UNIVERSITÄT
BERN

Faculty of Business, Economics and Social Sciences

Department of Social Sciences

University of Bern Social Sciences Working Paper No. 24

Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT

Marc Höglinger and Andreas Diekmann

This paper is forthcoming in *Political Analysis*.

December 15, 2016

http://ideas.repec.org/p/bss/wpaper/24.html
http://econpapers.repec.org/paper/bsswpaper/24.htm

University of Bern
Department of Social Sciences
Fabrikstrasse 8
CH-3012 Bern

Tel. +41 (0)31 631 48 11
Fax +41 (0)31 631 48 17
info@sowi.unibe.ch
www.sowi.unibe.ch

Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT

Marc Höglinger
Institute of Sociology, University of Bern, marc.hoeglinger@soz.unibe.ch

Andreas Diekmann
Chair of Sociology, ETH Zurich, diekmann@soz.gess.ethz.ch

December 15, 2016

Abstract

Validly measuring sensitive issues such as norm violations or stigmatizing traits through self-reports in surveys is often problematic. Special techniques for sensitive questions like the Randomized Response Technique (RRT) and, among its variants, the recent crosswise model should generate more honest answers by providing full response privacy. Different types of validation studies have examined whether these techniques actually improve data validity, with varying results. Yet, most of these studies did not consider the possibility of false positives, i.e. that respondents are misclassified as having a sensitive trait even though they actually do not. Assuming that respondents only falsely deny but never falsely admit possessing a sensitive trait, higher prevalence estimates have typically been interpreted as more valid estimates. If false positives occur, however, conclusions drawn under this assumption might be misleading. We present a comparative validation design that is able to detect false positives without the need for an individual-level validation criterion – which is often unavailable. Results show that the most widely used crosswise-model implementation produced false positives to a non-ignorable extent. This defect was not revealed by several previous validation studies that did not consider false positives – apparently a blind spot in past sensitive question research.

Keywords: Sensitive Questions, Sensitive Survey Techniques, Randomized Response Technique, Crosswise Model, Item Count Technique, Data Validity, Social Desirability, Measurement Error, Survey Design

Authors' note: We thank Ben Jann, the editor, as well as the two anonymous reviewers for their helpful comments, Thomas Hinz and Sandra Walzenbach for pointing us to potential problems of the crosswise-model RRT, and Murray Bales for proofreading the manuscript. For replication data see Höglinger and Diekmann (2016).

1. Introduction

Measurements of sensitive issues such as extreme political attitudes, deviant behavior, or stigmatizing traits through self-reports in surveys are often not reliable. Validation studies show that a considerable share of respondents falsely denies sensitive behavior when asked about it (e.g. Preisendörfer and Wolter 2014). Despite this serious flaw, research in deviance, political science, epidemiology, and many other areas relies heavily on self-report data. Finding ways to validly measure sensitive items is, therefore, very important.

Special techniques for sensitive questions such as the Randomized Response Technique (RRT, Warner 1965) are supposed to provide more valid data. Using some randomization procedure, such as dice, that introduces noise into the response process, this technique grants respondents full response privacy. While theoretically compelling, respondents in practice sometimes do not trust the special technique and still misreport. Alternatively, they do not comply with the relatively special and complicated RRT procedure. Hence, the RRT does not necessarily improve data quality. While a widely-cited meta-analysis (Lensvelt-Mulders et al. 2005) concluded that the RRT generates more valid data, the literature is not short of examples where RRT applications did not work as well as expected (e.g. Coutts and Jann 2011; Holbrook and Krosnick 2010; Höglinger, Jann, and Diekmann 2016; Wolter and Preisendörfer 2013).

The recently proposed crosswise-model RRT variant (Yu, Tian, and Tang 2008) has some desirable properties that should overcome certain problems found in other RRT variants. Recent applications include surveys on corruption and involvement in narcotics trade (Corbacho et al. 2016; Gingerich et al. 2015) or a survey on illicit drug use in Iran (Shamsipour et al. 2014). In the crosswise model, respondents are asked two questions simultaneously, a sensitive one (e.g. "Are you an active member of the Egyptian Muslim Brotherhood?") and a non-sensitive one (e.g. "Is your mother's birthday in January or February?"). Respondents do not indicate their answers to the two questions but only whether their two answers were identical (two times "yes", or two times "no") or different (one "yes", the other "no"). Because a respondent's answer to the non-sensitive question is unknown, an "identical" or "different" response does not reveal their answer to the sensitive question. However, as the overall prevalence of a "yes" answer to the birthday question is known, the collected data can be used for analysis by taking the systematic measurement error introduced by the special procedure into account. Compared to other RRT variants, the crosswise model is relatively easy to explain and does not need an explicit randomizing device which makes it especially suitable for self-administered survey modes such as paper-and-pencil or online. Further, the response options "identical" and "different" are obviously ambiguous which circumvents the problem encountered in some forced-response RRT implementations whereby distrustful respondents unconditionally choose the "no" response irrespective of the RRT instructions or their true

answer (Coutts et al. 2011). And, indeed, the crosswise model has been judged favorably in a series of validation studies because it elicited higher and seemingly more valid prevalence estimates of sensitive behavior or attitudes than direct questioning (Hoffmann and Musch 2015; Jann, Jerke, and Krumpal 2012; Korndörfer, Krumpal, and Schmukle 2014; Shamsipour et al. 2014; Hoffmann et al. 2015; Gingerich et al. 2015).

However, we argue that these results must be interpreted with great care because these validations had severe limitations. The majority of RRT evaluations are *comparative validation studies* where prevalence estimates of special sensitive question techniques and standard direct questioning (DQ) are compared under the more-is-better assumption: Assuming that respondents only falsely deny but never falsely admit an undesirable sensitive trait or behavior, higher prevalence estimates are interpreted as more valid estimates (e.g. Lensvelt-Mulders et al. 2005).[1] The more-is-better assumption is plausible for items that are unequivocally judged as socially undesirable, and where underreporting is the only likely source of misreporting. However, the social desirability of some items such as cannabis use or the number of sexual partners might be interpreted in the completely opposite way by a different subpopulation (e.g. Smith 1992). Moreover, some respondents actually might falsely admit sensitive behavior, i.e. they respond as if they possess a sensitive trait although they do not. We call this type of misreporting false positives. While quite unlikely for direct questioning, the occurrence of false positives cannot be ruled out a priori with special sensitive question techniques that require respondents to follow complex procedures. First, intentional or unintentional non-compliance with the RRT procedure likely leads to false negatives as well as false positives. Second, because the RRT guarantees full response privacy, respondents might be more prone than in the direct questioning mode to answer carelessly, including falsely giving a socially undesirable response. If false positives occur, however, the more-is-better assumption is no longer tenable since a higher prevalence estimate of a socially undesirable trait might not be the result of more but of less valid data.

