directly. The RRT, if successful, should therefore yield higher estimates of the sensitive behaviors. In our study we evaluate a

total of five variants of the RRT, two implementations of the forced-response design, one implementation of the unrelated-question design, and two variants of the crosswise-model design with different randomizing devices, all of them tailored to online surveys and carefully designed and pretested.

2 Data and Methods

Online survey on cheating in exams and plagiarism

We conducted an online student survey with a randomized experimental design to test and compare the different sensitive-question techniques (Höglinger, Jann, and Diekmann 2014). The survey was administered in spring 2011 to all Bachelor's and Master's degree students enrolled at two major Swiss universities, the University of Bern and ETH Zurich. Students received an invitation email with a unique access link to a questionnaire on "Exams and written assignments" that included, among other questions, five sensitive questions. These questions covered behaviors such as copying from other students in an exam or handing in a plagiarized paper. Table 1 lists the five sensitive questions in the order they were presented to the respondents.¹ Admitting having cheated in an exam (first three questions) may not be viewed as highly sensitive, but disclosing plagiarism (last two questions) can have serious consequences such as the expulsion from the university or the annulment of an earned degree. Hence, we expect respondents are less inclined to admit plagiarism than cheating.

[Table 1 about here.]

¹ Respondents who did not yet have any exams or did not yet hand in a paper skipped the corresponding questions.

Three rounds of cognitive pretesting using the retrospective think-aloud technique (Presser et al. 2004) were carried out during questionnaire development ($N = 19$). The main focus of the cognitive pretests was on evaluating the respondents' compliance with the instructions to the sensitive-question techniques and their understanding of the mechanisms by which the techniques protect their privacy. All rounds of cognitive pretests were followed by questionnaire improvements. For instance, training questions and additional screens with explanations were introduced to improve respondents' understanding and minimize noncompliance. At each university we also ran a quantitative pretest prior to starting the main survey. At the University of Bern, a 10% sample of students ($N = 957$) was invited for the pretest on March 2, 2011, of which 28% completed the survey ($N = 269$). Based on respondents' comments and item non-response patterns some further improvements were made to the questionnaire. At ETH Zurich, a random sample of 500 students was invited for the pretest on Mai 2, 2011, of which 40% completed the survey ($N = 200$). As no further adjustments were made to the questionnaire after the pretest at ETH Zurich, we will include the answers of these respondents in the subsequent data analysis.²

On March 15 and March 16, 2011, the university's student administration office submitted the invitation emails for the main survey at the University of Bern.³ Out of 8,610 invited students, 2,486 completed the interview by May 24, 2011, when the survey was concluded. Another 287 students started the interview, but did not complete it. At ETH Zurich, the research team

² Excluding these observations does not change our main findings. Results without these observations are available in the online supplement.

³ Data protection regulation of the University of Bern did not allow the student administration office to provide the students' email addresses to the research team. Furthermore, it did not allow sending reminder emails to students who did not respond.

submitted the invitation emails for the main survey on May 9 and May 10, 2011. After three weeks, a reminder email was sent to students who had not yet responded. Including the pretest sample, 10,800 students were invited at the ETH Zurich, of which 4,005 completed the interview until June 20, 2011, when the survey was stopped. Another 576 students started the interview without completing it. In total, 19,410 students were invited, 6,491 completed the interview, and 863 started the survey without completing it (about half only looked at the first page of the questionnaire). Excluding the incomplete interviews, the overall response rate was 33.4% (RR1, AAPOR 2011).⁴ Median response time for the interviews was 12 minutes.

In the subsequent analysis we include all respondents who completed their interview at least to the point where the sensitive questions began (6,701 of 7,354 students). We also exclude the 392 respondents who skipped all sensitive questions because they had not yet had an exam and did not yet hand in a paper (or, in 4 cases, because of a technical failure). Finally, we exclude 272 respondents whose mother tongue is not German and who did not assess their German to be at least “good”.⁵ The resulting sample size is 6,037.

Experimental conditions

Respondents were randomly assigned to one of six experimental conditions: direct questioning, one of two implementations of the forced-response RRT, one implementation of the unrelated-question RRT, or one of two implementations of the crosswise-model RRT. Table 2 provides an

⁴ At the University of Bern, the response rate was considerably lower (28.9%) than at ETH Zurich (37.1%). The difference is due to the fact that no reminder emails were sent to students at the University of Bern.

⁵ The survey was only available in German and given the complexity of the instructions to the sensitive-question techniques we believe that it is sensible to exclude respondents whose German is poor. However, including these observations in the analysis does not change our main findings (results available in the online supplement).

overview of the six experimental conditions and their sample sizes. The wording of the sensitive questions was identical in all conditions. Due to item non-response and because not all respondents had to answer all sensitive questions (e.g., if they did not yet hand in a paper) sample sizes differ by experimental condition and question, an overview of which is provided in table A1 in the appendix.

[Table 2 about here.]

The direct questioning condition (DQ) served as a benchmark for the evaluation of the different variants of the RRT. A screen announcing several sensitive questions, stating the importance of honest answers for the success of the study, and providing a privacy assurance statement, preceded the sensitive questions. The five sensitive questions (see table 1) then followed one by one on separate screens. Each question could be answered with “yes” or “no”.

The first variant of the RRT used a forced-response design (Boruch 1971; Greenberg et al. 1969) and a virtual random wheel as randomizing device (FR Wheel). First, a screen announcing several sensitive questions and the use of a special technique to guarantee respondents’ privacy was displayed. Then, the procedure of the sensitive-question technique and how it protects respondents’ privacy was explained. The respondents then had to answer a training question about whether they had ever ridden public transit without paying the fare, which was followed by a screen with additional explanations on how the RRT protects the respondents’ answers. After that, the five sensitive questions followed one by one on separate screens.

For each question, respondents had to apply a virtual random wheel to generate a random instruction (figure 1). The random wheel had twelve sectors labeled “Question”, “Yes”, or “No”. Respondents could spin the wheel by clicking a “Rotate wheel” button. After stopping at a random position, the resulting instruction (“Answer Question”, “Directly tick Yes”, or “Directly tick No”) was displayed in the middle of the wheel (the wheel could only be spun once). Respondents were randomized between a low privacy protection distribution of random instructions (9 “Answer Question”, 1 “Directly tick No”, 2 “Directly tick Yes”) and a high privacy protection scheme (8 “Answer Question”, 2 “Directly tick No”, 2 “Directly tick Yes”).

[Figure 1 about here.]

The virtual random wheel corresponds to the classic spinner used in some early variants of the RRT (see Fox 1986, p. 39). Peeters (2005; also see Peeters, Lensvelt-Mulders, and Lasthuizen 2010) presented a first online implementation of such a spinner. We expect that the respondents do not trust the virtual random wheel because the outcome could easily be tracked or even predetermined (it was not). The same problems exist with virtual dice or coins (Coutts and Jann 2011; Snijders and Weesie 2008; Lensvelt-Mulders et al. 2006).

Because the virtual random wheel is not trustworthy, we developed a new randomizing device for online mode that cannot be tracked. This new randomizing device was used in our second variant of the forced-response RRT (FR Number), which was otherwise identical to the first RRT variant. The randomizing device worked as follows: Respondents were presented twelve fields on the screen, numbered from 1 to 12. They were told to privately choose a field and memorize

their choice (without clicking on it). Then, they were told to click a „Show instructions“ button to uncover the instructions hidden within the fields and follow the instruction that appeared in the field they chose (figure 2). As above, possible instructions were “Answer Question”, “Directly tick Yes”, or “Directly tick No”. The instructions were randomized across fields.

