Doctoral Thesis

Revealing the Truth? Validating the Randomized Response Technique for Surveying Sensitive Topics

Author(s):
Höglinger, Marc

Publication Date:
2016

Permanent Link:
https://doi.org/10.3929/ethz-a-010636127

Rights / License:
In Copyright - Non-Commercial Use Permitted
Revealing the Truth?
Validating the Randomized Response Technique for Surveying Sensitive Topics

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

MARC HÖGLINGER
lic. phil., University of Zurich
born on 16 May 1979
citizen of Grüssch, GR

accepted on the recommendation of

Prof. Dr. Andreas Diekmann (examiner)
Prof. Dr. Thomas Hinz (co-examiner)
Prof. Dr. Peter Preisendörfer (co-examiner)

2016
Preface and acknowledgments

My journey into sensitive question research has led me into the fields of survey methodology, measurement and its errors, psychology and lying, statistics and randomness, and even into the casino. I searched the Internet for all possible randomization devices and sources such as random wheels, virtual dice, or atmospheric noise. I analyzed birthday patterns across the year and days of months, distributions of phone number digits and developed online cheating games in which people can cheat for money without ever being detected. In sum, I did a lot of fascinating things and I met many interesting and nice people. Most of the scientifically meaningful outcomes of this journey are presented in this dissertation.

I thank Andreas Dickmann, my supervisor, for giving me the opportunity to undertake this journey and to conduct research in a very inspiring environment, to stimulate my scientific curiosity, and to give me the room necessary for developing my own ideas. I always knew I had his full support for my scientific endeavor. Also, I cordially thank Thomas Hinz and Peter Preisendörfer for supporting this dissertation as co-supervisors – I am very pleased they agreed to take on this important engagement.

I learned a lot from Ben Jann, my predecessor at the chair, especially regarding statistical data analysis and programming and I am glad he joined me for some of my work. Without his support, my analyses and documentation would not have been as sound and complete as they are (hopefully) now. My colleagues from the chair, Joël Berger, Heidi Bruderer, Jennifer Gewinner, Julia Jerke, Matthias Naef, Manuela Vieth, and Stefan Wehrli helped me with lots of minor and major challenges I encountered during my research. I particularly thank them for patiently testing my countless questionnaire drafts. Ivar Krumpal and Felix Wolter, two other RRT enthusiasts, accompanied my journey into sensitive questions right from the start and I had many inspiring discussions, conference visits, and enjoyable journeys with them. Kurt Ackermann, Domenico Angelone, Debra Hevens- stone, Wojtek Przepiorka, Heiko Rauhut, Tobias Wolbring, and Chris Young pro-
vided valuable input to some of my research. Claudia Jenny improved the English of many of my texts and questionnaires. Jonas Blatter accomplished the professional typesetting of the final manuscript. I thank them all. In addition, I want to mention Elisabeth Coutts who pushed sensitive question research at the chair at a time when I still did not even know the meaning of the acronym RRT and whom I sadly never met while she was still among us.

The German Research Foundation, the Chair of Sociology of ETH Zürich, and the Institute of Sociology of the University of Bern provided generous funding that made my work possible. Also, I am very grateful to the thousands of participants who spared their time to take part in one of my surveys.

And, finally, a big thank you for making my days brighter goes to Gregory, Miranda, and Jessica.

Zurich, March 2016

Marc Höglinger
Summary

Validly measuring sensitive issues such as norm-violating behavior or stigmatizing traits with survey self-reports poses a big challenge. Various studies have shown that the share of respondents who misreport can be considerable. Despite this serious flaw, research on social norms and deviance, epidemiology, political science, and many other areas relies heavily on self-report data. This dissertation deals with validating special sensitive question techniques, more precisely, variants of the Randomized Response Technique (RRT, Warner 1965), that are intended to overcome this problem. The RRT should obtain truthful answers to sensitive questions by granting respondents full response privacy through some randomization procedure. Full response privacy means there is no possibility to infer from a single respondent’s response his or her actual answer to a sensitive question. In turn, respondents are supposed to answer honestly. However, methodological studies are so far inconclusive about whether the RRT fulfills its theoretical promise and consistently leads to more valid self-reports.

In my dissertation, I present different validation studies assessing RRT implementations that were all carefully designed and tailored to the online mode. The results regarding the evaluated RRT implementations are, in sum, devastating. None of them succeeded in eliciting more valid data than standard direct questioning. Quite to the contrary, many RRT implementations revealed significantly more misclassification than direct questioning. In particular, an application of the allegedly promising recent crosswise-model RRT variant (Yu, Tian, and Tang 2008) was found to produce sizeable shares of false positives, i.e. respondents misclassified as possessing a sensitive trait even though they actually did not—a misclassification type that had so far largely been overlooked. Based on these results, the RRT in its various variants cannot be recommended without first further clarifying which variant actually works in which implementation and in which context.

The dissertation’s second main contribution lies in clarifying what different validation strategies reveal about a particular sensitive questioning technique’s
validity. I show that validation studies which do not consider the possibility of false positives can be seriously misleading. I found that a widely used implementation of the recent crosswise-model RRT produced considerable false positives – a defect that a series of previous studies not considering false positives did not reveal. On the contrary, these studies interpreted the resulting higher prevalence estimates of sensitive behavior or traits – with more or less caution – as more valid estimates under the so-called more-is-better assumption. This assumption states that socially desirable responding is the only source of misclassification, hence, respondents only falsely deny sensitive traits (false negatives) but never falsely admit them (false positives). Consequently, the more respondents a particular technique classifies as having the sensitive trait, the more valid the data. However, as the occurrence of false positives in the crosswise-model implementation showed, the more-is-better assumption might not be warranted and the blind reliance on it is a serious weakness of most previous sensitive question research.

The third contribution is the development of two novel designs that allow the validation of special sensitive question techniques (be they the RRT or others) in a meaningful way and that overcomes the mentioned weakness of most earlier validations. The first design is an experimental individual-level validation where self-reports about cheating in an incentivized dice game can be validated. The second is a comparative validation that is able to detect systematic false positives thanks to the introduction of one or more zero-prevalence items. Both designs are easy to apply and replicate because they do not need a preexistent external individual-level validation criterion, which is often unavailable. Therefore, the two validation designs represent useful tools for future systematic sensitive question research.

The first study ("A comparative RRT validation", chapter 2 deals with developing and evaluating RRT variants that are suitable for online use. The online mode seemed a promising field for the use of the RRT, and there were only a few validation studies in this area. Chapter 3 ("The Benford RRT and an exploration of privacy") takes a detailed look at the Benford RRT, an implementation that seemingly worked well, and at the notion of privacy – the core principle of why the RRT should make respondents answer more honestly. Then, we realized that the evaluation methods hitherto in use, including our own, had severe weaknesses and that, although these weaknesses are repeatedly mentioned and discussed in the literature, they had almost never been properly addressed. Therefore, I designed a second study using a cheating experiment that enabled validation of respondents' self-reports about whether they had cheated on an individual level ("More is not always better: an individual-level validation", chapter 4). The results were very informative, not only regarding the validity of particular RRT variants (a crosswise-model implementation produced seriously biased
data), but especially because they showed that blindly relying on the more-is-better assumption, as done so far by most validation studies, is no longer tenable. The third study ("An enhanced comparative validation design for sensitive question research", chapter 5) presents a comparative validation able to detect false positives or, in other words, to test the more-is-better assumption. In contrast to the individual-level validation from chapter 4, it is, however, very straightforward to implement, more flexible, and closer to a substantive survey application. In this sense, it is an easy-to-apply validation strategy that is replicable and might be very useful for future evaluations of RRT implementations and even of other special sensitive question techniques.
Kurzfassung