Aggregate-level validation studies that compare estimated prevalence estimates to a known aggregate criterion such as official voting turnout rates (Rosenfeld, Imai, and Shapiro 2015) are preferable because they do not need the direct questioning estimate as a benchmark. However, they too do not allow a final conclusion to be drawn about a sensitive question technique's validity because if the sensitive question technique under investigation produces false negatives as well as false positives, both errors level each other out to an unknown degree. Hence, a seemingly more accurate estimate on the aggregate level might not be the result of more valid data on the individual level. Only *individual-level validations*, i.e. studies that compare self-reports to observed behavior or traits at the individual level, have the potential to identify false negatives as well as false

[1] This assumption is alternatively called "one sided lying", see e.g. Corbacho et al. (2016). The same holds, albeit in the opposite direction, for desirable traits or behaviors (less-is-better applies then).

positives. However, for many topics or items of interest they are impossible to carry out because one needs a validation criterion from typically hard-to-access sources such as sensitive individual record data. As a consequence, individual-level validations are rare, usually deal with special populations, and often cannot be replicated. Moreover, many do not consider false positives in their analysis even though they could (see online Appendix A for details).

Given that one reason for the apparent blind spot in sensitive question research is the difficulty of carrying out individual-level validation studies, we propose an alternative comparative design which is able to detect systematic false positives without needing an individual-level validation criterion. This is achieved by introducing one or more zero-prevalence items among the sensitive items. If a sensitive question technique systematically leads to false positives, the estimates of the zero-prevalence items will be non-zero and the more-is-better assumption is no longer tenable. If, however, the estimates for the zero-prevalence item are correct, and thus no false positives are produced, relying on the more-is-better assumption is warranted on much firmer ground.

We present results of an application of such an enhanced comparative validation in a survey on "Organ donation and health" ($N = 1,685$). Questions on having received a donor organ and on having suffered from Chagas disease, two items with nearly zero prevalence in the surveyed population, served as zero-prevalence items. The results show that what is currently the most widely used implementation of the crosswise-model RRT produced positive, i.e. wrong, prevalence estimates of the zero-prevalence items, and hence generated false positives to a non-ignorable extent.

2. Data and design

Our analysis sample consisted of $1,685$ members of a non-representative German online access panel that took part in a survey on "Organ donation and health".[2] To validate the sensitive question techniques we asked respondents a series of five health-related items with varying degrees of sensitivity: a question on whether they had ever donated blood, on their willingness to donate organs after death, on excessive drinking in the last two weeks, on whether they had ever received a donated organ, and on whether they had ever suffered from Chagas disease (Table 1). The last two items "ever received a donated organ" and "ever suffered from Chagas disease" have a close to zero prevalence in the surveyed population and are used to test for systematic false positives.

One-third of the respondents were randomly assigned to the direct questioning (DQ) version of the sensitive questions, and two-thirds to the crosswise-model variant (CM).[3] The crosswise-model RRT implemented was an unrelated question version as used in Jann, Jerke, and Krumpal (2012)

[2] See online Appendix B for data and design details, and Höglinger and Diekmann (2016) for replication data.
[3] To counterbalance the lower statistical efficiency of the CM.

Table 1: Sensitive questions surveyed

Item	Wording
Never donated blood*	"Have you ever donated blood?"
Unwilling to donate organs*	"Are you willing to donate your organs or tissues after death?"
Excessive drinking	"In the last two weeks, have you had five or more drinks in a row (a drink is a glass of wine, a bottle of beer, etc.)?"
Received a donated organ	"Have you ever received a donated organ (kidney, heart, part of a lung or liver, pancreas)?"
Suffered from Chagas disease	"Have you ever suffered from Chagas disease (Trypanosomiasis)?"

* Reverse coded for the purpose of analysis

and in most other previous studies using the crosswise model. Respondents were asked two questions at the same time: A sensitive question and an unrelated non-sensitive question. Respondents then had to indicate whether their answers to the two questions were identical or different. Due to the mixing with the non-sensitive question, a respondent's answer to the sensitive question remains completely private. The CM procedure was carefully introduced to the respondents and a practice question preceded the sensitive items which were asked in randomized order.

3. Results

For the comparative validation we estimated the self-report prevalence of the surveyed sensitive items for direct questioning (DQ) and the crosswise model (CM), as well as the corresponding difference (Figure 1).[4] The CM prevalence estimates are not significantly different to DQ for the item "never donated blood", but 5 percentage points higher for "unwilling to donate organs" (albeit not at a conventional significance level, $p = 0.066$), and 12 percentage points higher for "excessive drinking". This fits the pattern found in previous studies where the CM consistently produced higher prevalence estimates of sensitive behavior than DQ, which was typically interpreted as more valid estimates.

Looking at the two zero-prevalence items "ever received a donated organ" and "ever suffered from Chagas disease", we see that the DQ estimates are zero, as expected. In contrast, the corresponding CM estimates are with 8% (received organ) and 5% (Chagas disease) substantially and significantly above zero. The respective false positive rates of 8% and 5% reveal a non-ignorable

[4]For estimation we transformed the CM response variable to correct for the systematic error introduced by the randomization procedure and performed a least-squares regression with robust standard errors (see online Appendix B for details).

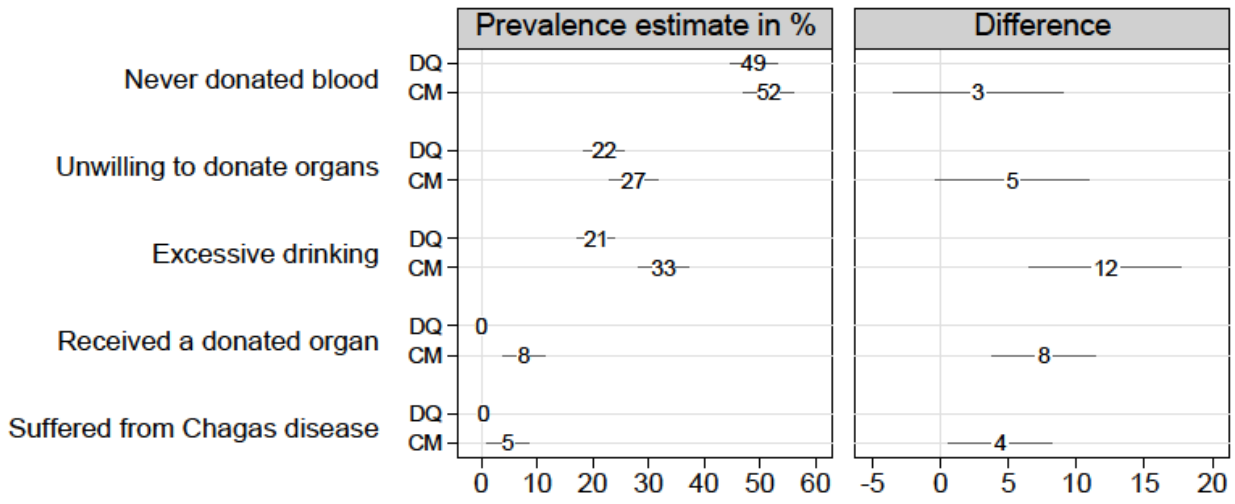

Figure 1: Comparative validation of sensitive question techniques (lines indicate a 95% confidence interval, *N* from 518 to 549 for DQ, and from 1,120 to 1,123 for CM)

amount of misclassification that cannot be explained by random error or by respondents' ignorance of their true status because, in the latter case, also the DQ estimates would deviate from zero.[5] The CM's inaccurate prevalence estimates are largely due to a false positive bias caused by this special sensitive question technique.[6] The more-is-better assumption is obviously not tenable for the CM. Hence, the CM's higher prevalence estimates for being unwilling to donate organs after death and for excessive drinking must not be interpreted as being the result of more respondents honestly giving the correct socially undesirable answer and of more valid data.