[Figure 2 about here.]

Our implementation of the unrelated-question RRT (UQ Benford) used a design with the Benford distribution of the first digits of house numbers as a randomizing device.⁶ In a first step, respondents were asked to think of an acquaintance and use the first digit of this person’s house number as their personal random number (figure 3). Then, for each sensitive item, respondents were asked to either answer the sensitive question or answer an unrelated auxiliary question, depending on their personal random number (figure 4; the used auxiliary questions are listed in the appendix). Again, respondents were randomized between two levels of privacy protection, a low protection scheme in which the sensitive question had to be answered if the personal random number was 1, 2, 3, 4, or 5 and a high protection scheme in which the sensitive question had to be answered if the personal random number was 1, 2, 3, or 4.

[Figure 3 about here.]

[Figure 4 about here.]

⁶ See Diekmann (2012) for the application of the Benford distribution as a simple randomizing device. Greenberg et al. (1969) first proposed the unrelated-question design for the RRT. For an overview see Fox and Tracy (1986).

Diekmann (2012) provides evidence that first digits of house numbers follow “Benford’s Law”. According to the law, the probability of 1, 2, 3, or 4 is 0.699; the probability of 1, 2, 3, 4, or 5 is 0.778. These probabilities are likely to be underestimated by respondents, so that the privacy protection by the procedure might be perceived higher than it actually is (called the “Benford illusion” by Diekmann).⁷

As in the other conditions, a screen announcing some sensitive questions and the use of a special technique to guarantee respondents’ privacy introduced the sensitive questions. Then, the screen where respondents had to select their personal random number appeared, which was followed by the screen for the first sensitive item. After the first sensitive item, a screen with some additional explanations was shown, which was followed by the screen for the remaining sensitive items. No training question was included in this condition.

Our first implementation of the crosswise-model RRT (CM Question) used an unrelated-question design as implemented in Jann, Jerke, and Krumpal (2012). For each sensitive item, respondents were presented two questions at the same time, the sensitive question and an unrelated non-sensitive question. Respondents were then instructed to indicate whether their answers to the two questions were the same (both “yes” or both “no”) or different (one “no”, the other “yes”) (figure 5). Again, respondents were randomized between two levels of privacy protection, using variations on the unrelated questions (see the appendix). The sequence of screens was similar as in UQ Benford except that the introductory screen already contained a brief explanation of the

⁷ Using two dice as randomizing device is a similar strategy since many respondents erroneously assume a uniform distribution of the added outcomes (Moriarty and Wiseman 1976).

crosswise-model RRT procedure and was directly followed by the screen for the first sensitive question.

[Figure 5 about here.]

Our second implementation of the crosswise-model RRT (CM Number) was analogous to FR Number, except that random answers (“Yes” or “No”) were included in the fields instead of forced-response instructions. Respondents were told to privately choose a field (without clicking on it) and then press a button to uncover the random answers in the fields. They then had to indicate whether the random answer in the field they chose was the same or different than their answer to the sensitive question (figure 6). Respondents were randomized between two levels of privacy protection. Furthermore, we varied whether “yes” or “no” was more frequent (which is formally arbitrary, but might be perceived differently). In the low protection scheme there were either 10 “yes” and 2 “no” or 2 “yes” and 10 “no”; in the high protection scheme there were either 9 “yes” and 3 “no” or 3 “yes” and 9 “no”.

[Figure 6 about here.]

Data analysis

Analysis of data collected by the RRT can be accomplished by means of simple variable transformations. Let Y be the observed outcome variable with $Y = 1$ if a respondent answers “yes” (or “the same” in the crosswise-model RRT) and $Y = 0$ if a respondent answers “no” (or

“different” in the crosswise-model RRT). Likewise, let S be the sensitive item with $S = 1$ if the sensitive item applies and $S = 0$ else. In the forced-response RRT, the respondents are instructed to answer “yes” with known probability p^{yes} , answer “no” with known probability p^{no} , or answer the sensitive question truthfully with probability $(1 - p^{yes} - p^{no})$. Assuming that respondents comply with the instructions, the overall probability of a “yes” answer in the forced-response RRT is

$$\Pr(Y = 1) = (1 - p^{yes} - p^{no}) \Pr(S = 1) + p^{yes}$$

where $\Pr(S = 1)$ is the unknown probability that sensitive item applies. Solving for $\Pr(S = 1)$ shows that taking the mean of

$$\tilde{Y} = \frac{Y - p^{yes}}{(1 - p^{yes} - p^{no})}$$

provides a consistent estimate of $\Pr(S = 1)$. The same transformation can also be employed for data from the unrelated-question RRT, setting $p^{yes} = p^u p^{yes,u}$ and $p^{no} = p^u (1 - p^{yes,u})$, where p^u is the known probability of being directed to the unrelated question and $p^{yes,u}$ is the known probability of a “yes” answer to the unrelated question. Finally, for the crosswise-model RRT, the corresponding transformation is

$$\tilde{Y} = \frac{Y + p^{yes,u} - 1}{(2p^{yes,u} - 1)}$$

where $p^{yes,u}$ is again the probability of a “yes” answer to the unrelated question.⁸

Standard methods can be used to estimate expected values from these transformed variables, yielding the same point estimates and standard errors as the basic formulas usually found in the RRT literature (Fox and Tracy 1986; Chaudhuri 2011). An equivalent approach, followed in the analyses below, is to estimate a least-squares regression on \tilde{Y} across the whole sample including dummy variables for the different sensitive-question techniques (with $\tilde{Y} = Y$ for direct questioning), employing heteroscedasticity robust formulas for standard errors (Jann 2008). Such an integrated model is convenient because it readily provides tests for differences among techniques. Furthermore, additional covariates can be included in the model to analyze effects of predictors of sensitive behaviors.⁹

3 Results

We first compare the sensitive-question techniques in terms of their resulting prevalence estimates for the sensitive behaviors. Assuming that respondents only falsely deny but never falsely admit a sensitive behavior, higher prevalence estimates from the sensitive-question techniques than from direct questioning (DQ) indicate that more respondents answered

⁸ As in the original Warner scheme, $p^{yes,u}$ must be unequal 0.5 for the crosswise-model RRT estimate to be identified.

⁹ An alternative approach would be to use suitably modified maximum-likelihood logistic regression (Jann 2005; also see Jann, Jerke, and Krumpal 2012 for the crosswise-model RRT). We prefer the linear regression approach here because it imposes fewer assumptions about the data generation process. For example, logistic regression may break down if respondents do not comply with the RRT instructions. Yet another approach is nonlinear least-squares estimation (e.g., Cameron and Trivedi 2005, chapter 5.8). Using maximum-likelihood logistic regression or nonlinear least-squares estimation does not change our main findings (results available in the online supplement; exceptions are noted in the text).

truthfully. Hence, we interpret a positive difference to DQ as evidence for a technique's superior validity ("more-is-better" assumption; Lensvelt-Mulders et al. 2005).

We start our analysis with results where we distinguish between DQ, the forced-response or unrelated-question RRT (FR/UQ), and the crosswise-model RRT (CM), but do not separate the single implementations. Such a simplification appears sensible, as there are major design differences between FR/UQ and CM. The left panel in figure 7 depicts the point estimates of the proportion of respondents admitting a particular sensitive behavior and the corresponding 95%-confidence intervals by technique (also see table A2 in the appendix). Differences in the prevalence estimates between FR/UQ or CM and DQ are shown in the right panel.