Die erste Studie ("A comparative RRT validation", Chapter 2) behandelt die Entwicklung und Evaluation von RRT-Implementierungen, welche für den Online-Modus geeignet sind. Online-Befragungen sind ein vielversprechender Anwendungsbereich für die RRT und erst vereinzelte Studien widmeten sich diesem Thema. Chapter 3 ("The Benford RRT and an exploration of privacy") wirft einen genauerer Blick auf eine einzelne RRT-Implementierung, Benford RRT, welche gut zu funktionieren schien. Zudem untersucht sie den Aspekt
des Antwortschutzes ("privacy") näher – das Kernprinzip, wie die RRT Responder
denten dazu bringen soll, wahrheitsgetreu zu antworten. Im Anschluss real-
isierten wir, dass bisherige Validierungs-Strategien, inklusive unsere eigene er-
ste Studie, eine beschränkte Aussagekraft und grosse Schwächen haben, welche
praktisch nie ernsthaft angegangen wurden. Für die Folgestudie entwickelte
ich deshalb ein Schummel-Experiment, welches erlaubt, Selbstangaben zum
Schummeln auf individueller Ebene zu validieren ("More is not always better:
an individual-level validation", Chapter 4). Die Ergebnisse waren sehr auf-
schlussreich bezüglich der Validität einzelner RRT-Implementierungen (eine Im-
plementierung des Crosswise Modells produzierte sehr hohe Missklassifikation,
keine RRT-Implementierung generierte validere Daten als die direkte Befragung),
aber insbesondere zeigte sich unmissverständlich, dass blindes Vertrauen in die
"More is better"-Annahme – wie bei den meisten Validierungs-Studien prakti-
tiziert – unhalbar ist. Die dritte Studie ("An enhanced comparative validation
design", Chapter 5) stellt eine vergleichende Validierung vor, welche es durch
die Identifikation von falsch Positiven erlaubt, die "More is better"-Annahme
zu überprüfen. Im Vergleich zur Validierung auf Individual-Ebene in Chap-
ter 4 ist dieses Design einfacher zu implementieren, flexibler, und näher an einer
realen Anwendung in einer Bevölkerungsbefragung. In diesem Sinne ist es ein
sehr einfach einsetzbares Validierungs-Design, das einfach replizierbar ist und
für zukünftige Evaluationen von RRT-Implementierungen und anderer spezifischer
Fragetechniken zur Erhebung heikler Themen sehr nützlich sein dürfte.
Contents

1 Introduction 1

2 Sensitive Questions in Online Surveys: An Experimental Evaluation of the RRT and the Crosswise Model 7
  2.1 Introduction .................................................. 8
  2.2 Data and Methods ........................................ 13
  2.3 Results ....................................................... 21
  2.4 Discussion and Conclusions ............................. 30
  2.A Appendix .................................................. 33

3 A New Randomizing Device for the RRT Using Benford’s Law: An Application in an Online Survey 35
  3.1 Introduction ................................................ 36
  3.2 The Randomized Response Technique (RRT) ........ 36
  3.3 Benford RRT: A new randomizing device using Benford’s law 40
  3.4 An application in a survey on student cheating ........ 44
  3.5 Conclusions ............................................... 47
  3.A Appendix .................................................. 49

4 More Is Not Always Better: An Experimental Individual-Level Validation of the RRT and the Crosswise Model 53
  4.1 Introduction ................................................ 54
  4.2 Data and Methods ........................................ 58
  4.3 Results ...................................................... 67
  4.4 Conclusions ............................................... 73
  4.A Appendix .................................................. 76
5 False Positives Undermine the Crosswise-Model RRT: An Enhanced Comparative Validation Design for Sensitive Question Research 79
5.1 Introduction 80
5.2 Data and methods 86
5.3 Results 90
5.4 Discussion and conclusion 100
5.5 Appendix 102

6 Summary and Conclusions 105

References 111

Curriculum vitae 121
List of Tables

2.1 Sensitive questions on student misconduct (translated from German) ........................................ 13
2.2 Experimental conditions and number of observations ............................................... 15
2.3 Norms against academic misconduct ................................................................. 22
2.4 Summary of effects of evaluation criteria on prevalence estimates ........................ 30
2.A.1 Prevalence estimates by experimental condition (in percent; standard errors in parentheses) .................. 33
2.A.2 Comparison of experimental conditions on various measures .............................. 34

3.A.1 Comparison of prevalence estimates of cheating between direct questioning (DQ) and Benford RRT .......................................................... 49
3.A.2 Comparison of break-off rate, response time and respondents' evaluation of the sensitive question technique between direct questioning (DQ) and Benford RRT .................................................. 49
3.A.3 Comparison of Benford RRT prevalence estimates of cheating between designs with differing probability \( p \) with which respondents are instructed to answer the sensitive question .................. 51
3.A.4 Comparison of break-off rate, response time and respondents' evaluation of the sensitive question technique with differing probability \( p \) .................................................. 51

4.1 Descriptive statistics of the sample ................................................................. 60
4.2 Sensitive questions ......................................................................................... 61
4.3 Number of observations by dice game variant and sensitive question technique .................................................. 62
4.A.1 Prevalence estimates by sensitive question technique as displayed in figure 4.1 .................................................................................................. 76
4.A.2 Cheating rates in the prediction game and the roll-a-six game as displayed in figure 4.2 .................................................................................. 77
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.A.3</td>
<td>Individual-level validation results in the prediction game and the roll-a-six game as displayed in figure 4.3</td>
<td>77</td>
</tr>
<tr>
<td>5.1</td>
<td>Sensitive questions</td>
<td>87</td>
</tr>
<tr>
<td>5.2</td>
<td>Sensitivity assessment of surveyed items</td>
<td>90</td>
</tr>
<tr>
<td>5.3</td>
<td>Effects of CM implementation details on the false positive rate</td>
<td>98</td>
</tr>
<tr>
<td>5.4</td>
<td>Bivariate associations between respondents’ behavior and personal characteristics and false positive rate</td>
<td>99</td>
</tr>
<tr>
<td>5.A.1</td>
<td>Comparative validation of sensitive question techniques</td>
<td>102</td>
</tr>
<tr>
<td>5.A.2</td>
<td>Aggregate and individual-level validation</td>
<td>102</td>
</tr>
<tr>
<td>5.A.3</td>
<td>Comparison of the elicited and theoretical “yes” prevalence to unrelated questions used in the CM</td>
<td>103</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Screen shot of the forced-response random wheel implementation ("FR Wheel", translated from German) ................. 16
2.2 Screen shot of the forced-response pick-a-number implementation ("FR Number", translated from German) ............ 17
2.3 Screen shot of the unrelated-question Benford implementation ("UQ Benford"), screen 1 (translated from German) ........ 18
2.4 Screen shot of the unrelated-question Benford implementation ("UQ Benford"), screen 2 (translated from German) ........ 18
2.5 Screen shot of the unrelated-question crosswise-model implementation ("CM Question", translated from German) ... 19
2.6 Screen shot of the pick-a-number crosswise-model implementation ("CM Number", translated from German) .......... 19
2.7 Prevalence estimates and difference to DQ by experimental condition .................................................. 24
2.8 Comparison of experimental conditions on various measures .......................................................... 26
3.1 Comparison of first digits of house numbers from the Swiss phone directory with the Benford distribution .......... 42
3.2 Benford RRT in the unrelated-question RRT design .......... 43
3.3 Comparison of prevalence estimates of cheating between direct questioning and Benford RRT ................. 46
3.4.1 Comparison of Benford RRT prevalence estimates of cheating ............................................. 50
4.1 Comparative validation of sensitive question techniques ........ 68
4.2 Aggregate-level validation of sensitive question techniques .................. 70
4.3 Individual-level validation of sensitive question techniques ................................................... 71
5.1 Screen shot of the direct questioning implementation ........ 88
5.2 Screen shot of the CM implementation .................................................. 89
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Comparative validation of sensitive question techniques</td>
<td>91</td>
</tr>
<tr>
<td>5.4</td>
<td>Aggregate-level validation and individual-level validation</td>
<td>94</td>
</tr>
<tr>
<td>5.5</td>
<td>Effect of random answering and unrelated question bias on the false</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>positive rate for zero-prevalence items</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Revealing the truth is the ultimate goal of science. In the empirical sciences, every investigation rests on measurements of the phenomenon of interest and on the validity of those measurements. The social sciences base a lot of, if not most, research on individuals’ self-reports. James Coleman noted this quite some time ago already: "most research techniques which analyze behavioral data take a shortcut in data-collection, and base their methods on individuals’ reports of their own behavior" (Coleman 1969, p. 109, emphasis in the original). The heavy reliance on questionnaires in research areas such as deviance, epidemiology, social norms, or political behavior and attitudes, which has actually increased rather than decreased since Coleman’s comment, motivates research into the validity of self-report measures. Self-reports are especially critical if they concern sensitive topics such as embarrassing questions on extreme political attitudes, sexual behavior, deviant and illegal activities, or health status. If researchers want to draw accurate conclusions based on self-reports, they have to first ensure they have succeeded in revealing the truth at a much more elementary level: that respondents answer their survey questions honestly and accurately.