In addition, we carried out an individual-level validation using a barely sensitive question on whether respondents had (not) completed the "Abitur", the German general university entrance qualification. Answers were validated using previously collected self-report information. While some limitations apply to this validation, the found false positive rate of 7% corroborates the findings from the zero-prevalence comparative validation above. Most interestingly, the misclassification of the CM was not revealed in an aggregate-level validation we simulated. The aggregate prevalence estimate did not deviate significantly from the true value because the false negatives and false positives canceled each other out almost completely. This demonstrates the weakness of even an aggregate-level validation strategy (see online Appendix C for details).

Finally, we investigated the causes and correlates of false positives in the CM. However, the data did not reveal any pattern that would clearly point to a particular explanation we tested. We could, however, identify some candidate causes of false positives whose effect should be investi-

[5]None of the 548 respondents indicated having received a donated organ in the DQ condition, only 2 of 547 respondents indicated having suffered from Chagas disease.

[6]See below and online Appendix C for a discussion of potential causes such as random answering, problematic unrelated questions, or omitting a "don't know" response option.

gated more systematically in future studies: Some problematic, unrelated questions possibly not producing the expected "yes" answer probability, omitting a "don't know" response option, and respondents speeding over the CM instructions. Still, each of these factors accounts for only a share of the false positives that occurred and, very likely, the resulting false positive rate was caused by a mix of different mechanisms (see online Appendix C for details).

4. Discussion and conclusion

We introduced an enhanced comparative sensitive question validation design that is able to detect false positives and thereby allows for testing the more-is-better assumption on which comparative validations rely. The suggested design does not need an individual-level validation criterion, making it easily applicable in a broad array of substantive survey topics and populations of interest. Systematic false positives are detected by introducing one or more (near) zero-prevalence items among the sensitive items surveyed with a particular sensitive question technique.

Validating an implementation of the recently proposed crosswise-model RRT (CM) we found that the CM produced false positives to a non-ignorable extent. Our evidence is based on a comparative validation with zero-prevalence items and an additional individual-level validation using a non-sensitive question. Previous validation studies appraised the crosswise model for its easy applicability and seemingly more valid results. However, none of them considered false positives. Our results strongly suggest that in reality the crosswise model as implemented in those studies does not produce more valid data than DQ.

Further, our validation design allowed us to analyze various potential causes and correlates of false positives. For instance, by excluding responses elicited using some potentially problematic unrelated questions, false positives could be reduced considerably for one item. Still, this as well as other candidate causes could account for only a share of the false positives that actually occurred, suggesting that a mix of mechanisms might be responsible for the substantial amount of false positives. Possibly, better designed crosswise-model implementations are less plagued by false positives. Most conveniently, our validation design allows for testing such design improvements in an easy and reproducible way.

Note that the comparative validation with a zero-prevalence item only detects false positives if they occur systematically across different items. In this sense, it allows for a limited, but still much more meaningful validation than the comparative and aggregate-level validations used so far. To draw final conclusions regarding the validity of a particular technique, it should be complemented by individual-level validation studies. However, the fact that the presented design does not need a hard to achieve individual validation criterion makes it an easy and broadly applicable tool for developing and evaluating special sensitive question techniques and even for sensitive question

research in general.

To conclude, in our view the main lesson from this study is not so much that the crosswise-model RRT we implemented did not work as expected but that, had we not considered false positives in our analysis, we would have never revealed this fact. False positives might also occur in other RRT variants, and even with other sensitive question techniques such as the item count technique, forgiving wording or other question format changes. Because validation studies have so far largely neglected this possibility, we simply do not know. Sensitive question research must stop relying blindly on the more-is-better assumption and explicitly consider the possibility of false positives. The zero-prevalence comparative validation presented here as well as some recently proposed experimental individual-level validation strategies (e.g. Höglinger and Jann 2016) provide useful tools for overcoming this blind spot in future studies.

References

Corbacho, Ana, Daniel Gingerich, Virginia Oliveros, and Mauricio Ruiz-Vega. 2016. "Corruption as a Self-Fulfilling Prophecy: Evidence from a Survey Experiment in Costa Rica". *American Journal of Political Science* (online first).

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* 40:169–193.

Coutts, Elisabeth, Ben Jann, Ivar Krumpal, and Anatol-Fiete Näher. 2011. "Plagiarism in Student Papers: Prevalence Estimates Using Special Techniques for Sensitive Questions". *Journal of Economics and Statistics* 231:749–760.

Gingerich, Daniel W., Virginia Oliveros, Ana Corbacho, and Mauricio Ruiz-Vega. 2015. "When to protect? Using the crosswise model to integrate protected and direct responses in surveys of sensitive behavior". *Political Analysis*: online first.

Hoffmann, Adrian, Birk Diedenhofen, Bruno Verschuere, and Jochen Musch. 2015. "A Strong Validation of the Crosswise Model Using Experimentally-Induced Cheating Behavior". *Experimental Psychology* 62:403–414.

Hoffmann, Adrian, and Jochen Musch. 2015. "Assessing the Validity of Two Indirect Questioning Techniques: A Stochastic Lie Detector versus the Crosswise Model". *Behavior Research Methods* (online first).

Höglinger, Marc, and Andreas Diekmann. 2016. *Replication Data for: Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT*. Harvard Dataverse. doi:10.7910/DVN/SJ2RP1.

Höglinger, Marc, and Ben Jann. 2016. *More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model*. University of Bern Social Sciences Working Paper No. 18. University of Bern. https://ideas.repec.org/p/bss/wpaper/18.html.

Höglinger, Marc, Ben Jann, and Andreas Diekmann. 2016. "Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model". *Survey Research Methods* 10 (3): 171–87. doi:10.18148/srm/2016.v10i3.6703.

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, Julia Jerke, and Ivar Krumpal. 2012. “Asking Sensitive Questions Using the Crosswise Model. An Experimental Survey Measuring Plagiarism”. *Public Opinion Quarterly* 76:32–49.

Korndörfer, Martin, Ivar Krumpal, and Stefan C. Schmukle. 2014. “Measuring and Explaining Tax Evasion: Improving Self-Reports Using the Crosswise Model”. *Journal of Economic Psychology* 45:18–32.

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.

Preisendörfer, Peter, and Felix Wolter. 2014. “Who is Telling the Truth? A Validation Study on Determinants of Response Behavior in Surveys”. *Public Opinion Quarterly* 78:126–146.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2015. “An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions”. *American Journal of Political Science*: (online first).

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 & Misuse* 49:1303–1310.

Smith, Tom W. 1992. “Discrepancies between Men and Women in Reporting Number of Sexual Partners: A Summary from Four Countries”. *Social Biology* 39:203–211.

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 vs. Direct Questioning Using Individual Validation Data”. *Sociological Methods & 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.

Online Appendix

Online Appendix to "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT". For the full survey documentation, including questionnaire screenshots and data as well as extended analyses and analysis scripts see Höglinger and Diekmann (2016).

Contents

A. Recent individual-level validation studies

Of the handful of RRT individual-level validations published since 2000 only Höglinger and Jann (2016) and John et al. (2016) actually considered false positives in their analysis. The others surveyed only "guilty" respondents, i.e. true positives, which inhibits testing for false positives (van der Heijden et al. 2000; Moshagen et al. 2014; Wolter and Preisendörfer 2013), or used designs that allowed for identifying false positives to be identified in principle, but did not make use of this opportunity (Hoffmann et al. 2015; Kirchner 2015). This, too, indicates a profound lack of awareness of the potential occurrence of false positives in sensitive question research.