[Figure 7 about here.]

CM produced the highest prevalence estimates for all five sensitive questions. Furthermore, the difference between CM and DQ is highly significant for all items but the last ("handing in someone else's paper").¹⁰ The size of the absolute differences between CM and DQ follows a rough pattern with larger differences for high prevalence items and smaller differences for low prevalence items. Such a pattern is consistent with what we would expect from a successful sensitive-question technique that manages to elicit truthful answers from respondents who misreport when asked directly.

¹⁰ Analyzing the data using maximum-likelihood logistic regression as suggested in Jann, Jerke, and Krumpal (2012) leads to a significant difference also for the last item ($p = 0.023$; results available in the online supplement).

FR/UQ, in contrast, yielded higher prevalence estimates than DQ only for three items and only one of these differences is significant (“using crib notes in exam,” $p = 0.013$). Furthermore, for two items, FR/UQ produced lower estimates than DQ. One of these negative differences is statistically significant at the 5% level ($p = 0.018$ for “taking drugs to enhance exam performance”), the other has a p -value of 0.094 (“handing in someone else’s paper”). Such negative differences indicate that either trust is crowded out by the technique so that respondents who would admit a sensitive behavior in direct questioning refrain from doing so in RRT, or that respondents are reluctant to provide a “yes” if instructed to do so by the randomizing device and, hence, deviate from the RRT instructions.

Overall, CM yielded higher prevalence estimates than direct questioning, while the aggregated FR/UQ variants did not. However, as pointed out above, details of the implementation such as the choice of the randomizing device might affect respondents’ compliance with the sensitive-question procedures. In figure 8, we therefore present detailed results by implementations of the RRT (also see table A3 in the appendix). As sample sizes are smaller, variance among the results is generally higher and confidence intervals are wider than in figure 7. Nonetheless, some revealing patterns are observable.

[Figure 8 about here.]

For the crosswise-model RRT we see that the classic implementation using unrelated questions (CM Question) produced the highest prevalence estimates among all techniques for four out of five items. The difference to DQ is substantial for all items and highly significant for three of

them. The results for the second implementation of the crosswise-model RRT that used the pick-a-number device to generate a random answer (CM Number) are mixed. The DQ estimates are exceeded only for two items (statistically significant in just one case), the results for the remaining three items are very similar to the DQ estimates.

Results for the forced-response and unrelated-question RRT implementations are even more mixed. In only three out of 15 comparisons did one of the three implementations yield a significantly higher prevalence estimate than DQ, but then there were also three cases in which one of the three implementations produced significantly lower estimates than DQ. In fact, in these three cases the RRT yielded a negative prevalence estimate. This suggests that there was substantial noncompliance with the RRT instructions, that is, that many respondents answered “no” even though the procedure instructed them to respond “yes.” The forced-response RRT implementations (FR Wheel and FR Number) produced lower estimates than DQ in most cases. However, the unrelated-question RRT implementation (UQ Benford) yielded higher estimates than DQ for two items (statistically significant in one case), and produced very similar estimates to DQ for the remaining three items.¹¹

¹¹ As discussed above, the design parameters of the RRT were varied among respondents, leading to different levels of respondent protection. We found no evidence whatsoever that the level of respondent protection affected the respondents’ answers to the sensitive questions (results available in the online supplement). Note, however, that the variations in the design parameters were only moderate. Furthermore, we do find some evidence that the level of respondent protection affected the self-reported trust in the privacy protection by the survey (correlation: $r = 0.032$, $p = 0.026$) and the perceived protection of answers by the special technique (correlation: $r = 0.034$, $p = 0.019$) (see below for details on these variables). For CM Number, we additionally varied whether random answer “yes” or “no” was more frequent. Although formally arbitrary, we find weak evidence that this variation affected respondents’ behavior. Prevalence estimates tended to be somewhat higher in the condition in which “yes” was more frequent ($p = 0.027$ across all five sensitive questions). Note that in CM Question we used a design in which random answer “no” was always more frequent.

We now turn to the evaluation of the sensitive-question techniques on various alternative quality criteria such as item-nonresponse, ease of use, or respondents' understanding of the procedure. The left panel of figure 9 displays results for quality criteria available for all techniques including direct questioning, the right panel contains results from additional criteria available only for RRT (also see table A4 in the appendix).

[Figure 9 about here.]

Sensitive-question techniques place additional burden on respondents, which might lead to higher break-off rates and item non-response. In fact, we observe slightly increased break-off rates (measured as the proportion of respondents who reached the introductory screen for the sensitive questions but did not complete the interview) from about 1% for DQ to about 2% or 3% for the RRT (although the difference between DQ and UQ Benford is not statistically significant). Likewise, we observe slightly increased levels of item-nonresponse (measured as the proportion of sensitive questions that remained unanswered) from about half a percent for DQ to about 1% or 2% for the RRT (the difference between DQ and UQ Benford again being insignificant). We conclude that the sensitive-question techniques increase break-off and item non-response only slightly.

Of greater concern is the fact that the sensitive-question techniques require much more answering time than DQ (third graph on the left in figure 9). Answering time is measured as the median response time required to complete the five sensitive questions, including all screens with instructions and explanations. Using the RRT causes a threefold to fourfold increase in

median answering time (around 3 minutes for the whole block) compared to DQ (below 1 minute). Even if we exclude all instruction and training screens, using the RRT still causes a twofold to threefold increase in median answering time compared to DQ (not shown).

A crucial aspect of sensitive-question techniques is that they are supposed to increase respondents' trust in the protection of their privacy. After all, this is the primary motivation for these techniques. At the end of the interview, we asked the respondents about how much they trusted in the protection of privacy by the survey (see the appendix for the wording of the question). The fourth graph on the left in figure 9 shows the percentage of respondents who answered "rather much" or "very much." Levels of trust were significantly lower for all sensitive-question techniques (around 75%) than for DQ (over 80%). An explanation for this surprising finding might be that there is a crowding-out effect. The usage of a special technique raises suspicion and makes respondents aware of privacy concerns they might not have had if asked directly. In a way, using a special technique signals to the respondents that they should, in fact, be concerned. The crowding-out effect was highest for the forced-response RRT implementation with the virtual random wheel (below 70% trust), which makes sense since this randomization device is, in fact, not trustworthy. We also asked the respondents about how likely they thought it was that, based on this survey, one could discover whether an individual survey participant engaged in one of the sensitive behaviors. The lowest graph on the left in figure 9 displays the percentage of respondents who thought that such disclosure was "rather likely" or "very likely." For DQ the percentage was about 30%, which is significantly higher than for the RRT, with percentages between 20% and 25% (with the exception of the unrelated-question implementation of the crosswise-model RRT, for which the difference to DQ is not significant; p

= 0.087). Hence, even though general privacy concerns were lower among respondents in the DQ condition, they rightly judged the risk of disclosure to be higher in DQ than in the RRT conditions.