Measurement accuracy is by no means only a challenge when surveying sensitive topics. But the problem poses itself here in a more accentuated manner. Results from validation studies, that is, studies in which the researcher knows the true answers, illustrate that the proportion of respondents who do not answer sensitive questions truthfully can be substantial. For example, 42 percent (face-to-face interviews) and 33 percent (mail survey) of respondents did not admit they had been convicted in court (Preisendörfer and Wolter 2014). Likewise, 75 percent of respondents who had committed welfare or unemployment benefit fraud denied having done so in face-to-face interviews (van der Heijden et al. 2000). As a result of such misreporting, the prevalence of sensitive behaviors is likely
to be underestimated and estimated correlations between sensitive characteristics and other variables might be biased.

The strategies for mitigating this problem include, for example, choosing a more anonymous interview mode such as paper-and-pencil or online. In part, this seems to work since underreporting of undesirable behavior persisted, but was lower in self- relative to interviewer-administered surveys (Kreuter, Presser, and Tourangeau 2008). Appropriate wording and the choice of the context of questions are also believed to increase self-reports’ accuracy, but their effects are often inconsistent (Krumpal and Nüh 2012). Confidentiality and anonymity assurances are nowadays standard, although it is questionable to what degree respondents perceive them as credible and whether they actually foster a trustworthy interview atmosphere. Many of these strategies are applied habitually, but there is barely any sound evidence on whether and to what extent they manage to improve self-reports’ validity.

A very different and at the time it was presented completely novel approach to tackling the problem of misreporting in sensitive questions was suggested by Stanley Warner in 1965: introducing systematic random error into respondents’ answers to preserve their privacy. Granting response privacy to respondents should, in turn, make them answer more honestly. The method, which he named the Randomized Response Technique (RRT), has been further developed since then, with several new variants being proposed. While the method is a brilliant theoretical solution in that it guarantees the privacy of respondents’ answers and simultaneously allows for accurate aggregate estimates, it only works if respondents properly follow the special RRT procedure. Whether this is the case is an empirical question. Results of published RRT validation studies that set out to answer this have so far been mixed and, as I will show, mostly rest on assumptions that might not be warranted. Hence, we cannot trust their conclusions too much.

The dissertation project presented in the following originally set off to develop RRT implementations that succeed in eliciting truthful answers to sensitive questions, and to apply these methods in substantive studies on academic misbehavior. Yet things turned out slightly differently. After developing some new RRT implementations for online surveys, I validated them together with some other promising implementations using the standard approach: a comparative validation. However, it was soon realized that the standard validation strategy on which most sensitive question research relies has some serious weaknesses. Hence, the project gained a new focus: developing validation designs that actually allow sensitive question techniques to be validated. What set out to develop methods to reveal the truth from survey respondents turned into a project of developing sensitive question validation designs that are able to reveal the truth about
these methods. I believe I partly succeeded in the latter. While this dissertation certainly does not say the last word about the RRT's validity, I hope the work presented here will bring research on sensitive questions a small step closer to the truth. In particular, I provide some useful instruments, i.e., validation designs, that can make efforts to reveal the truth in this area a little more focused and straightforward.

A comparative RRT validation (chapter 2)

Below, I briefly summarize the following substantive chapters. Chapter 2 presents a comparative validation study that assesses five RRT implementations specifically developed for the online mode in a survey on students' cheating and plagiarism (N = 6,037). Even though online surveys may provide more privacy than interviewer-administered surveys, misreporting of socially undesirable items is still an issue (Kreuter, Presser, and Tourangeau 2008) and the application of special sensitive question techniques could be valuable. However, RRT implementation must be tailored to the online mode. Because online survey respondents do not interact with an interviewer who can help them, the RRT procedure must be easy to understand so that a short written explanation enables respondents to follow the procedure and convinces them that the RRT really protects their answers. Further, the randomizing device must be right at hand and ready to use because conventional devices such as a coin or dice that require respondents to pause the survey and leave the screen seem to lead to substantial non-compliance (Coutts and Jann 2011). I therefore used devices that were directly implemented in the online questionnaire: a virtual wheel of fortune (as implemented in Peeters, Lensvelt-Mulders, and Lasthuizen 2010) and the newly developed "pick-a-number" device, a completely trustworthy and reliable randomizing device for the online mode. In addition, I implemented the Benford RRT that uses a special unrelated question as a randomizing device (Diekmann 2012). Prevalence estimates of academic misconduct differed considerably between the various implementations, suggesting that small details have a sizeable impact on RRT estimates. Among all tested implementations, including direct questioning, the unrelated question crosswise-model RRT yielded the highest estimates of sensitive behavior. Hence, a necessary condition for superior validity was fulfilled. However, as the later research in chapter 4 and 5 showed, the higher prevalence was likely not due to an increase in more honest answering but, in contrast, due to false positives.
The Benford RRT and an exploration of privacy (chapter 3)

The article reprinted in chapter 3 takes a closer look at a particular RRT implementation that seemingly fared better than the others in the comparative validation: the Benford RRT, originally suggested in Diekmann (2012). In addition, the chapter looks at finer details of the underlying main rationale of the RRT: privacy protection. The concept of privacy protection in special sensitive question techniques is discussed in a little more detail by reviewing the methodological literature and analyzing the effect of different privacy protection levels on RRT prevalence estimates and on respondents’ perception of privacy. The analyses are based on the data set from the study presented in chapter 2.¹ The starting point is the notion that actual (statistical) privacy protection and protection as perceived by respondents are two different things and whether and how much they are related is an empirical question. The theoretical RRT literature abounds with articles on the privacy-efficiency tradeoff of different RRT techniques and on how to improve one or the other (e.g. Greenberg et al. 1977; Lensvelt-Mulders, Hox, and van der Heijden 2005; Leyseefer and Warner 1976; Zhimin and Zaizai 2012). I examine the relationship between the actual level of privacy and respondents’ perceived privacy. The results suggest that perceived privacy protection is mostly driven by design details other than mere objective privacy protection as defined by the technical design parameters of the RRT.