B. Design, data, and analysis details

Sample and survey details

Respondents were members of the PsyWeb-Panel, a non-representative online access panel administered by three German universities (see `https://psyweb.uni-muenster.de`). Of 10,000

members invited by email, 1,722 accessed our online questionnaire on "Organ donation and health" consisting of various questions on organ donation attitudes and behavior and containing an experimental information treatment on beliefs related to organ donation willingness.[1] After excluding one respondent who assessed his language skills (in German) as "rather poor"[2], we were left with 1,685 respondents who completed the survey part containing the sensitive questions. The median response time was 10.4 minutes, with the questionnaire version using the crosswise model taking one minute longer than the one using direct questioning. Break-off rates were almost identical for both the DQ version with 4% and the crosswise model (CM) with 5%. The sample consisted of German residents, with a median age of 47 years, 64% females, 54% married or living together with a partner, and 96% with German citizenship. Further, 46% worked full-time, 20% part-time, 5% were occasionally employed, 7% in training, and 22% not employed or on leave, while 13% were university students. Their educational background was quite above-average with 76% having completed the general or subject-specific university entrance qualification (about equivalent to a High School diploma).

The sensitive question techniques implemented

To validate the sensitive question techniques, one-third of the respondents were randomly assigned to the direct questioning (DQ) version of the sensitive questions (Figure B.1), and two-thirds to the crosswise-model variant (CM). The unbalanced assignment partly counterbalances the lower statistical efficiency of the crosswise-model RRT. The sensitive questions were preceded by a screen announcing some sensitive questions, stating the importance of honest answers for the success of the study and providing some privacy assurance.

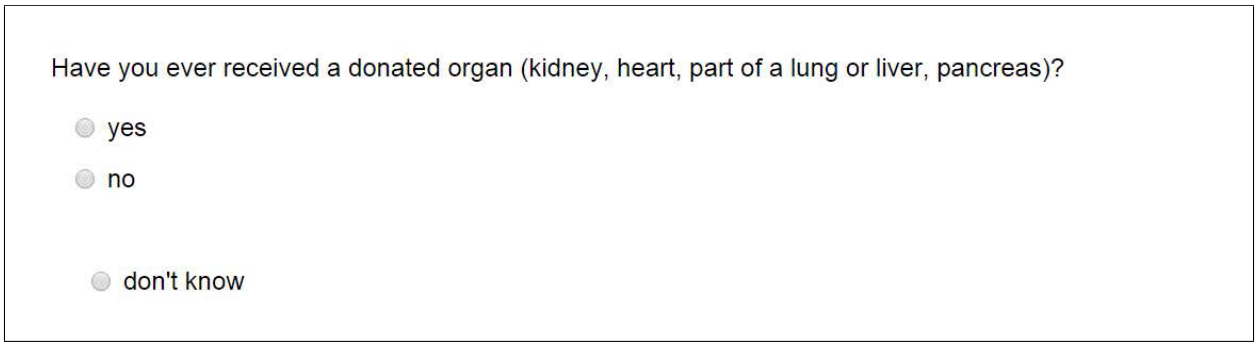

Figure B.1: Screen shot of the direct questioning implementation (translated from German)

[1] Because we used a fully-crossed experimental design, these treatments, which are not discussed here, have no impact on the sensitive question technique validation.

[2] We additionally performed most analyses excluding the 47 respondents who had assessed their language skills as only "medium" and not as "good" or "very good". The results are basically identical. See the online supplement for the corresponding analyses.

The crosswise-model RRT implemented was an unrelated question version as previously used in Jann, Jerke, and Krumpal (2012) and in most other studies using the crosswise model. Respondents were asked two questions at the same time: A sensitive question and an unrelated non-sensitive question (see Figure B.2). Respondents then had to indicate whether their answers to the two questions were identical (both "No" or both "Yes") or different (one "Yes", the other "No"). The CM procedure was carefully introduced to the respondents. On the first screen, we outlined the procedure and briefly explained how the technique protects individual answers. In addition, respondents were referred for further information about the RRT to a Wikipedia article which they could directly access by clicking on a button, with 18% of respondents making use of this possibility. On the second screen, respondents were shown a practice question on whether they had completed the "Abitur". Then, the five sensitive items followed.

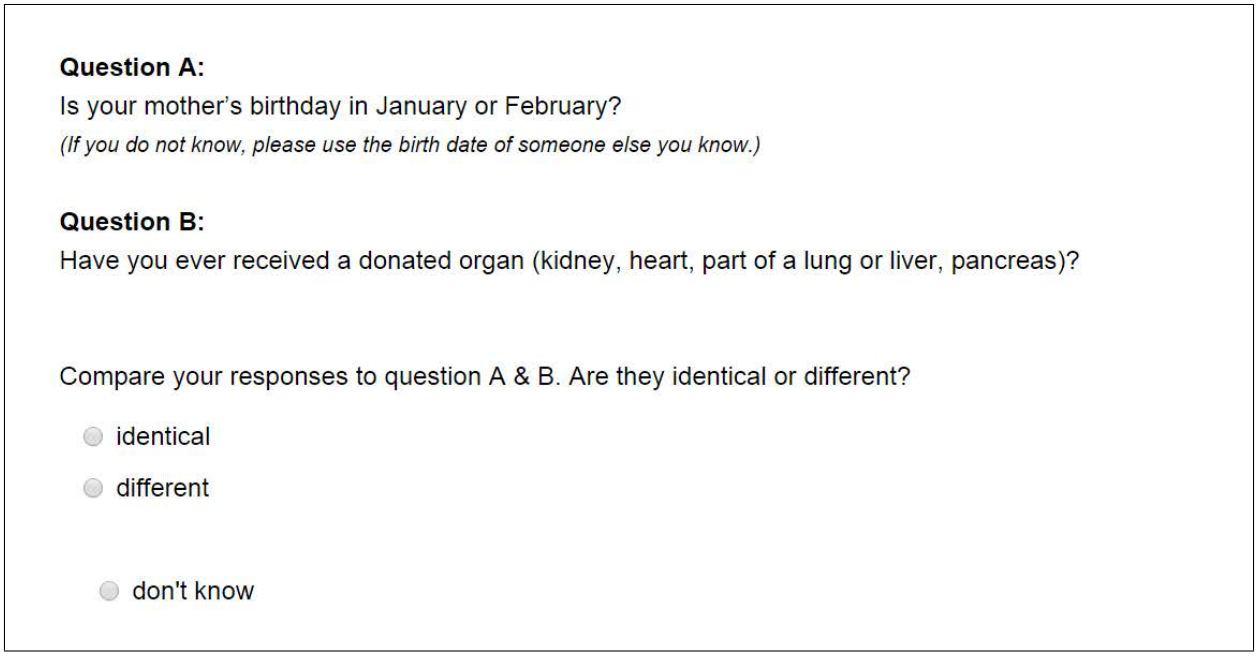

Figure B.2: Screen shot of the CM implementation (translated from German). Unrelated questions (Question A) were randomized across items and every question was only used once for each respondent.