The plots on the right in figure 9 display additional results on a number of specific questions answered by respondents in the RRT conditions. We asked the respondents whether the employed technique was cumbersome, whether they thought that they applied the technique correctly, whether they were convinced that the technique protected their answers, whether they thought that the technique was a reasonable approach to protect respondents' privacy, and whether they believed that they understood how the technique protects their answers. The majority of respondents did not find the techniques cumbersome, but the percentage of respondents who answered that the technique was "rather" or "very" cumbersome was slightly higher in the conditions in which an explicit randomization device was employed (about 12% to 14%; FR Wheel, FR Wheel, CM Number) than in the conditions where no such device was used (between 8% and 10%; UQ Benford, CM Question). Furthermore, between 92% and 97% of respondents believed that they applied the technique correctly ("rather" or "definitely"); they seemed to have the least problems with CM Question, the most with FR Wheel. The third plot on the right in figure 9 shows the percentage of respondents who were convinced that the technique protects their answers ("rather" or "definitely"). As expected, the virtual random wheel was trusted least (57%), but also UQ Benford (62%) was trusted significantly less than the other implementations (67% to 75%), presumably because many respondents didn't understand its rationale (see below). Consequently, the respondents also deemed these two techniques least reasonable to protect respondents' privacy (fourth plot on the right in figure 9; shown is the

percentage of respondents who thought the technique was “rather” or “very” reasonable).

Finally, only between 57% and 66% of respondents claimed that they understood the rationale behind the techniques (“rather” or “definitely”). UQ Benford seems to be the implementation that was most difficult to understand.

We also analyzed correlations among the different quality criteria. Strongest correlations are found among the items measuring general trust in the survey, whether the technique protects one’s answers, whether the technique was considered reasonable, and whether the principle of the technique was understood. Most notably, understanding correlated with general trust ($r = 0.24$), protection ($r = 0.46$), and reasonableness ($r = 0.31$) (all correlations being highly significant with $p < 0.001$; computations based on dichotomized items as used for figure 9). This illustrates that a good understanding of a technique’s principle is crucial for developing trust in the technique’s privacy protection, which, we assume, is a precondition for increasing the likelihood of answering truthfully. Due to these associations, we conclude that levels of understanding of about 60% or 65%, as found in this study, are insufficient. Yet, when regressing the respondents’ answers to the sensitive questions on the level of trust we only find weak evidence for the assertion that trust increases the likelihood of admitting sensitive behaviors. Only for FR Wheel we find a marginally significant positive effect of trust ($p = 0.025$; using a joint test across all items).¹²

5 Discussion and Conclusions

¹² We also ran regressions on the other self-reported quality criteria (risk of disclosure, cumbersomeness, correct application, protection, reasonableness, understanding; results available in the online supplement). The only notable results were that, for RRT Benford, perceived cumbersomeness came along with increased prevalence estimates ($p < 0.001$) and correct application came along with decreased prevalence estimates ($p = 0.032$) (using joint tests across all five items).

Two main findings result from our study. First, the RRT in the forced-response or unrelated-question design, as implemented in our study, did not yield higher estimates than direct questioning. It often produced lower and sometimes even negative estimates. This questions the viability of the RRT for online surveys. The reason for the low RRT estimates might lie in respondents' noncompliance with the RRT instructions. More specifically, we assume that many respondents answer „no“ even if instructed to provide an automatic “yes,” because they are reluctant to give a false “yes” answer and always answering “no” is obviously the best self-protective answer strategy in the RRT.¹³ Although a lot of effort has been put into pretesting and finding good implementations of the RRT, no convincing evidence could be found that the technique yields more valid estimates than direct questioning. Even a completely anonymous randomizing device such as the pick-a-number procedure did not help to overcome the method's weaknesses. An exception was the Benford RRT implementation that performed somewhat better than the other RRT implementations.

Second, the crosswise-model RRT produced significantly higher prevalence estimates than direct questioning for four out of five sensitive items. Assuming the “more-is-better” assumption is valid, the crosswise-model RRT succeeded in eliciting more truthful answers to the sensitive questions than direct questioning and, hence, produced more valid estimates. The crosswise-model RRT, therefore, seems to be a promising alternative to the conventional RRT. Main advantages of the crosswise-model RRT are that no one is forced to provide a “false” answer and that the optimal self-protective answer strategy is far less obvious than for the forced-response or

¹³ This assumption is supported by estimates including noncompliance correction (see the appendix).

unrelated-question RRT.¹⁴ A drawback of the crosswise-model RRT compared to forced-response or unrelated-question RRT is its lower statistical efficiency (see the standard errors in table A3).

Our results also show that the performance of both the conventional RRT and the crosswise-model RRT seems to depend on the specific form of implementation, an aspect that is ignored in most other evaluation studies. Therefore, an important task for developing more reliable sensitive-question techniques for online mode is to figure out the implementation details that affect the validity of these techniques. In our study we emphasized the choice of randomizing device. For instance, the crosswise-model RRT implementation employing the unrelated-question design produced different results than the crosswise-model RRT implementation that used an explicit randomizing device. Moreover, for all evaluated implementations we found rather low levels of trust and understanding by respondents. In our view, this is problematic because trust and understanding are essential preconditions for increasing the likelihood of respondents answering truthfully.

A limitation of our study is that it is based on a sample of university students and results may not be generalizable to other populations. Furthermore, our aforementioned interpretation of the study's results rests on the "more-is-better" assumption, a limitation shared with most other studies on the RRT. Although the CM Question design produced the highest estimates, this does not necessarily mean that the results from the crosswise-model RRT are more valid than those from the other techniques. Higher estimates than DQ may be a necessary condition for the

¹⁴ Detection of the optimal self-protective answer strategy would require a thorough understanding of Bayesian updating and the crosswise-model RRT principle by respondents. If $p^{yes,u} < 0.5$, the optimal self-protective answer is "the same"; if $p^{yes,u} > 0.5$, the optimal self-protective answer is "different".

validity of a technique's results, given the sensitivity of the questions used in our study, but whether higher estimates are also a sufficient condition is questionable. It is possible that higher estimates do not come about due to an increased share of respondents who answer truthfully, but that some other mechanisms are at play. For example, in the crosswise-model RRT, if many respondents are confused and provide random answers, prevalence estimates will be biased towards 50% (although the pattern of results does not suggest that this is what is going on in our study; see the section on the effect of random answers in the appendix). Furthermore, as indicated above, the forced-response RRT scheme might perform well for "guilty" respondents, but estimates are distorted because the "innocent" are reluctant to provide a "false" positive answer and thus deviate from the RRT instructions. As it is difficult to disentangle the precise mechanisms based on studies in which true scores are not known, there is a clear need for validation studies. Initial results from a validation study we are currently running indicate that reliance on the more-is-better assumption might be inadequate, also for the crosswise-model RRT. Hence, we explicitly warn against drawing premature conclusions from comparative studies that do not offer the possibility to validate results against known values.

Eliciting truthful answers to sensitive questions remains a big challenge in online surveys. Although levels of misreporting seem to be somewhat lower than in interviewer-assisted surveys, the available validation studies show that also in online mode misreporting is substantial. Better strategies than direct questioning are necessary. That RRT approaches offer a viable solution cannot be confirmed without qualification by our study. However, the development and testing of such techniques in online mode is still at an early stage. Our study showed how resulting prevalence estimates depend on implementation details. That results differ so much by

implementation appears discouraging at first sight. In our view, however, it indicates that the RRT does have potential, if a good implementation can be found. Future studies should hence focus on identifying the factors that render an RRT implementation successful. From our results we conclude that a successful implementation should be nontechnical, easy to understand, and simple to apply, that no respondents should be forced into providing “false” positive answers, and that no obvious self-protective answering strategy should be available.