More is not always better: an individual-level validation (chapter 4)

The study in chapter 4 addresses the biggest weakness of the first study – and of most previous validations: the blind reliance on the more-is-better assumption. In an individual-level validation carried out on Amazon Mechanical Turk (N = 6,152), I consider both sides of the (misreporting) coin: false negatives, i.e. respondents falsely denying sensitive behavior, but also false positives, i.e. respondents falsely admitting sensitive behavior. This is achieved by comparing respondents’ self-reports on cheating in dice games with actual cheating behavior, thereby distinguishing between false negatives, false positives, and accurate responses. Even though the possibility of false positives and their implications for sensitive question technique assessments have been discussed in previous studies (e.g. Lee 1993; Wolter and Preisendorfer 2013; Moshagen et al. 2014; chapter 2), I am aware of only one RRT study since 2000 (John et al. 2013) that took this concern seriously by explicitly analyzing for false positives. I assess several variants of the RRT, including the crosswise-model. The results indicate that the forced-

¹ The samples analyzed in these two chapters differ slightly, resulting in marginally different prevalence estimates.
response RRT and the unrelated-question RRT, as implemented in our survey, fail to reduce the level of misreporting compared to conventional direct questioning. For the crosswise-model RRT, we do observe a reduction of false negatives but at the same time, however, there is a sizeable increase in false positives which led to a higher aggregate prevalence estimate. These higher estimates, interpreted as more valid estimates under the more-is-better assumption, turned out to be the result of false positives, a so far seldom considered type of misclassification. Overall, the crosswise-model produced less valid data than any other evaluated sensitive question technique. This finding is in striking contrast with the positive assessments of the crosswise-model in a series of earlier comparative validation studies (Hoffmann and Musch 2015; Jann, Jerke, and Krumpal 2012; Korndörfer, Krumpal, and Schmukle 2014; Kundt 2014; Kundt, Misch, and Nerré 2014; Shamsipour et al. 2014; chapter 2) and in one individual-level validation not considering false positives (Hoffmann et al. 2015). Hence, the finding demonstrates the importance of considering false negatives as well as false positives when validating sensitive question techniques.

An enhanced comparative validation design (chapter 5)

While the study presented in chapter 4 makes an innovative and significant contribution to sensitive question research by presenting a replicable experimental design able to detect false positives, it has two limitations the study in chapter 5 set out to remedy. First, the individual-level validation in chapter 4 is based on one quite particular item, cheating in an experimental dice game. Second, even though it is quite heterogeneous, the Amazon Mechanical Turk population studied is not representative of the general population nor of many populations of interest for surveys with sensitive questions. In many instances, it might be desirable to validate sensitive question techniques on a particular survey topic and using a particular population of interest where an individual-level validation criterion is just not available. For this, I developed a comparative validation design that is able to detect systematic false positives. This was achieved by introducing zero-prevalence items among the sensitive items, i.e. items with close to zero prevalence in the population. If a method produces an estimate of zero for these items, there are no systematic false positives and the more-is-better assumption is placed on much firmer grounds. If, however, the estimate is not zero, there definitely are false positives and the more-is-better assumption must be refuted. Applying this design in a survey on organ donation and health \( (N = 1,685) \), I replicate the finding that the unrelated question crosswise-model implementation generates considerable false positives. This corroborates the results in chapter 4. The fact that the comparative validation with a zero-prevalence item does not
need an individual-level validation criterion makes it an easy and broadly applicable tool for the development and evaluation of special sensitive question techniques and even for sensitive question research in general. In this sense, it offers a solution to the dilemma that individual-level validations are the most meaningful validations, yet are often impossible to carry out and hard to replicate.
Chapter 2

Sensitive Questions in Online Surveys: An Experimental Evaluation of the RRT and the Crosswise Model

Abstract Self-administered online surveys may 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. We therefore conducted an experimental study in which different implementations of the RRT, including two implementations of the so-called crosswise model, were tested and compared to direct questioning. Our study is an online survey (N = 6,037) on sensitive behaviors by students such as cheating in exams and plagiarism. Results vary considerably between different implementations, indicating that practical details have a strong effect on the performance of the RRT. Among all tested implementations, including direct questioning, the unrelated-question crosswise-model RRT yielded the highest estimates of student misconduct.

This chapter is an edited version of Höglinger, Marc, Ben Jann and Andreas Dirksmann. 2014. "Sensitive questions in Online Surveys: An Experimental Evaluation of the Randomized Response Technique and the Crosswise Model." University of Bern Social Sciences Working Paper No. 9. https://ideas.repec.org/p/bss/wpaper/9.html. We thank Debra Hevenstone for her comments on an earlier draft of this article.
2.1 Introduction

Many empirical studies in the fields of deviance, epidemiology, political opinions, or attitudes are based on self-reports about sensitive behavior or potentially stigmatizing traits. Surveying sensitive topics and obtaining accurate answers to sensitive questions, however, 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 Freisendörfer 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.

2.1.1 Question sensitivity and social-desirability bias

Following Tourangeau and Yan (2007) three types of sensitive questions may be distinguished. First, a question might be perceived as too intrusive and personal. For such a question high rates of nonresponse, but not necessarily a high degree of misreporting, might be expected. Second, a question can involve a threat of disclosure and subsequent sanctions by third parties. For such a question we would expect deliberate misreporting by respondents as a means of self-protection, unless anonymity is guaranteed in a credible way. Third, and more generally, a question can be sensitive in the sense that it refers to the violation of a social norm. In such a case we may expect that respondents tend to answer in accordance with the social norm, leading to so-called social-desirability bias. The misreporting might be due to deliberate "impression management" (Paulhus 1984), or to more subtle processes such as self-deception. Furthermore, the degree to which a question is perceived as sensitive and the answers that are considered as socially desirable or undesirable may depend on context and may differ between subpopulations. Questions on academic misconduct, for instance, the topic we survey in the present study, are perceived as more or less sensitive depending on respondents' personal attitudes, their beliefs about the risk of disclosure and possible sanctions, and their perception of social norms against academic misconduct.
2.1.2 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, as the misclassification resulting from the random noise breaks the link between individual answers and the true value of the sensitive variable and therefore eliminates the risk of disclosure as well as the opportunity for impression management. A widely used RRT variant is the forced-response design proposed by Boruch (1971) and Greenberg et al. (1969), in which respondents employ a randomizing device (e.g., dice, coins) to determine whether they should answer the sensitive question ("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 higher prevalence 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 Freisendörfer 2013). Furthermore, for self-administered online mode, empirical evidence on the performance of the RRT is still scarce and inconclusive.