Due to the mixing with the non-sensitive question, a respondent's answer to the sensitive question remains completely private. Nevertheless, at the aggregate level prevalence estimates for the sensitive question are possible because the probability distribution of the unrelated non-sensitive question is known. The unrelated questions used were about the birthdates of respondents' parents and of an arbitrarily chosen acquaintance such as "Is your mother's birthday in January or February?". Unrelated questions were randomly paired with the sensitive items for each respondent. Note that half the respondents received unrelated questions with a probability of a "yes" answer

of .15 to .20, the other half received inverted questions with a "yes" answer probability of .80 to .85 (see Table C.3 for a list of the unrelated questions used). Further, in both the DQ and the CM condition half the respondents were shown a "don't know" response option, whereas the other half were not.

The zero-prevalence items

As zero-prevalence items to test for systematic false positives served a question on having "ever received a donated organ" and on having "ever suffered from Chagas disease (Trypanosomiasis)". We deliberately chose zero-prevalence items that suited the survey topic and had near-zero prevalence in the surveyed population without being completely impossible so that they appeared meaningful to respondents. We did not find any statistics on living organ recipients in Germany. However, using the average number of transplanted organs in Germany from the last ten years ($4,400/year$) to extrapolate over the last 30 years and making the unrealistic but most conservative assumption that all patients who received an organ since 1985 are still alive and that each received only one organ, we can estimate an upper bound of organ recipients presently alive of 132,000, which corresponds to 0.16% of the population.

For the second item, Chagas disease, some epidemiological findings were available. Chagas disease is a parasitic disease spread mostly by insects and potentially leading to heart and digestive disorders that is endemic in most countries in South and Middle America. In Western Europe, however, the disease is nearly non-existent, the exception being Latin American migrants for whom studies found prevalence rates of slightly above 10% for samples from Florence and Geneva. Strasen et al. (2014) estimate an incidence rate for Germany of between 0.0001% and 0.0004%.

Data analysis

To correct for the systematic error that is introduced by the randomization procedure of the crosswise model, the response variable must be transformed. Let Y be the observed response variable with $Y = 1$ if the response is "identical" and $Y = 0$ for "different". S is the actual answer to the sensitive item with $S = 1$ if the answer to the sensitive item is "yes", and $S = 0$ for "no". $p^{yes,u}$ is the known probability of a "yes" answer to the unrelated question. The probability of the response "identical" then is

$$\Pr(Y = 1) = \Pr(S = 1) \cdot p^{yes,u} + (1 - \Pr(S = 1)) \cdot (1 - p^{yes,u})$$

Solving for $\Pr(S = 1)$ results in the transformed response variable $\tilde{Y}$ for the CM:

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

For the direct questioning data, we set $p^{yes,u}$ to 1 so that $\tilde{Y}$ equals the untransformed response variable with $Y = S = 1$ if the answer is "yes" and $Y = S = 0$ if the answer is "no". For the prevalence estimates, we used least-squares regressions on this transformed response variable with robust standard errors (i.e. Fox and Tracy 1986). Data analysis was carried out using the Stata program rrreg (Jann 2008) which readily accommodates the outlined procedure. In addition, we performed all analyses using a logistic regression as well as a non-linear least-squares estimation. The results are essentially identical (see the online supplement for the corresponding analyses and Höglinger, Jann, and Diekmann 2016 for a more thorough discussion of RRT estimation strategies). Figures and tables of the estimated parameters were generated using the Stata programs coefplot (Jann 2014) and esttab (Jann 2007).

C. Additional results

Sensitivity of the items

To assess the sensitivity of the five surveyed items, towards the end of the survey we asked participants to rate how touchy answering them might be. Most items were not assessed as particularly sensitive by the majority of respondents (see Table C.1). The question on blood donation was assessed as "quite touchy" or "very touchy" by only 2% of respondents, the question on organ donation willingness by 23%, and the one on excessive drinking by 43%, apparently being the most sensitive item. The zero-prevalence item on whether one had received a donated organ was assessed as sensitive by 11%, the one on having suffered from Chagas disease by 15%. The five items covered quite a range of sensitivity, but in general appeared not too sensitive to most respondents.

Individual-level validation

As a complementary individual-level validation of the sensitive question techniques, we used a barely sensitive question on whether respondents had (not) completed the "Abitur", the general university entrance qualification. The question was presented as a practice question in the CM condition and appeared as a normal question in the DQ condition. Answers were validated using previously collected information on respondents' basic characteristics when they registered for the online panel. Some limitations apply to this validation. First, the question was presented as a practice question in the CM but not in the DQ condition. It is therefore possible that respondents

Table C.1: Sensitivity assessment of surveyed items

Sensitive item	Respondents assessing an item as "quite touchy" or "very touchy"
Never donated blood	2%
Unwilling to donate organs after death	23%
Excessive drinking last two weeks	43%
Received a donated organ	11%
Suffered from Chagas disease	15%

Notes: Question wording: "Please indicate for the following questions, how touchy answering them might be for some respondents". Answer categories were "not touchy at all", "relatively not touchy", "partly", "quite touchy", and "very touchy". *N* from 1,630 to 1,634

exercised relatively less care in answering it in the CM compared to DQ. To minimize this as far as possible, we asked respondents in the CM condition to "nevertheless, carefully follow the procedure" and to "answer the question truthfully", regardless of the fact that it is not sensitive and for practice. Second, the format differed between the question posed in our survey and the elicitation in the panel's registration form. In the survey, the question read "Have you completed the 'Abitur?"' with the response options "yes" and "no". In the registration form, respondents had to select their educational achievement from among several categories.[3] Third, respondents had registered for the panel up to five years prior to our survey and so it is possible that a few had completed the "Abitur" in the meantime and had not updated the corresponding panel information. However, this would only decrease the false-positive rate. Moreover, the latter two sources of error are constant in both the DQ and the CM condition, hence by comparing the validation results between DQ and CM they are controlled for.

Note that as for the items of the comparative validation the "Abitur" item was reverse-coded, such that the potentially socially undesirable response is the "yes" response, i.e. which corresponds to admitting not having completed the "Abitur". Results of the aggregate-level validation (upper panel of Figure C.3, also see Table C.2) show that the prevalence estimates of respondents not having completed the "Abitur" are nearly identical for DQ and the CM. Both are a negligible two percentage points above the corresponding validation values denoted by the diamond symbol (difference not significant). According to this, one would conclude that both techniques produce valid estimates equally well. This result does not seem surprising given that the question on whether one has completed the "Abitur" is neither barely sensitive nor ambiguous. Yet looking at results

[3]Because there is some disagreement in general understanding on whether one of the offered categories, the subject-specific university entrance qualification ("Fachhochschulreife"), is considered as "Abitur" or not, we excluded the 14% of respondents who selected it, restricting the validation to respondents who unequivocally indicated having completed the "Abitur" or not.

of the individual-level validation (middle and lower panel) tells a very different story. Note that the sensitive outcome is "having not completed the Abitur". Hence, the false negative rate is the share of respondents misclassified as having completed the "Abitur" even though they have not. It amounts to 9% in DQ and up to 29% for the CM. The false positive rate is the percentage of respondents incorrectly classified as not having completed the "Abitur" even though they have. It is not significantly different from zero in the DQ condition but a considerable 7% in the CM. Hence, the CM shows more missclassification than DQ in both directions. Note that the CM's high false negative and high false positive rates level each other out, resulting in an accurate aggregate prevalence estimate.