The possibilities to carry out validation studies, that is, studies in which the respondents’ answers can be compared to known true values, either at the aggregate or at the individual level, might be limited. Nonetheless, given their clear advantage over comparative studies in terms of explanatory power, it appears advisable to invest more research effort in that direction.

Data and supporting materials

The data and documentation of the survey are available online at <http://ideas.repec.org/p/bss/wpaper/8.html>. The analysis scripts and supplementary results are available online at <http://ideas.repec.org/p/bss/wpaper/9.html>.

Appendix

Auxiliary questions for UQ Benford (translated from German)

The auxiliary questions were:

- “Is your mother’s birthday in the months of January through June?” (0.521)
- “Is your mother’s birthday in an even-numbered month? (Feb., Apr., Jun., Aug., Oct., Dec.)” (0.495)
- “Is your mother’s birthday in the first half of the month? (from the 1st up to and including the 15th of the month)” (0.493)
- “Is your mother’s birthday on an even-numbered day? (2nd, 4th, 6th, etc. of the month)” (0.490)
- “Is your mother’s birth year even-numbered? (treat 0 as an even number)” (0.500)

The numbers in parentheses are our estimates of the probability that a question applies (based on the Swiss birth distribution between 1941 and 1965 as available from the Swiss Federal Statistical Office for the first two questions and based on a uniform distribution for the other questions). The questions were randomly paired with the sensitive questions for each respondent. Respondents were instructed to take the birthday of another person if they did not know their parent’s birthday (see figure 4).

Unrelated questions for CM Question (translated from German)

In the low privacy protection scheme the unrelated questions were:

- “Is your mother’s birthday in January or February?” (0.167)
- “Is your mother’s birthday between the 1st and the 6th of the month (inclusive)?” (0.197)
- “Is your father’s birthday in January or February?” (0.167)
- “Is your father’s birthday between the 1st and the 6th of the month (inclusive)?” (0.197)

- “Please think of your parents’ phone number (or the phone number of someone else you know): Is the last digit of this number equal to 1 or 2?” (0.202)

In the high privacy protection scheme the unrelated questions were:

- “Is your mother’s birthday in January, February, or March?” (0.260)
- “Is your mother’s birthday between the 1st and the 7th of the month (inclusive)?” (0.230)
- “Is your father’s birthday in January, February, or March?” (0.260)
- “Is your father’s birthday between the 1st and the 7th of the month (inclusive)?” (0.230)
- “Please think of your parents’ phone number (or the phone number of someone else you know): Is the last digit of this number equal to 1, 2, or 3?” (0.304)

The numbers in parentheses are our estimates of the probability that a question applies (based on the Swiss birth distribution between 1941 and 1965 as available from the Swiss Federal Statistical Office for the first and third question in each scheme, based on the Swiss phone directory of 2008 for the last, and based on a uniform distribution for the other questions). The questions were randomly paired with the sensitive questions for each respondent. Respondents were instructed to take the birthday of another person if they did not know their parent’s birthday (see figure 4)

Wording of questions on respondent’s evaluation of the survey (translated from German)

The questions and their answer options (in parentheses) were:

- Trust in anonymity: “Please be honest: How much do you trust in our measures for anonymity and privacy protection of the participants of this survey?” (“not at all,” “rather not,” “partly,” “rather much,” “very much”)
- Disclosure risk: “How likely do you think is it that based on this survey one can reconstruct whether a specific participant engaged in one of sensitive behaviors we asked about?” (“impossible,” “very unlikely,” “rather unlikely,” “rather likely,” “very likely”)
- Techniques is cumbersome: “How cumbersome was the application of this special survey technique to you?” (“not at all,” “slightly,” “somewhat,” “rather,” “very”)
- Applied technique correctly: “Do you think that you applied the special survey technique correctly in each case?” (“definitely not,” “rather not,” “partly,” “rather yes,” “yes, definitely”)
- Technique protects: “What is your personal opinion: Does the special survey technique provide 100% protection of your answers to the sensitive questions?” (“definitely not,” “rather not,” “partly,” “rather yes,” “yes, definitely”)
- Techniques is reasonable: “How reasonable do you think is the use if this survey technique to protect the answers of survey participants to sensitive questions?” (“not at all,” “slightly,” “somewhat,” “rather,” “very”)
- Understood principle: “Do you understand why the employed survey technique provides 100% protection of your answers?” (“definitely not,” “rather not,” “partly,” “rather yes,” “yes, definitely”)

Noncompliance-corrected prevalence estimates for the forced-response and unrelated-question

RRT

The noncompliance-correction model we employ is as follows (for related approaches see: Clark and Desharnais 1998; Moshagen and Musch 2012; Moshagen, Musch, and Erdfelder 2012). Let π be the proportion of respondents who committed the sensitive behavior and are ready to admit it in the RRT. These respondents are assumed to comply with the RRT instructions, providing a truthful “yes” or, depending on the outcome of the randomizing device, an automatic “yes” or an automatic “no.” The remaining $(1 - \pi)$ respondents either did not commit the sensitive behavior or are not ready to admit it in the RRT. They will answer “no” to the sensitive question and they will provide an automatic “no” if instructed to do so by the randomizing device. However, if the instruction is to provide an automatic “yes,” a proportion γ of them will not comply and answer “no” nonetheless. Given these assertions, the overall probability to observe a “yes” answer is $\Pr(Y = 1) = \pi(1 - p^{no}) + (1 - \pi)(1 - \gamma)p^{yes}$, where p^{yes} and p^{no} are as defined above. Standard maximum-likelihood methods can be used to estimate π and γ , as long as p^{yes} and p^{no} are varied among respondents (assuming that π and γ are independent from p^{yes} and p^{no}).

Pooling all RRT conditions, the noncompliance-corrected prevalence estimates for the five sensitive items are: 16.5 (6.5), 11.5 (6.1), 17.1 (5.6), 13.8 (6.6), 6.8 (5.9) (in percent; standard errors in parentheses). The noncompliance proportion γ is estimated as (in percent): -17.1 (34.8), -5.9 (31.6), 90.8 (37.2), 53.1 (39.5), 36.2 (31.9). As indicated above, the RRT performed particularly bad for the last three items. These are the items for which the noncompliance correction has a substantial effect, indicating that noncompliance with the instruction might, in fact, be the reason for the bad performance. However, also note that the estimates are only weakly identified and standard errors are large.

Effect of random answers in the crosswise-model RRT

If there is a constant share of respondents who randomly tick “the same” or “different” then one would expect that the absolute difference between estimates from DQ and crosswise-model RRT is similar across items. We observe, however, considerable variation in the absolute differences. Assuming that there is a proportion ρ of respondents providing random answers, the probability to observe answer “the same” can be written as $\Pr(Y = 1) = (p^{yes,u}\pi + (1 - p^{yes,u})(1 - \pi))(1 - \rho) + \rho/2$, where $p^{yes,u}$ is as defined above and π is the probability that the sensitive question is answered with “yes”. Assuming that the crosswise-model RRT does not make respondents more willing to admit sensitive behavior, we can plug-in the direct questioning estimates for π and then derive ρ , the proportion of random answers that would be required to drive the crosswise-model RRT estimate up to the observed level. Pooling both crosswise-model RRT conditions, we get values for ρ of 30% for the 1st item, 14% for the 2nd and 3rd, 11% for the 4th, and 3% for the 5th. Such variability in the share of random answers appears unrealistic (although learning might be an explanation). The hypothesis that the crosswise-model RRT estimates are driven by an increase of “yes” answers to the sensitive questions seems more plausible to us. Under this hypothesis we would roughly expect the observed pattern, with lower absolute differences for low prevalence items.