2.1.3 RRT 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 attenuated. 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 computer-assisted telephone interviews (CATI). However, misreporting remained substantial also in online mode, indicating that the application of sensitive-question techniques such as the RRT could be valuable. Moreover, respondents might actually be more attentive to privacy concerns in online surveys than in CATI or paper-and-pencil interviews (Couper
2000). Results from the few studies comparing RRT to direct questioning in online mode are not very promising. Coutts and Jann (2011) found no higher prevalence estimates for six socially undesirable behaviors using five different forced-response RRT implementations. Quite the contrary, prevalence estimates were often lower than with direct questioning, or even negative due to considerable noncompliance with the RRT procedure. Snijders and Weesie 2008 found similar results with numerous negative prevalence estimates using a forced-response RRT design with a virtual die. Ostapczuk and Musch (2011) as well as Peeters (2005), both using a forced-response RRT design, found no differences in prevalence estimates between RRT and direct questioning. Holbrook and Kroosnick (2010) surveyed voting in the US, a socially desirable behavior, and found unrealistically high voter turnout estimates using various RRT implementations. The only online study we are aware of in which the RRT actually outperformed direct questioning is the study by Jong, Pieters, and Fox (2010), which used a special multi-item RRT design.¹

2.1.4 Reasons for the failure of the RRT in online mode

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 van 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 forced-response variant of the RRT, 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 (Coutts and Jann 2011; Jong, Pieters, and Fox 2010; Holbrook and Kroosnick 2010) are problematic because they require respondents to leave the computer and pause with the survey. This might induce respondents

¹ Furthermore, Moshagen and Musch (2012) found higher prevalence estimates if cheating correction (see footnote 11) was applied. Without cheating correction, however, the RRT estimates were not significantly different from the direct-questioning estimates.
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 (Coutts and Jann 2011; Peeters 2005; Snijders and Weesie 2008) can be manipulated or tracked by experimenters, and thus might not be judged trustworthy by the respondents. Because the randomizing devices employed in most of the published 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.

2.1.5 The crosswise-model RRT

Yu, Tian, and Tang (2008) introduced the crosswise-model RRT as a promising alternative to conventional RRT variants. 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. Respondents should easily understand that the crosswise-model RRT protects their privacy since the possible answers, “the same” or “different”, are obviously ambiguous. Furthermore, there is no clear 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 to the respondents as two questions have to be answered simultaneously and no affirmative or negative answer has to be given. A first empirical application of the crosswise-model RRT in a small-scale 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, evidence on the performance of the crosswise-model RRT is still scarce and the technique has not yet been tested in online mode.

2.1.6 Our study

In our study we compare different variants of the RRT, including the crosswise model, to direct questioning in an 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: Crown and Spiller 1998; McCabe, Trevino, and Butterfield 2001. 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 serious consequences in the case of disclosure. Both universities where the study was conducted have formal rules explicitly stating that cheating on exams and plagiarism will result in disciplinary actions and—depending on the severity of the misconduct and on the context—in sanctions such as a failing grade, expulsion from the respective course or field of study, temporary or indeterminate expulsion from the university, or revocation of an academic title. 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 implementations, if successful, should therefore yield higher estimates of the sensitive behaviors.

The goals of our study are as follows. First, we want to provide evidence on the performance of the RRT in online surveys in general, as convincing evidence that the RRT provides more valid estimates than direct questioning in online surveys is still lacking. Second, we want to evaluate whether the poor performance of the RRT in some of the previous online studies is due to the lack of a good randomizing device. Therefore, we compare a traceable virtual randomizing device, as has been used in previous studies, against a novel virtual randomizing device that cannot be tracked. Third, previous evidence indicates that the often-used forced-response RRT might be subject to noncompliance because respondents are reluctant to provide a “false” forced answer. We therefore compare the forced-response RRT to a design in which respondents answer an unrelated question instead of providing a forced response, a design that might mitigate the noncompliance problem as all respondents provide an answer to a “real” question. Fourth, the unrelated-question RRT still has the problem that there is a clear self-protective answering strategy (always say “no”). The crosswise-model RRT might overcome this problem. Furthermore, we think that the crosswise-model RRT is particularly well suited for use in self-administered online surveys due to its simplicity. We therefore evaluate how the crosswise-model RRT compares to the other RRT variants and whether the promising results of earlier studies can be replicated in online mode. Fifth, a limitation of the classic crosswise-model RRT is that it requires the researcher to come up with sensible unrelated questions for which the probability distribution is known. We therefore evaluate the
performance of a new implementation of the crosswise-model RRT in which the unrelated-questions are replaced by a (non-traceable) virtual randomizing device.

2.2 Data and Methods

2.2.1 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. The survey was implemented using the EFS Survey 8.0 platform by Globalpark AG (see www.unipark.de). It 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 2.1 lists the five sensitive questions in the order they were presented to the respondents.

Table 2.1: Sensitive questions on student misconduct (translated from German)

<table>
<thead>
<tr>
<th>Item</th>
<th>Wording</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copying from other students in exam</td>
<td>In your studies, have you ever copied from other students during an exam?</td>
</tr>
<tr>
<td>Using crib notes in exam</td>
<td>In your studies, have you ever used illicit crib notes in an exam (including notes on mobile phones, calculators or similar)?</td>
</tr>
<tr>
<td>Taking drugs to enhance exam performance</td>
<td>In your studies, have you ever used prescription drugs to enhance your performance in an exam?</td>
</tr>
<tr>
<td>Including plagiarism in paper</td>
<td>In your studies, have you ever handed in a paper containing a passage intentionally adopted from someone else’s work without citing the original?</td>
</tr>
<tr>
<td>Handing in someone else’s paper</td>
<td>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?</td>
</tr>
</tbody>
</table>

For details on the questionnaire development (several rounds of pretesting, cognitive and quantitative, were carried out) and the fieldwork see the data documentation (Höglinger, Jann, and Diekmann 2014a). 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% (AAPOR 2011).\footnote{At the University of Bern, the response rate was considerably lower (28.9% of 8,610 invited students) than at ETH Zurich (37.1% of 10,800 invited students). At the University of Bern, due to data protection regulation, the student administration office submitted the invitations. Reminder emails were not possible. At ETH Zurich, the research team submitted the invitations. A reminder email was sent to students who did not respond within three weeks. The difference in response rates is due to the effect of the reminder email. The sample at ETH Zurich includes 200 observations from the last quantitative pretest as a random sample was used for the pretest and no changes were made to the design and questionnaire after the pretest. Excluding these observations does not change our findings (results without these observations are available in the online supplement).} 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). Furthermore, we exclude 272 respondents whose mother tongue is not German and who did not assess their German to be at least "good".\footnote{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).} The resulting sample size is 6,037.

### 2.2.2 Experimental conditions

Respondents were randomly assigned to one of six experimental conditions: direct questioning, one of two implementations of the forced-response RRT, an implementation of the unrelated-question RRT, or one of two implementations of the crosswise-model RRT. Table 2.2 provides an 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 slightly differ by experimental condition and question (available sample sizes per experimental condition are between 963 and 983 respondents for the items on behavior in exams and between 710 and 725 respondents for the items on plagiarism).

The direct questioning condition (DQ) served as a benchmark for the evaluation of the different RRT variants. 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.
Table 2.2: Experimental conditions and number of observations

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

five sensitive questions (see table 2.1) then followed one by one on separate screens. Each question could be answered with “yes” or “no”.

The first variant of the RRT (“FR Wheel”) used a symmetric forced-response design (Boruch 1971; Greenberg et al. 1969) and a virtual random wheel as randomizing device.\(^4\) 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 2.1). 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).\(^5\)

The virtual random wheel corresponds to the classic spinner used in some early variants of the RRT (see Fox and Tracy 1986, p. 39). Peeters (2005; also see Peeters, Lensvelt-Mulders, and Lasthuizen 2010) presented a first online im-

---

\(^4\) In a symmetric design the forced response can be either “yes” or “no”. Such a design seems to be preferable over an asymmetric design, in which the forced response is always “yes” (or always “no”, depending on context) (Ostapczuk et al. 2009).