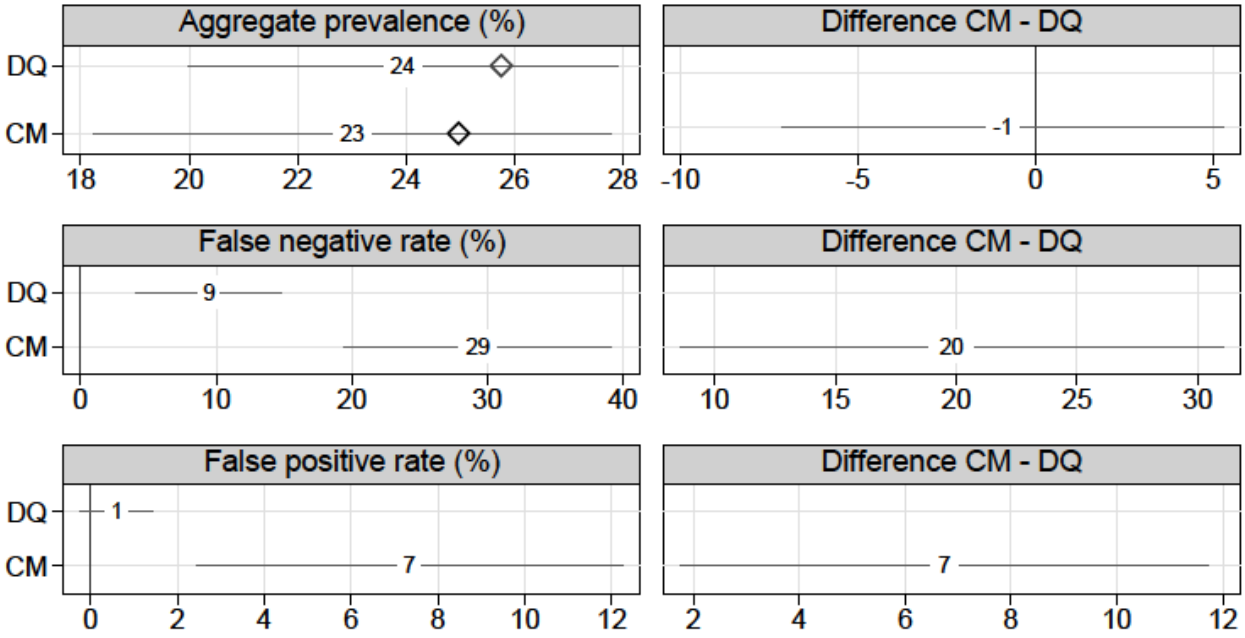

Figure C.3: Aggregate-level validation (upper panel) and individual-level validation (middle and lower panels). Diamond symbols denote the aggregate validation values of "no Abitur" (lines indicate a 95% confidence interval). $N = 458$ for DQ and $N = 953$ for CM

In sum, these results corroborate the findings from the zero-prevalence comparative validation. As mentioned, our individual-level validation had some limitations, mainly that we cannot rule out that the higher misclassification in the CM is caused to some extent by the fact the question was presented as a practice question in the CM condition. But what is most remarkable is not so much the finding that there was again misclassification in the CM, but that the substantial misclassification was not revealed in the aggregate-level validation. This demonstrates the serious weakness of such a validation strategy.

Table C.2: Aggregate and individual-level validation as displayed in Figure C.3 (standard errors in parentheses)

	Aggregate prevalence	False negative rate	False positive rate
Direct questioning (DQ)	23.94	9.48	0.60
	(2.02)	(2.73)	(0.43)
Crosswise model (CM)	23.01	29.29	7.34
	(2.43)	(5.03)	(2.51)
Difference CM - DQ	-0.93	19.81	6.74
	(3.16)	(5.72)	(2.54)

Notes: $N = 1,361$. Aggregated validation values are 25.76 for DQ, and 24.97 for CM

Exploring the causes and correlates of false positives in the CM

Having shown that false positives occurred in the CM with a non-ignorable frequency, we now look at some potential causes and mechanisms underlying this type of misclassification. We can think of two main causes: Careless answering and a bias in the unrelated question outcome that served as a randomizing device. Socially desirable responding can be excluded because the less incriminating answer to the zero-prevalence items is "no", i.e. denying having received a donated organ or having suffered from Chagas disease. Hence, it is hard to imagine why respondents would deliberately give a false "yes" answer to these questions.

The first, careless answering, might be the result of respondents not complying with the CM procedure to evade the effort involved or because they simply were unable to cope with the special procedure's complexity. Due to the privacy-protecting nature of the CM, false answers can never be revealed and so respondents might be more inclined to careless answering in the CM than in the direct questioning mode where answers are potentially verifiable (for this argument, also see Wolter and Preisendörfer 2013). Assuming that careless answering results in random responses, i.e. ticking the response options "different" and "identical" with equal probability[4], the share of respondents randomly answering needed to produce the bias found in our data would be twice the actual false positive rate: 15% for the "received organ" item and 10% for "Chagas disease" (see the left panel of Figure C.4).[5] Randomly answering always produces more false positives than

[4]Because the order of the response options was randomized across respondents and also because half the respondents received inverted unrelated questions, hence the correct response ("identical" or "different") was exactly the inverse, this assumption is quite plausible.

[5] The function for the false positive bias is derived from the transformed response variable $\tilde{Y}$ for the CM:

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

We introduce the probability r of answering randomly, hence of giving the response "identical" with a probability of

negatives for a prevalence that in reality is below 0.5, which is typical for sensitive items.[6] Hence, in principle it could explain the overestimation bias found in our study as well as the consistently higher estimates from previous validations.

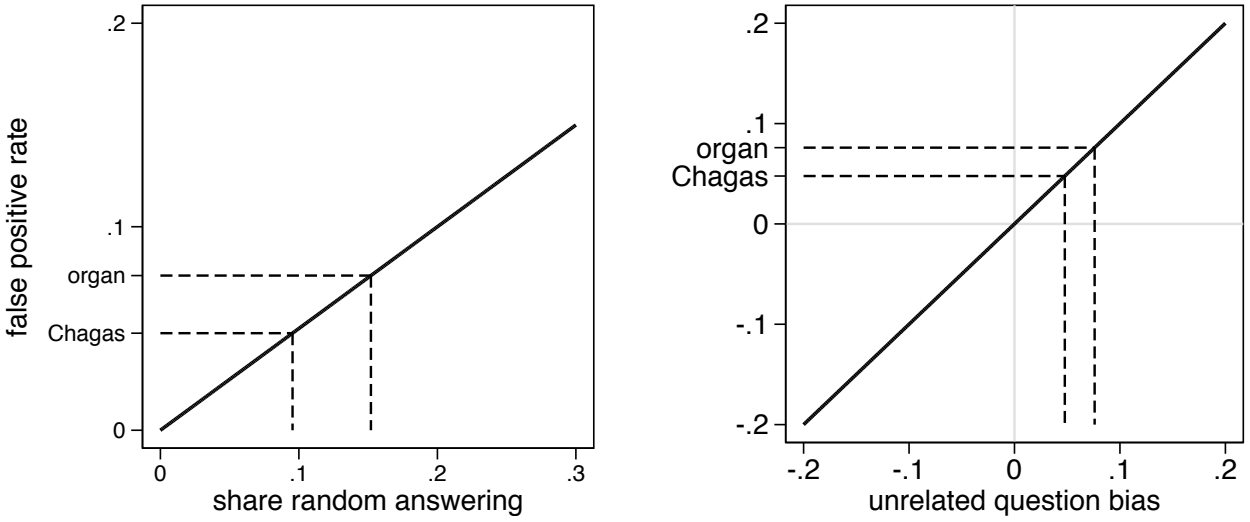

Figure C.4: Effect of random answering and unrelated question bias on the false positive rate for zero-prevalence items. (Dashed lines indicate false positive rates found and the corresponding extent of error necessary to generate them.)