Tables

[Table A1]

[Table A2]

[Table A3]

[Table A4]

References

- AAPOR. 2011. *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition*. Lenexa, Kansas: The American Association for Public Opinion Research.
- Boruch, Robert F. 1971. "Assuring Confidentiality of Responses in Social Research: A Note on Strategies." *The American Sociologist* 6:308–311.
- Bowers, William J. 1964. *Student Dishonesty and Its Control in College*. New York: Columbia University.
- Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge & New York: Cambridge University Press.
- Chaudhuri, Arijit. 2011. *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Clark, Stephen J., and Robert A. Desharnais. 1998. "Honest Answers to Embarrassing Questions: Detecting Cheaters in the Randomized Response Model." *Psychological Methods* 3:160–168.
- Coutts, Elisabeth, and Ben Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)." *Sociological Methods & Research* 40:196–193.
- Crown, Deborah F., and M. Shane Spiller. 1998. "Learning from the Literature on Collegiate Cheating: A Review of Empirical Research." *Journal of Business Ethics* 17:683–700.

- de Jong, Martijn G., Rik Pieters, and Jean-Paul Fox. 2010. "Reducing Social Desirability Bias Through Item Randomized Response: An Application to Measure Underreported Desires." *Journal of Marketing Research (JMR)* 47:14–27.
- Diekmann, Andreas. 2012. "Making Use of "Benford's Law" for the Randomized Response Technique" *Sociological Methods & Research* 41:325–334.
- Edgell, Stephen E., Samuel Himmelfarb, and Karen L. Duchan. 1982. "Validity of Forced Responses in a Randomized-Response Model." *Sociological Methods & Research* 11:89–100.
- Fox, James Alan, and Paul E. Tracy. 1986. *Randomized Response: A Method for Sensitive Surveys*. Newbury Park, CA: Sage.
- Greenberg, Bernard G., Abdel-Latif A. Abul-Ela, Walt R. Simmons, and Daniel G. Horvitz. 1969. "The Unrelated Question Randomized Response Model: Theoretical Framework." *Journal of the American Statistical Association* 64:520–539.
- Höglinger, Marc, Ben Jann, and Andreas Diekmann. 2014. *Online Survey on "Exams and Written Papers". Documentation*. ETH Zurich and University of Bern (available from <http://ideas.repec.org/p/bss/wpaper/8.html>).
- Holbrook, Allyson L., and Jon 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.
- Jann, Ben. 2005. "rrlogit: Stata module to estimate logistic regression for randomized response data." Statistical Software Components S456203, Boston College Department of Economics.

- Jann, Ben. 2008. "rrreg: Stata module to estimate linear probability model for randomized response data." Statistical Software Components S456962, Boston College Department of Economics.
- Jann, Ben, Julia Jerke, and Ivar Krumpal. 2012. "Asking Sensitive Questions Using the Crosswise Model: Some Experimental Results." *Public Opinion Quarterly* 76:32–49.
- Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys." *Public Opinion Quarterly* 72:847–865.
- Landsheer, Johannes, Peter van der Heijden, and Ger van Gils. 1999. "Trust and Understanding, Two Psychological Aspects of Randomized Response." *Quality & Quantity* 33:1–12.
- Lensvelt-Mulders, Gerty J. L. M., and Hennie R. Boeije. 2007. "Evaluating Compliance with a Computer Assisted Randomized Response Technique: a Qualitative Study into the Origins of Lying and Cheating." *Computers in Human Behavior* 23:591–608.
- Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, and Cora J. M. Maas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods & Research* 33:319–348.
- Lensvelt-Mulders, Gerty J. L. M., Peter G. M. van der Heijden, Olav Laudy, and Ger van Gils. 2006. "A Validation of a Computer-Assisted Randomized Response Survey to Estimate the Prevalence of Fraud in Social Security." *Journal of the Royal Statistical Society Series A* 169:305–318.
- McCabe, Donald L., Linda Klebe Trevino, and Kenneth D. Butterfield. 2001. "Cheating in Academic Institutions: A Decade of Research." *Ethics & Behavior* 11:219–232.

- Moriarty, Mark, and Frederick Wiseman (1976). "On the Choice of a Randomization Technique with the Randomized Response Model." In *Proceedings of the Social Statistics Section*, 624–626. Washington, DC: American Statistical Association.
- Moshagen, Morten, Jochen Musch, and Edgar Erdfelder. 2012. "A Stochastic Lie Detector." *Behavior Research Methods* 44:222–231.
- Moshagen, Morten, and Jochen Musch. 2012. "Surveying Multiple Sensitive Attributes Using an Extension of the Randomized-Response Technique." *International Journal of Public Opinion Research* 24:508–523.
- Peeters, Carel F. W. 2005. "Measuring Politically Sensitive Behavior. Using Probability Theory in the Form of Randomized Response to Estimate Prevalence and Incidence of Misbehavior in the Public Sphere: a Test on Integrity Violations." PhD Dissertation, Faculty of Social Sciences, Vrije Universiteit, Amsterdam.
- Peeters, Carel F. W., Gerty J. L. M. Lensvelt-Mulders, and Karin Lasthuizen. 2010. "A Note on a Simple and Practical Randomized Response Framework for Eliciting Sensitive Dichotomous and Quantitative Information." *Sociological Methods & Research* 39:283–296.
- Preisendörfer, Peter, and Wolter Felix. 2014. "Who is Telling the Truth? A Validation Study on Determinants of Response Behavior in Surveys." *Public Opinion Quarterly* 78:126–146.
- Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer. 2004. "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly* 68:109–130.
- Shamsipour, Mansour, Masoud Yunesian, Akbar Fotouhi, Ben Jann, Afarin Rahimi-Movaghar, Fariba Asghari, and Ali Asghar Akhlaghi. 2014. "Estimating the Prevalence of Illicit

- Drug Use Among Students Using the Crosswise Model.” *Substance Use and Misuse* 49:1303–1310.
- Snijders, Chris, and Jeroen Weesie. 2008. “The Online Use of Randomized Response Measurement.” Paper presented at General Online Research 2008, Hamburg, Germany.
- Tourangeau, R., Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- Tourangeau, Roger, and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133:859–883.
- van der Heijden, Peter G. M., Ger van Gils, Jan Bouts, and Joop 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 and Research* 28:505–537.
- Warner, Stanley L. 1965. “Randomized-Response: A Survey Technique for Eliminating Evasive Answer Bias.” *Journal of the American Statistical Association* 60:63–69.
- Wolter, Felix, and Peter Preisendörfer. 2013. “Asking Sensitive Questions: An Evaluation of the Randomized Response Technique Versus Direct Questioning Using Individual Validation Data.” *Sociological Methods and Research* 42:321–353.
- Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. “Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis.” *Metrika* 67:251–263.

Tables

Table 1. Sensitive questions on student misconduct (translated from German)

Item	Wording
Copying from other students in exam	In your studies, have you ever copied from other students during an exam?
Using crib notes in exam	In your studies, have you ever used illicit crib notes in an exam (including notes on mobile phones, calculators or similar)?
Taking drugs to enhance exam performance	In your studies, have you ever used prescription drugs to enhance your performance in an exam?
Including plagiarism in paper	In your studies, have you ever handed in a paper containing a passage intentionally adopted from someone else's work without citing the original?
Handing in someone else's paper	In your studies, have you ever had someone else write a large part of a submitted paper for you or have you handed in someone else's paper as your own?