\(^5\) Respondents were randomized between a lower privacy protection scheme and a higher privacy protection scheme (9 “Answer Question” sectors versus 8 “Answer Question” sectors). Similar privacy protection variations were employed for the other RRT implementations. Results for the two protection schemes were very similar. We therefore do not report results from separate analyses.
Figure 2.1: Screen shot of the forced-response random wheel implementation (“FR Wheel”, translated from German)

Implementation of such a spinner. Because the outcome of a virtual random wheel could easily be tracked or even predetermined (it was not in our application), we would expect that respondents do not trust the virtual random wheel. The same problems exist with virtual dice or coins, which have been used frequently in past studies (Coutts and Jann 2011; Lensvelt-Mulders et al. 2006; Snijders and Weeze 2008). We included this condition in our study to evaluate empirically whether respondents actually do mistrust such a virtual randomizing device.

For our second variant of the forced-response RRT (“FR Number”) we developed a new randomizing device that is more credible than the virtual random wheel because it cannot be tracked. Apart from the randomizing device, “FR Number” was identical to “FR Wheel”. The new pick-a-number 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 them.

1. Please pick one of the twelve fields.

<table>
<thead>
<tr>
<th>Field</th>
<th>Answer Question</th>
<th>Directly tick Yes</th>
<th>Directly tick No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Now click the “Show instructions” button:

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

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

Yes  No

Please tick the corresponding answer on the right →

Figure 2.2: Screen shot of the forced-response pick-a-number implementation (“FR Number”, translated from German)

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 2.3). 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 2.4).7

Diekmann (2012) provides empirical evidence that first digits of house numbers provided by respondents follow “Benford’s Law”. According to the law, for

---

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

7 The auxiliary questions asked about the mother’s birthday being in the first half of the year, being in an even-numbered month, being in the first half of the month, being on an even-numbered day, or being in an even-numbered year, respectively. They were randomly paired with the sensitive questions for each respondent. See the data documentation in the online supplement for details.
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 7 for number 73, or number 348).

---

**Question 1**

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.)

---

Figure 2.4: Screen shot of the unrelated-question Benford implementation (“UQ Benford”), screen 2 (translated from German)

example, the probability of 1, 2, 3, or 4 is 0.699. 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).8

Our first implementation of the crosswise-model RRT (“CM Question”) used an unrelated-question design as employed 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 2.5).9

---

8 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).

9 Again, the non-sensitive questions were randomly paired with the sensitive questions for each respondent. The questions asked about the mother’s or father’s birthday being in a specific part of the year or a in specific part of the month, or about the last digit of the parent’s phone number. See the data documentation in the online supplement for details.
2.2. Data and Methods

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 instructions for the forced response RRT.

Question pair 1

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 2.6: Screen shot of the pick-a-number crosswise-model implementation ("CM Number", translated from German)

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

$$\hat{Y} = \frac{Y - p^{yes}}{1 - p^{yes} - p^{no}}$$

provides a consistent estimate of $\Pr(S = 1)$. The same transformation can also be employed for data from the unrelated-question RRT, setting $p^{yes} = p^* p^{yes,u}$ and $p^{no} = p^*(1 - p^{yes,u})$, where $p^*$ 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

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

where $p^{yes,u}$ is again the probability of a “yes” answer to the unrelated question.\(^\text{10}\)

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 2010). An equivalent approach, followed in the analyses below, is to estimate a least-squares regression on $\hat{Y}$ across the whole sample including dummy variables for the different sensitive-question techniques (with $\hat{Y} = Y$ for direct questioning), employing heteroscedasticity robust formulas for standard errors (Jann 2008). Such an integrated model is convenient because it readily provides

---

\(^{10}\) As in the original Warner scheme, $p^{yes,u}$ must be unequal 0.5 for the crosswise-model estimate to be identified.
tests for differences among techniques. Furthermore, additional covariates can be included in the model to analyze effects of predictors of sensitive behaviors.11

2.3 Results

2.3.1 Question sensitivity

A prerequisite for the validity of our evaluation of the different sensitive question techniques is that respondents perceive the questions we asked as sensitive. As mentioned above, the universities at which our study was conducted have formal rules about how to sanction cheating on exams and plagiarism. The sanctions can be severe and the students seem to be well aware of that fact (for example, 26% of our respondents believe that they will be expelled from their studies if they get caught plagiarizing in a Bachelor’s or Master’s thesis; overall, serious sanctions are expected by 89% of the respondents). We therefore assume that the threat of disclosure is of serious concern to our respondents. Furthermore, strong norms against academic misconduct appear to exist among the respondents so that socially desirable responses can be expected. Table 2.3 provides evidence on three dimensions of norm prevalence (see, e.g., Bicchieri 2006): the percentage of students who the respondents believe have never engaged in the specific behaviors (perceived descriptive norm), the percentage of respondents who think the specific behaviors are bad or very bad (personal norm), and the percentage of respondents who believe that most others consider the specific behaviors as bad or very bad (perceived general norm).

11 An alternative approach would be to use suitably modified maximum-likelihood logistic regression (Maddala 1983; Jann 2005; also see Jann, Ferke, 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). Interesting extensions to these approaches are so-called cheating-correction methods that exploit variations in design parameters (e.g., Clark and Desharnais 1998; Moshagen, Musch, and Endfelder 2012; Moshagen and Musch 2012; van den Hout, Böckenholt, and van der Heijden 2010) or response patterns across multiple items (Böckenholt and van der Heijden 2007; Jong, Pieters, and Stremersch 2012) to identify the proportion of respondents who do not comply with the RRT instructions, and correct the prevalence estimates accordingly. We do not employ such methods here because the variation in design parameters is too low in our study for the cheating-correction estimates to be efficient and also because additional assumptions are required (such as, e.g., that the variation in design parameters has no effect on the willingness to provide a truthful answer).
Table 2.3: Norms against academic misconduct

<table>
<thead>
<tr>
<th>Sensitive behavior</th>
<th>Descriptive norm</th>
<th>Personal norm</th>
<th>General norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copying from other students in exam</td>
<td>77%</td>
<td>39%</td>
<td>31%</td>
</tr>
<tr>
<td>Using crib notes in exam</td>
<td>81%</td>
<td>50%</td>
<td>35%</td>
</tr>
<tr>
<td>Taking drugs to enhance performance</td>
<td>87%</td>
<td>62%</td>
<td>50%</td>
</tr>
<tr>
<td>Including plagiarism in paper</td>
<td>89%</td>
<td>80%</td>
<td>69%</td>
</tr>
<tr>
<td>Handing in someone else’s paper</td>
<td>94%</td>
<td>94%</td>
<td>85%</td>
</tr>
</tbody>
</table>

*Notes:* Descriptive norm (perceived norm compliance): mean of respondents’ estimate of the percentage of students who never engaged in the behavior; Personal norm: percentage of respondents who think the behavior is rather bad or very bad; General norm: percentage of respondents who believe that most people think the behavior is rather bad or very bad; N between 5871 and 5921.

The results in table 2.3 reveal a consistent ordering of the five sensitive questions. Compared to the other behaviors, compliance to norms against copying from other students and using crib notes is perceived as relatively low, with an average estimated percentage of students who never engaged in these behaviors of 77% and 81%, respectively. Furthermore, only 39% to 50% of respondents consider these behaviors as bad or very bad, and 31% to 35% of respondents believe that most others consider these behaviors as bad or very bad. For plagiarism, perceived norm compliance is substantially higher (89% and 94%) and the vast majority of respondents think that these behaviors are bad or very bad (80% and 94%) and that most others consider these behaviors as bad or very bad (69% and 85%). The prevalence of the norm against taking drugs to enhance exam performance, for which no formal sanctions are defined at the two universities, lies between the prevalence of the norms against exam cheating and plagiarism. About 60% of respondents consider this behavior as bad or very bad.