Notes: With an expected "yes" probability for the unrelated questions of 0.18 as in the CM implemented. If the "yes" probability is inverted to 0.82 random answering has the same effect, but the effect of the unrelated question bias goes in the opposite direction.

The second potential cause, a bias in the unrelated question outcome, occurs if the unrelated questions do not produce the theoretically expected "yes" answer prevalence. We used unrelated questions about the birth dates of respondents' mother and father, and of arbitrarily chosen acquaintances. A bias in the "yes" probability could occur if there is actually a different prevalence of the underlying attribute in the study sample, which is quite unlikely for birthdate questions, or if respondents do not know the status of the attribute, i.e. the date of their parents' birth. In addition, for the question on an acquaintance's birthday which in one version read "Think of an acquaintance of yours whose birthday you know: Is this person's birthday in January or February?" respondents might be more inclined to choose an acquaintance whose actual birthdate falls within the specified time frame (January or February) or whose birthday falls about the time the survey was carried

.5, and a bias b:

$$\Pr(S = 1) + b = \frac{.5r + \Pr(Y = 1)(1 - r) + p^{yes,u} - 1}{(2p^{yes,u} - 1)}$$

After setting the "yes" prevalence $\Pr(S = 1)$ to zero and $\Pr(Y = 1)$ to $1 - p^{yes,u}$ as assumed, we arrive after some transformations at $b = r/2$.

[6]For estimates with a true prevalence above 0.5 the inverse holds: random answering leads to more false negatives and an underestimation in the aggregate. Complete random answering would lead in both cases to an estimate of 0.5.

out. To minimize such effects (and test them, see below), we randomized the unrelated questions across items and also used an inverted form for every unrelated question (instead of "in January or February", "in March to December, including December").

To generate the false positive rates found in our data, the "yes" answer bias must be of the same size, namely 8 and 5 percentage points (see the right panel of Figure C.4). We subjected the unrelated questions to a test by asking respondents of the DQ condition to explicitly answer the unrelated questions used in the CM.[7] A comparison of the so elicited "yes" prevalence with the theoretically expected prevalence showed a good match in general (see Table C.3). With the exception of three out of twelve questions, the differences were in the range of -5 to +3 percentage points and not significant. In part, very sizeable differences were found for the questions on "acquaintance's birthday in January or February" (36% instead of 16%, +20 percentage points bias), "acquaintance's birthday from the 1st to the 6th of the month" (31% instead of 20%, +11 percentage points), and for "father's birthday in March to December, including December" (77% instead of 84%, -7 percentage points). Interestingly, these prevalence estimates were all biased towards 50%, suggesting that choosing an answer at random might be the cause. Excluding responses based on these three potentially problematic unrelated questions indeed reduced false positive rates from 8% to 6% (received donated organ) and from 5% to 1% (Chagas disease, see the online supplement for the corresponding analysis). Apparently, some of the unrelated questions used might have been problematic. Most likely that is because they leave too much wiggle-space to respondents (the question on an acquaintance's birthday), or some respondents simply do not know the answer (the question on the father's birthday). A less unequivocal non-sensitive question or another randomizing device might therefore be preferable.

Note that, in contrast to random answering, a bias in the unrelated question outcome can lead to more false positives as well as more false negatives depending on the direction of the "yes" answer bias. This would not quite fit the pattern whereby the CM consistently produced more false positives. Still, the problematic questions identified with our test all showed a bias towards 50%, which would result in relatively more false positives. Therefore, the unrelated questions are likely responsible for some false positives, although they do not explain the whole bias.

Irrespective of the actual cause of the false positives (it might well be a mix of various mechanisms), we expected to find systematic patterns regarding implementation details of the CM as well as respondents' behavior and characteristics. In the following, we first present the effects of experimentally manipulated details of the CM implementation on false positives. Our analytical

[7] The questions were introduced as a "seemingly strange" task without detailing the purpose. To increase the certainly limited comparability, we employed a procedure as similar as possible and also randomized the question order. Of course, because the context of the questions when they were tested was very different to when they were used in the CM, we cannot directly infer that the same bias occurred in the CM. Still, the test provides some insights into the direction and possible size of the potential bias, and indicates potentially problematic questions.

Table C.3: Comparison of the elicited and theoretical "yes"-prevalence to unrelated questions used in the CM (standard errors in parentheses)

	"Yes" prevalence in test	Theoretical "yes" prevalence	Difference
Mother's birthday Jan-Feb	15.30	15.95	-0.65
	(2.20)		(2.20)
Mother's birthday 1st-6th	18.35	19.71	-1.36
	(2.37)		(2.37)
Father's birthday Jan-Feb	17.16	15.95	1.22
	(2.31)		(2.31)
Father's birthday 1st-6th	18.87	19.71	-0.85
	(2.41)		(2.41)
Acquaintance's birthday Jan-Feb	35.82	15.95	19.87*
	(2.93)		(2.93)
Acquaintance's birthday 1st-6th	30.57	19.71	10.85*
	(2.84)		(2.84)
Mother's birthday Mar-Dec	81.01	84.05	-3.05
	(2.45)		(2.45)
Mother's birthday 7th-31st	83.01	80.29	2.72
	(2.34)		(2.34)
Father's birthday Mar-Dec	77.38	84.05	-6.67*
	(2.64)		(2.64)
Father's birthday 7th-31st	75.60	80.29	-4.69
	(2.72)		(2.72)
Acquaintance's birthday Mar-Dec	82.75	84.05	-1.31
	(2.37)		(2.37)
Acquaintance's birthday 7th-31st	76.77	80.29	-3.52
	(2.65)		(2.65)

Notes: N from 250 to 268 per question. $^{*}p < 0.05$

strategy consisted of running bivariate regressions on the pooled response variables of the two zero-prevalence items, where answering "yes" is equivalent to giving a false positive. The results show that none of the experimental manipulations had a significant effect on false positives (Table C.4). The largest, albeit not significant effect (-4 percentage points, $p = 0.108$) was found for the introduction of a "don't know" response option.[8] All other manipulations such as reversing the

[8]Because only 0.7% (organ recipient) and 0.5% (Chagas) of the respondents provided with a "don't know" response option actually ticked it, the effect of the "don't know" option on false positives was not caused by respondents actually

Table C.4: Effects of CM implementation details on false positive rate (bivariate regression coefficients, standard errors in parentheses)

	Percentage points change
With "don't know" response option	-4.48
	(2.79)
Response order different - identical (vs. inverse)	-1.18
	(2.79)
Unrelated question on father (vs. mother)	-2.82
	(2.87)
Unrelated question on acquaintance (vs. mother)	2.69
	(2.91)
Unrelated question on birthday (vs. birth month)	2.04
	(2.73)
Yes-probability unrelated question .82 (vs. .18)	-2.10
	(2.79)
Item position (linear)	0.09
	(0.96)
Item position 1st or 2nd (vs. 4th or 5th)	-1.54
	(3.77)

Notes: Bivariate regressions on pooled responses to zero-prevalence items. Standard errors corrected for clustering in respondents. $N = 2,243$. $^{*}p < 0.05$

order of the response options from identical–different to different–identical, the type of the unrelated question (birthday of mother, father, or acquaintance; birthday vs. birth month), or inverting the "yes" probability of the unrelated question from on average $p = .18$ to $p = .82$ clearly had no effect. Moreover, no effects were found for the placement of the sensitive item, i.e. whether they were displayed as the first, second, third, fourth, or fifth item.