Table 2. Experimental conditions and number of observations

Experimental condition	Design	Randomizing device	<i>N</i>
DQ	direct questioning		1004
FR Wheel	forced-response RRT	random wheel	1010
FR Number	forced-response RRT	pick-a-number device	1014
UQ Benford	unrelated-question RRT	Benford procedure and unrelated question	998
CM Question	crosswise-model RRT	unrelated question	1008
CM Number	crosswise-model RRT	pick-a-number device	1003

Table A1. Number of valid answers by sensitive question and experimental condition

Sensitive question	Experimental condition					
	DQ	FR Wheel	FR Number	UQ Benford	CM Question	CM Number
Copying from other students in exam	979	972	980	977	983	968
Using crib notes in exam	979	971	979	973	980	965
Taking drugs to enhance exam performance	976	971	978	967	972	963
Including plagiarism in paper	723	720	718	722	725	710
Handing in someone else's paper	725	717	718	721	721	709

Table A2. Prevalence estimates by sensitive-question technique (in percent; standard errors in parentheses)

	Copying from other students in exam	Using crib notes in exam	Taking drugs to enhance exam performance	Including plagiarism in paper	Handing in someone else's paper
<i>Prevalence estimates</i>					
Direct questioning (DQ)	17.88 (1.23)	9.09 (0.92)	3.38 (0.58)	2.90 (0.62)	1.52 (0.45)
Forced-response or unrelated question RRT (FR/UQ)	19.60 (1.18)	12.69 (1.12)	0.74 (0.95)	3.86 (1.17)	-0.45 (1.09)
Crosswise-model RRT (CM)	27.42 (1.99)	14.65 (1.90)	9.96 (1.86)	8.02 (2.12)	3.16 (2.05)
<i>Differences</i>					
FR/UQ – DQ	1.73 (1.70)	3.60 (1.45)	-2.64 (1.11)	0.95 (1.32)	-1.97 (1.18)
CM – DQ	9.54 (2.34)	5.56 (2.11)	6.58 (1.95)	5.12 (2.21)	1.64 (2.10)
CM – FR/UQ	7.82 (2.32)	1.96 (2.20)	9.22 (2.09)	4.16 (2.42)	3.61 (2.32)
<i>N</i>	5859	5847	5827	4318	4311

Table A3. Prevalence estimates by experimental condition (in percent; standard errors in parentheses)

	Copying from other students in exam	Using crib notes in exam	Taking drugs to enhance exam performance	Including plagiarism in paper	Handing in someone else's paper
<i>Prevalence estimates</i>					
Direct questioning (DQ)	17.88 (1.23)	9.09 (0.92)	3.38 (0.58)	2.90 (0.62)	1.52 (0.45)
FR Wheel	22.80 (2.14)	11.28 (1.96)	-0.89 (1.67)	0.94 (2.01)	0.46 (2.00)
FR Number	18.78 (2.08)	13.86 (2.00)	-1.52 (1.64)	2.95 (2.07)	-4.25 (1.82)
UQ Benford	17.24 (1.91)	12.93 (1.83)	4.67 (1.63)	7.68 (1.98)	2.43 (1.81)
CM Question	30.06 (2.90)	18.37 (2.80)	15.26 (2.80)	7.61 (3.08)	6.12 (3.05)
CM Number	24.74 (2.73)	10.88 (2.56)	4.62 (2.45)	8.45 (2.92)	0.14 (2.73)
<i>Differences</i>					
FR Wheel – DQ	4.93 (2.47)	2.19 (2.17)	-4.27 (1.77)	-1.96 (2.10)	-1.06 (2.05)
FR Number – DQ	0.90 (2.41)	4.77 (2.20)	-4.90 (1.74)	0.04 (2.16)	-5.77 (1.88)
UQ Benford – DQ	-0.63 (2.27)	3.84 (2.05)	1.29 (1.73)	4.77 (2.08)	0.91 (1.87)
CM Question – DQ	12.18 (3.15)	9.28 (2.95)	11.88 (2.86)	4.70 (3.14)	4.60 (3.08)
CM Number – DQ	6.87 (2.99)	1.79 (2.72)	1.24 (2.52)	5.55 (2.99)	-1.38 (2.77)
<i>N</i>	5859	5847	5827	4318	4311

Table A4. Comparison of experimental conditions on various measures


	Break-off (%)	Item nonresponse (%)	Answering time (median in seconds)	Trust in anonymity (%)	Disclosure risk (%)
Direct questioning	1.20 (0.34)	0.55 (0.21)	33.00 (0.49)	80.61 (1.26)	28.82 (1.45)
FR Wheel	3.27 (0.56)	1.84 (0.40)	167.00 (2.12)	69.22 (1.48)	22.93 (1.35)
FR Number	2.76 (0.51)	1.91 (0.40)	162.00 (2.28)	73.15 (1.41)	19.49 (1.26)
UQ Benford	2.00 (0.44)	0.98 (0.27)	138.00 (1.71)	73.37 (1.41)	20.94 (1.30)
CM Question	2.78 (0.52)	1.39 (0.32)	116.00 (1.38)	76.37 (1.36)	25.38 (1.39)
CM Number	3.39 (0.57)	2.30 (0.45)	159.00 (2.12)	76.65 (1.36)	20.00 (1.28)
<i>N</i>	6037	6037	5961	5884	5874
	Technique is cumbersome (%)	Applied technique correctly (%)	Technique protects (%)	Technique is reasonable (%)	Understood principle (%)
FR Wheel	14.18 (1.12)	95.06 (0.70)	56.54 (1.59)	53.44 (1.60)	60.12 (1.57)
FR Number	12.99 (1.08)	92.41 (0.85)	67.35 (1.50)	59.28 (1.57)	66.16 (1.51)
UQ Benford	9.57 (0.94)	94.87 (0.71)	61.66 (1.56)	53.96 (1.60)	57.19 (1.59)
CM Question	8.59 (0.90)	97.03 (0.54)	67.42 (1.50)	59.90 (1.57)	62.22 (1.55)
CM Number	11.70 (1.03)	95.66 (0.66)	75.03 (1.39)	62.53 (1.56)	65.63 (1.53)
<i>N</i>	4867	4865	4862	4862	4865

Figures

Figure 1. Screen shot of the FR Wheel implementation (translated from German)

1. Please rotate the random wheel:

Rotate wheel



2. Now follow the instructions as indicated by the random wheel:

Answer Question
↓

Directly tick Yes
↓

Directly tick No
↓

In your studies, have you ever copied from other students during an exam?