In sum, although differences exist, in particular between exam cheating and plagiarism, there seem to be significant norms against those behaviors we study. Together with the possible sanctions in case of disclosure (for four of the five questions) we therefore suppose that the questions in our survey appeared sensitive to at least a substantial proportion of the respondents. For the more sensitive items (plagiarism), we expect a larger share of norm-offenders to misreport so that the sensitive question techniques, should they be successful in reducing misreporting, will have a stronger (relative) effect. Yet, because the true share of norm-offenders is likely lower for these behaviors, the observable absolute effect of the sensitive question techniques may be lower than for the less sensitive items.
2.3.2 Prevalence estimates by experimental conditions

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 truthfully. Hence, relying on the “more-is-better” assumption (Lensvelt-Mulders et al. 2005) we interpret a positive difference to DQ as evidence for a technique’s superior validity. We will come back to this assumption in the discussion.

The left panel in figure 2.7 depicts the point estimates of the proportion of respondents admitting a particular sensitive behavior and the corresponding 95%-confidence intervals by experimental condition (also see table 2.A.1 in the appendix). Differences in the prevalence estimates between a particular RRT implementation and DQ are shown in the right panel. The crosswise-model RRT implementation using unrelated questions (“CM Question”) produced the highest estimates of all implementations for four out of the five items. Furthermore, the difference between “CM Question” and DQ is substantial for all items and highly significant for three of them (“copying from others”, “using crib notes”, and “taking drugs to enhance performance”). The size of the absolute differences between “CM Question” 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. 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 less favorable. 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 two forced-response RRT implementations (“FR Wheel” and “FR Number”) are disillusioning. In only two out of ten comparisons did these implementations yield a significantly higher prevalence estimate than DQ (“RRT Wheel” for “copying from others”, “RRT Number” for “using crib notes”). On the other hand, there are three cases in which one of these implementations produced significantly lower estimates than DQ. In fact, in these three cases the prevalence estimate is negative (significantly negative in one case). This suggests that there was substantial noncompliance with the RRT instructions, that is, that many respondents answered “no” even though the procedure instructed

---

12 Negative estimates do not make sense, of course, and are a result of the data violating our assertion about how they came about. Forcing the prevalence estimate into $[0,1]$ could easily be achieved (e.g. using maximum-likelihood techniques; see footnote 11), but doing so would obscure the fact that there is a problem with the data.
Figure 2.7: Prevalence estimates and difference to DQ by experimental condition

them to respond “yes.” Unfortunately, due to the nature of the RRT, it is not possible to identify noncompliance with the RRT instructions at the individual level, which hampers an in-depth analysis of instruction noncompliance. Finally, 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.13

---

13 As discussed above, the design parameters of the sensitive question techniques were varied among respondents, leading to somewhat different levels of respondent protection. We found no evidence whatsoever that these variations affected the respondents’ answers to the sensitive questions (results available in the online supplement). However, we do find some weak 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 always random answer “no” was more frequent.
2.3. Results

In sum, the unrelated-questions crosswise-model RRT ("CM Question") consistently produced higher prevalence estimates than direct questioning for all sensitive items. The alternative implementation of the crosswise-model RRT ("CM Number"), however, produced higher prevalence estimates for only two out of five items. Prevalence estimates from the two forced-response RRT implementations are comparable to the direct-questioning estimates, or are even lower. This casts serious doubt on the validity of the estimates from the forced-response RRT implementations. The unrelated-question RRT implementation ("UQ Benford") performed similar to "CM Number". Comparing the relative effects of the techniques between sensitive questions does not promise much insight given the poor overall performance of most of the techniques. However, relative effects for the technique with the highest face validity, "CM Question", indicate that, as expected, effects are weaker for the less sensitive questions on cheating in exams (70 to 100% increase in prevalence estimates compared to direct questioning) than for the more sensitive questions on plagiarism (160 to 300% increase). Surprisingly, however, the effect is strongest for the question on taking drugs to enhance exam performance (350% increase). Yet, there is too little statistical power to draw firm conclusions about the differences among these relative effects (an overall test has a p-value of 0.104; among the 10 possible contrasts, only the difference in relative effects between taking drugs and copying from other students, p = 0.014, and between taking drugs and using crib notes, p = 0.035, are significant at the 5% level).

2.3.3 Alternate quality criteria

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 2.8 displays results for quality criteria available for all techniques including direct questioning, the right panel contains results from additional criteria available only for the RRT implementations (also see table 2.A.2 in the appendix).

The RRT places 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 did not complete the interview among the respondents who reached the introductory screen for the sensitive questions) from about 1% for DQ to about 2% or 3% for the RRT implementations (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
Figure 2.8: Comparison of experimental conditions on various measures
the RRT implementations (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 all RRT implementations require much more answering time than DQ (third graph on the left in figure 2.8). 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 should increase respondents' trust in the protection of their privacy. After all, this is the assumed mechanism by which these techniques are supposed to increase honest answering. At the end of the interview, we asked the respondents about how much they trusted in the protection of privacy by the survey ("Please be honest: How much do you trust in our measures for anonymity and privacy protection of the participants of this survey?"). The fourth graph on the left in figure 2.8 shows the percentage of respondents who answered "rather much" or "very much." Levels of self-reported 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 salience 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 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 one could discover whether a survey participant engaged in one of the sensitive behaviors ("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?"). The lowest graph on the left in figure 2.8 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 2.8 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 ("How cumbersome was the application of this special survey technique to you?"); whether they thought that they applied the technique correctly ("Do you think that you applied the special survey technique correctly in each case?"); whether they were convinced that the technique protected their answers ("What is your personal opinion: Does the special survey technique provide 100% protection of your answers to the sensitive questions?"); whether they thought that the technique was a reasonable approach to protect respondents' privacy ("How reasonable do you think is the use if this survey technique to protect the answers of survey participants to sensitive questions?"); and whether they believed that they understood how the technique protects their answers ("Do you understand why the employed survey technique provides 100% protection of your 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 Number", "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 Number". The third plot on the right in figure 2.8 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 did not 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 2.8; 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 self-reported 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 2.8). 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 sensitive questions).

To test for effects of respondents' perceptions of the sensitive question techniques on prevalence estimates we ran regressions on all self-reported quality criteria. Table 2.4 summarizes the results from these regressions. The only notable results are that, for "UQ Benford", perceived cumbersomeness is associated with increased prevalence estimates ($p < 0.001$) and correct application is associated with decreased prevalence estimates ($p = 0.032$) and, for "CM Number", perceived reasonableness of the technique to protect privacy is associated with decreased prevalence estimates ($p = 0.028$; using joint tests across all five sensitive questions). However, we could not find a robust effect of any of the surveyed quality criteria on prevalence estimates in general, that is, across more than one RRT implementation.

