PA

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

Marc Höglinger[1] and Andreas Diekmann[2]

[1] *Winterthur Institute of Health Economics, Zurich University of Applied Sciences, Gertrudstrasse 15, 8401 Winterthur, Switzerland. Email: marc.hoeglinger@gmail.com*
[2] *ETH Zürich, Department of Humanities, Social and Political Sciences, Clausiusstrasse 50, 8092 Zurich, Switzerland. Email: diekmann@soz.gess.ethz.ch*

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 nonignorable 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.

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).

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. An online appendix for this article is available on the journal's website. For replication data see Höglinger and Diekmann (2016).

Political Analysis (2017)
vol. 25:131–137
DOI: 10.1017/pan.2016.5

Published
9 February 2017

Corresponding author
Marc Höglinger

Edited by
Jonathan Katz

The recently proposed crosswise-model (CM) 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 CM, respondents are asked two questions simultaneously, a sensitive one (e.g., "Are you an active member of the Egyptian Muslim Brotherhood?") and a nonsensitive 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 nonsensitive 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 CM 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 CM 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 (DQ) (Hoffmann and Musch 2016; 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 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 DQ, 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 noncompliance 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 DQ 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 2016) are preferable because they do not need the DQ estimate as a benchmark. However, they too

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).

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

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 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 nonzero 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 CM RRT produced positive, i.e. wrong, prevalence estimates of the zero-prevalence items, and hence generated false positives to a nonignorable extent.

2 Data and design

Our analysis sample consisted of 1,685 members of a nonrepresentative 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

2 See online Appendix B for data and design details, and Höglinger and Diekmann (2016) for replication data.

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 DQ version of the sensitive questions, and two-thirds to the CM variant.[3] The CM RRT implemented was an unrelated question version as used in Jann, Jerke, and Krumpal (2012) and in most other previous studies using the CM. Respondents were asked two questions at the same time: A sensitive question and an unrelated nonsensitive question. Respondents then had to indicate whether their answers to the two questions were identical or different. Due to the mixing with the nonsensitive 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 DQ and the 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 nonignorable 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

3 To counterbalance the lower statistical efficiency of the CM.

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).

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.

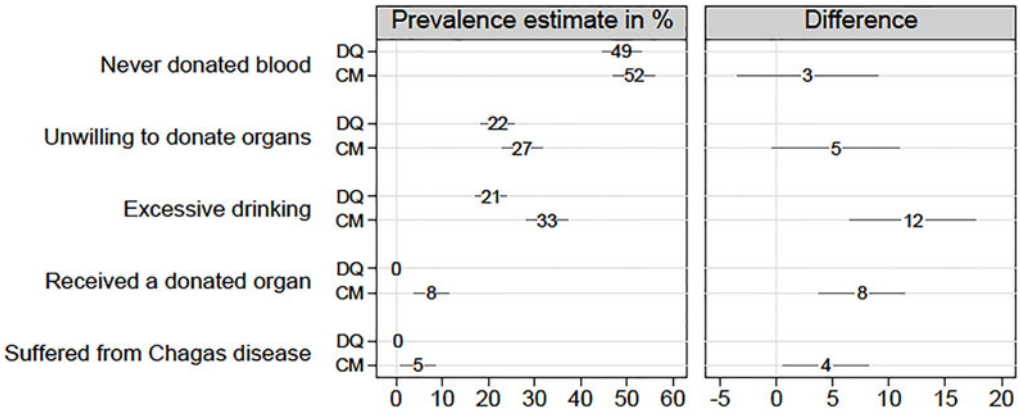

Figure 1. Comparative validation of sensitive question techniques (lines indicate a 95% confidence interval, *N* from 518 to 549 for DQ, and from 1,120 to 1,123 for CM).

tested. We could, however, identify some candidate causes of false positives whose effect should be investigated 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 CM RRT, we found that the CM produced false positives to a nonignorable extent. Our evidence is based on a comparative validation with zero-prevalence items and an additional individual-level validation using a nonsensitive question. Previous validation studies appraised the CM for its easy applicability and seemingly more valid results. However, none of them considered false positives. Our results strongly suggest that in reality the CM 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 CM 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 CM 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.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2016.5.

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* 60:1077–1092.

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* 24:132–156.

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. 2016. Assessing the validity of two indirect questioning techniques: A stochastic lie detector versus the crosswise model. *Behavior Research Methods* 48:1032–1046.

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):171–187.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. Measuring voter turnout by using the randomized response technique: Evidence calling into question the methods 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. 2016. An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science* 60:783–802.

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 versus 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.