Please tick the corresponding answer on the right →

Yes ☐

No ☐

Figure 2. Screen shot of the FR Number implementation (translated from German)

1. Please pick one of the twelve fields.

1 Answer Question	2 Directly tick Yes	3 Answer Question	4 Answer Question	5 Answer Question	6 Directly tick No	7 Answer Question	8 Answer Question	9 Answer Question	10 Directly tick Yes	11 Answer Question	12 Answer Question
--------------------------------	----------------------------------	--------------------------------	--------------------------------	--------------------------------	---------------------------------	--------------------------------	--------------------------------	--------------------------------	-----------------------------------	---------------------------------	---------------------------------

2. Now click the "Show instructions" button: [Show instructions](#)

3. Please follow the instruction displayed in the field you picked:

Answer
Question

↓

Directly tick
Yes

↓

Directly tick
No

↓

In your studies, have you ever copied from other students during an exam?
Please tick the corresponding answer on the right →

Yes
☐

No
☐

Figure 3. Screen shot of the UQ Benford implementation, screen 1 (translated from German)

Please generate a random number that determines whether you have to answer question A or question B on the subsequent screens:

- 1. For this purpose, think of an acquaintance of yours who doesn't live in your household and whose address and house number you know.**
- 2. Take the first digit of this person's house number (for instance "3" for number 3, number 37, or number 348).**
- 3. Memorize this digit - it is your personal random number for the following questions.**

Figure 4. Screen shot of the UQ Benford implementation, screen 2 (translated from German)

Please answer question A or question B according to your random number:

If your random number is 1, 2, 3, or 4 →

A In your studies, have you ever copied from other students during an exam?

If your random number is 5, 6, 7, 8, or 9 →

B Is your mother's birthday in the first half of the year (January to June)?
(If you don't know, please take the birthday of another person you know.)

☐ Yes

☒ No

Figure 5. Screen shot of the CM Question implementation (translated from German)

Question A: Is your mother's birthday in January or February?
(If you don't know, please take the birthday of another person you know.)

Question B: In your studies, have you ever copied from other students during an exam?

Compare your answers to the two questions: Are the answers the same or different?

☐ same (both Yes or both No)

☐ different (one Yes, and the other No)

Figure 6. Screen shot of the CM Number implementation (translated from German)

1. Please answer the following question for yourself:

In your studies, have you ever copied from other students during an exam?

2. Now generate a random answer by picking one of the twelve fields.

1 No	2 No	3 No	4 Yes	5 No	6 No	7 No	8 Yes	9 No	10 No	11 No	12 Yes
----------------	----------------	----------------	-----------------	----------------	----------------	----------------	-----------------	----------------	-----------------	-----------------	------------------

3. Please click the "Show random answer" button:

4. Compare your own answer with the random answer in the field you picked:
Are the answers the same or different?

☐ same (both Yes or both No)

☐ different (one Yes, and the other No)

Figure 7. Prevalence estimates and difference to DQ by sensitive-question technique

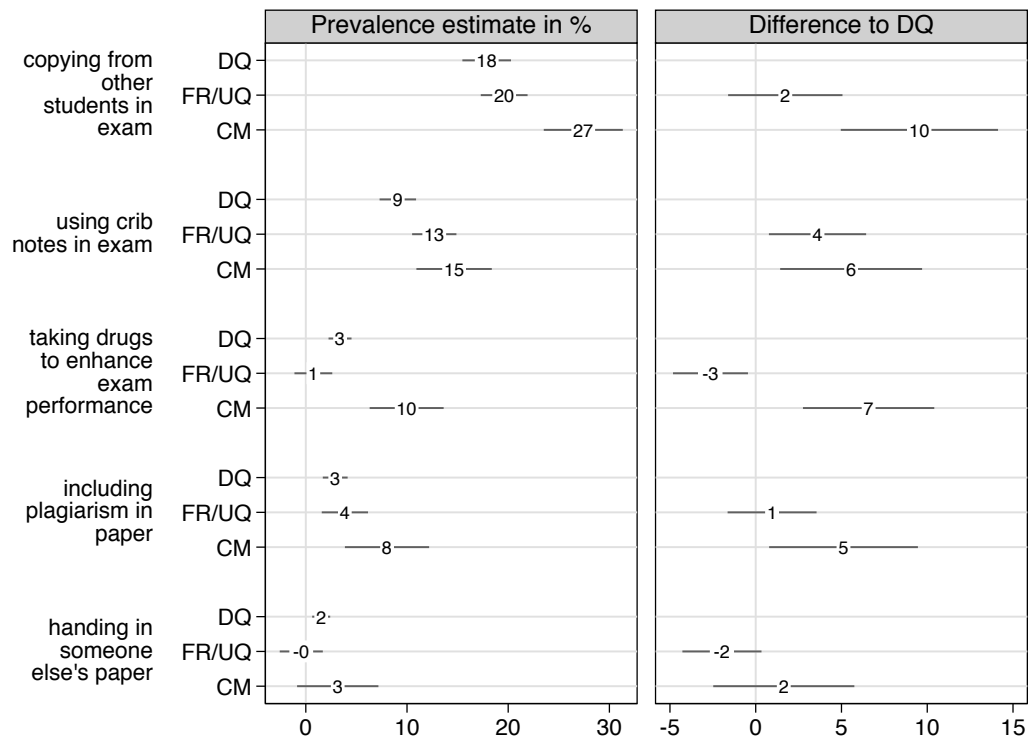


Figure 8. Prevalence estimates and difference to DQ by experimental condition

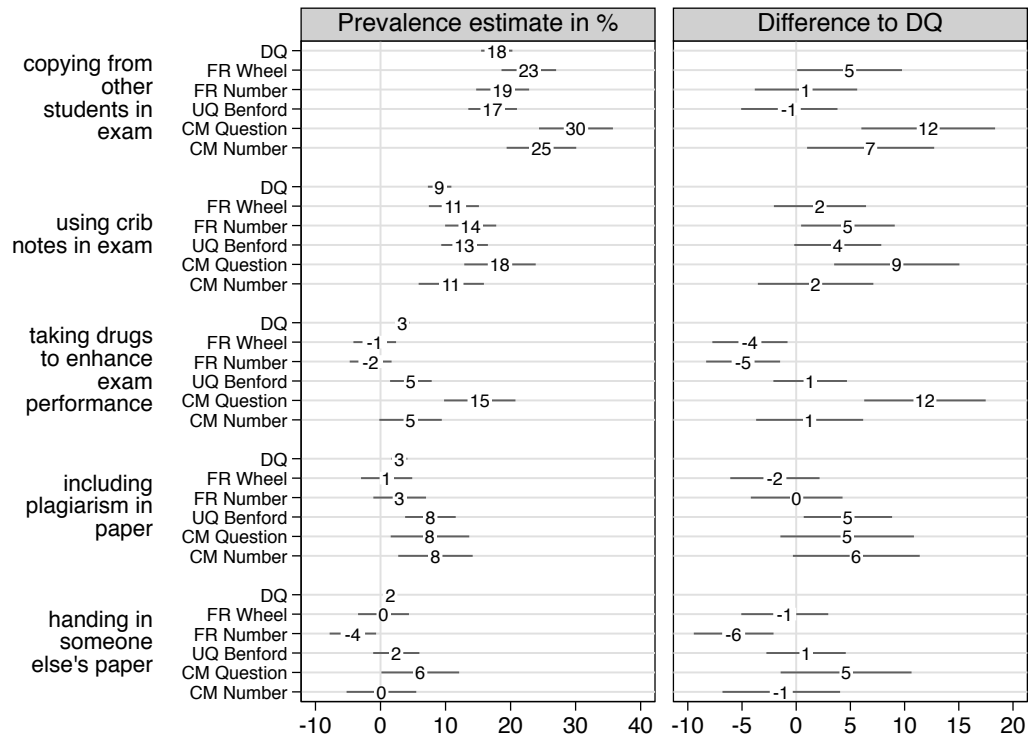


Figure 9. Comparison of experimental conditions on various measures