In the final step, we explored bivariate associations between giving a false positive and respondents' behavior and personal characteristics. Again, the results are far from conclusive (Table C.5). Being among the 10% of respondents who passed the CM introduction page with the explanations on the special technique the fastest was positively related to giving a false positive (+9 percentage points, albeit not significant at a conventional level, $p = 0.063$). This suggests that speeding respondents did not carefully read the instructions and thus did not fully understand the CM procedure, and consequently gave more false positive responses. But, somehow in contrast to this finding, being among the 10% fastest respondents in answering the five sensitive items was

making use of this option. It was the response behavior of those who ticked the "different" or "identical" response that was altered by simply having this option offered.

Table C.5: Bivariate associations between respondents' behavior and personal characteristics and false positive rate (bivariate regression coefficients)

	Percentage points change
Among the fastest 10% on CM introduction screen	9.05
	(4.87)
Among the fastest 10% answering sensitive items (without intro)	-4.33
	(4.46)
Clicked button referring to the RRT Wikipedia link	6.05
	(3.90)
Social desirability (Crown-Marlowe scale)	1.62*
	(0.80)
Completed the university entrance qualification	-5.17
	(3.53)
Age	-0.03
	(0.10)
Female	-1.73
	(2.95)

Notes: Bivariate regression on pooled zero-prevalence items. Standard errors corrected for clustering in respondents. N from 2,208 to 2,243. $^{*}p < 0.05$

clearly not positively associated with false positives. Clicking on the button provided to access the Wikipedia page with further RRT information on the introduction screen also showed no significant association. Scoring high on the Crowne-Marlowe social desirability scale (Crowne and Marlowe 1960) was positively related to giving a false positive (+1.6, $p = 0.042$, $scaleSD = 1.7$), meaning that respondents more prone to socially desirable responding were also more likely to give a false positive. We have no explanation for this finding because, if any social desirability bias existed, it would instead work against falsely admitting having suffered from Chagas disease or having received a donated organ. Finally, having completed the university entrance qualification is not systematically related to false positives, nor are age or gender.

Note that the statistical power of the previous analyses was relatively weak due to the low prevalence of the false positives. In addition, we tested several potential causes and covariates without having a clear theory about how they are related to false positives in the CM. Hence, the risk of both alpha and beta errors increased considerably and the findings presented in this section must be interpreted as exploratory. However, in light of the novelty of the finding that the CM produced false positives and a unique possibility to analyze the potential causes these results are, in our view, nevertheless valuable for informing future studies dealing with improving the crosswise model or related techniques. In sum, the analysis of the causes and correlates of false positives

did not reveal any pattern that would clearly point to a particular explanation. We could, however, identify some candidate causes of false positives whose effect should be investigated more systematically in future studies: The unrelated questions used and their respective bias, not offering a "don't know" response option (albeit the reason is unclear), and respondents speeding over the CM instructions. Still, each of these factors accounts for only a share of the false positives that occurred and, very likely, false positives might have been caused by a mix of different mechanisms.

D. **Table underlying the figure in the main text**

Table D.6: Comparative validation of sensitive question techniques as displayed in Figure 1 (standard errors in parentheses)

	Never donated blood	Unwilling to donate organs	Excessive drinking	Received a donated organ	Suffered from Chagas disease
Levels					
Direct questioning (DQ)	48.82	22.01	20.58	0.00	0.37
	(2.14)	(1.82)	(1.73)	(.)	(0.26)
Crosswise model (CM)	51.58	27.30	32.71	7.60	4.77
	(2.33)	(2.23)	(2.28)	(1.95)	(1.91)
Difference					
CM – DQ	2.76	5.29	12.13	7.60	4.40
	(3.16)	(2.88)	(2.86)	(1.95)	(1.92)
N	1669	1641	1672	1669	1669

References

Crowne, Douglas P., and David Marlowe. 1960. "A New Scale of Social Desirability Independent of Psychopathology". *Journal of Consulting Psychology* 24:349–354.

Fox, James Alan, and Paul E. Tracy. 1986. *Randomized response: A method for sensitive surveys*. Newbury Park, CA: Sage.

Hoffmann, Adrian, Birk Diedenhofen, Bruno Verschuere, and Jochen Musch. 2015. "A Strong Validation of the Crosswise Model Using Experimentally-Induced Cheating Behavior". *Experimental Psychology* 62:403–414.

Höglinger, Marc, and Andreas Diekmann. 2016. *Replication Data for: Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT*. Harvard Dataverse. doi:10.7910/DVN/SJ2RP1.

Höglinger, Marc, and Ben Jann. 2016. *More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model*. University of Bern Social Sciences Working Paper No. 18. University of Bern. https://ideas.repec.org/p/bss/wpaper/18.html.

Höglinger, Marc, Ben Jann, and Andreas Diekmann. 2016. "Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model". *Survey Research Methods* 10 (3): 171–87. doi:10.18148/srm/2016.v10i3.6703.

Jann, Ben. 2007. "Making regression tables simplified". *Stata Journal* 7:227–44.

— . 2014. "Plotting regression coefficients and other estimates". *Stata Journal* 14:708–37.

— . 2008. *rrreg: Stata module to estimate linear probability model for randomized response data*. S456962. Boston College Department of Economics.

Jann, Ben, Julia Jerke, and Ivar Krumpal. 2012. "Asking Sensitive Questions Using the Crosswise Model. An Experimental Survey Measuring Plagiarism". *Public Opinion Quarterly* 76:32–49.

John, Leslie K., George Loewenstein, Alessandro Acquisti, and Joachim Vosgerau. 2016. *When and Why Randomized Response Techniques (Fail to) Elicit the Truth*. Harvard Business School Working Paper No. 16-125. Harvard Business School. http://www.hbs.edu/faculty/Pages/item.aspx?num=51059.

Kirchner, Antje. 2015. "Validating Sensitive Questions: A Comparison of Survey and Register Data". *Journal of Official Statistics* 31:31–59.

Moshagen, Morten, Benjamin E. Hilbig, Edgar Erdfelder, and Annie Moritz. 2014. “An Experimental Validation Method for Questioning Techniques That Assess Sensitive Issues”. *Experimental Psychology* 61:48–54.

Strasen, Jörn, Tatjana Williams, Georg Ertl, Thomas Zoller, August Stich, and Oliver Ritter. 2014. “Epidemiology of Chagas Disease in Europe: Many Calculations, Little Knowledge”. *Clinical Research in Cardiology* 103:1–10.

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 & Research* 28:505–537.

Wolter, Felix, and Peter Preisendörfer. 2013. “Asking Sensitive Questions: An Evaluation of the Randomized Response Technique vs. Direct Questioning Using Individual Validation Data”. *Sociological Methods & Research* 42:321–353.