In sum, compared to direct questioning, all RRT implementations come at large costs with respect to answering time, but increases in break-off rates and item-nonresponse are only small. Using sensitive question techniques seems to undermine respondents' general trust in the survey, but at the same time respondents consider the risk of disclosure lower if questioned by the RRT than by direct questioning. Perhaps the most striking result is that only between 57% and 75% of respondents claim that they understood how the RRT protects their answers. However, none of the surveyed subjective evaluation criteria shows a consistent correlation with the propensity to admit a sensitive behavior.
Table 2.4: Summary of effects of evaluation criteria on prevalence estimates

<table>
<thead>
<tr>
<th></th>
<th>DQ</th>
<th>FR Wheel</th>
<th>FR Number</th>
<th>UQ Benford</th>
<th>CM Question</th>
<th>CM Number</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust in anonymity</td>
<td>(+)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>5879</td>
</tr>
<tr>
<td>Disclosure risk</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>5869</td>
</tr>
<tr>
<td>Technique is cumbersome</td>
<td>n.s.</td>
<td>n.s.</td>
<td>+++</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>4861</td>
</tr>
<tr>
<td>Applied technique correctly</td>
<td>n.s.</td>
<td>n.s.</td>
<td>-</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>4861</td>
</tr>
<tr>
<td>Technique protects</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>(-)</td>
<td>n.s.</td>
<td></td>
<td>4859</td>
</tr>
<tr>
<td>Technique is reasonable</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>-</td>
<td></td>
<td>4858</td>
</tr>
<tr>
<td>Understood principle</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td></td>
<td>4861</td>
</tr>
</tbody>
</table>

Notes: + mostly positive; − mostly negative; joint F test: (*) p < .1, * p < .05, ** p < .01, *** p < .001, where + stands for + or −; n.s.: joint F test not significant; computations based on dichotomized evaluation criteria as used for figure 2.4; detailed results are available in the online supplement.

2.4 Discussion and Conclusions

Three main findings result from our study. First, different implementations of the RRT, even of the same variant but using different randomizing devices, can produce quite diverse estimates of sensitive behaviors. It is, therefore, difficult to draw a final conclusion about the RRT based on the evaluation of just one implementation, an aspect that is ignored in most studies. The high variability of results across implementations is not very helpful for clarifying whether the RRT is a suitable sensitive question technique for online surveys. However, it clearly shows that drawing final conclusions based on just one or two implementations might be premature (e.g., Holbrook and Krosnick 2010).

Second, the forced-response RRT variants (“FR Wheel”, “FR Number”), as implemented in our study, did not yield systematically higher estimates than direct questioning. They even produced negative estimates in some cases. This questions the viability of the forced-response RRT variant for online surveys. The reason for these low or even negative 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 forced-response RRT. Although a lot of effort has been put into pretesting and finding good implementations, no convincing evidence could be found that forced-response RRT variants yield 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. The unrelated-question RRT implementation “UQ Benford” performed somewhat better, generating similar estimates as DQ for three items and higher estimates for two items. However, with respect to respondents’ assessment of the technique in terms of understanding, protection, and reasonableness, “UQ Benford” fared among the worst of the techniques we evaluated.

Third, the unrelated-question crosswise-model RRT implementation (“CM Question”) produced higher prevalence estimates than direct questioning for all sensitive questions (significantly higher in three cases). Assuming the “more-is-better” assumption is valid, “CM Question” succeeded in eliciting more truthful answers to the sensitive questions than direct questioning and, hence, produced more valid estimates. “CM Question”, therefore, seems to be a promising alternative to conventional RRT variants. 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 most other RRT variants.14 A drawback of the crosswise-model RRT compared to forced-response or unrelated-question RRT, however, is its lower statistical efficiency (compare the confidence intervals in figure 2.7 or the standard errors in table 2.1A.2). Another critical point is that results for the crosswise-model RRT implementation employing an explicit randomizing device (“CM Number”) are inconclusive as this implementation yielded higher estimates than DQ for only two items (statistically significant in one case). That is, also for the crosswise-model RRT the details of implementation seem to matter.

That the unrelated-question crosswise-model RRT performed well did not come as a big surprise given the preliminary positive findings of some earlier studies. However, whether its results can be considered more valid than the results from DQ depends on the viability of the “more-is-better” assumption, a limitation shared with most other studies on sensitive question techniques. Higher estimates are a necessary condition for the validity of a technique’s results if – as suggested by a number of validation studies (e.g., Kreuter, Presser, and Tourangeau 2008; Preisendörfer and Wolter 2014; van der Heijden et al. 2000) – DQ is affected by

---

14 Detection of the optimal self-protective answer strategy would require a thorough understanding of Bayesian updating and the crosswise-model principle by respondents. If \( p_{\text{true}} < 0.5 \), the optimal self-protective answer is “the same”; if \( p_{\text{true}} > 0.5 \), the optimal self-protective answer is “different”.

underreporting. Yet, higher estimates may not be sufficient. It is possible that higher estimates come about by some other mechanisms than an increase in the share of respondents who answer truthfully. For example, if many respondents are confused by the instructions of the crosswise-model RRT and provide random answers, prevalence estimates will be biased towards 50% (although, in this case, we would expect a percentage-point deviation from the DQ results that is more or less constant across items, a pattern which is not observed in our study). Therefore, even though good opportunities for validation are notoriously hard to find, the next step in this research program should be a study in which respondents' answers are compared to known true values. Furthermore, 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.

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. In our study we emphasized the choice of the randomizing device and the basic RRT design. Our results suggest that using an explicit randomizing device such as a virtual random wheel or the pick-a-number device does not work so well and that using unrelated questions might be preferable. 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. Overall, 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.
## 2.A Appendix

The data and documentation of the survey and the analysis scripts are provided in the online supplement at ftp://repec.sowi.unibe.ch/files/wp8/ and ftp://repec.sowi.unibe.ch/files/wp9/.

Table 2.A.1: Prevalence estimates by experimental condition (in percent; standard errors in parentheses)

<table>
<thead>
<tr>
<th>Levels</th>
<th>Copying from other students in exam</th>
<th>Using crib notes in exam</th>
<th>Taking drugs to enhance exam performance</th>
<th>Including plagiarism in paper</th>
<th>Handing in someone else’s paper</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct questioning (DQ)</strong></td>
<td>17.88 (1.23)</td>
<td>9.09 (0.92)</td>
<td>3.38 (0.58)</td>
<td>2.90 (0.62)</td>
<td>1.52 (0.45)</td>
</tr>
<tr>
<td><strong>FR Wheel</strong></td>
<td>22.80 (2.14)</td>
<td>11.28 (1.96)</td>
<td>-0.89 (1.67)</td>
<td>0.94 (2.01)</td>
<td>0.46 (2.00)</td>
</tr>
<tr>
<td><strong>FR Number</strong></td>
<td>18.78 (2.08)</td>
<td>13.86 (2.00)</td>
<td>-1.52 (1.64)</td>
<td>2.95 (2.07)</td>
<td>-4.25 (1.82)</td>
</tr>
<tr>
<td><strong>UQ Benford</strong></td>
<td>17.24 (1.91)</td>
<td>12.93 (1.83)</td>
<td>4.67 (1.63)</td>
<td>7.68 (1.98)</td>
<td>2.43 (1.81)</td>
</tr>
<tr>
<td><strong>CM Question</strong></td>
<td>30.06 (2.90)</td>
<td>18.37 (2.80)</td>
<td>15.26 (2.80)</td>
<td>7.61 (3.08)</td>
<td>6.12 (3.05)</td>
</tr>
<tr>
<td><strong>CM Number</strong></td>
<td>24.74 (2.73)</td>
<td>10.88 (2.56)</td>
<td>4.62 (2.45)</td>
<td>8.45 (2.92)</td>
<td>0.14 (2.73)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Differences</th>
<th>Copying from other students in exam</th>
<th>Using crib notes in exam</th>
<th>Taking drugs to enhance exam performance</th>
<th>Including plagiarism in paper</th>
<th>Handing in someone else’s paper</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FR Wheel – DQ</strong></td>
<td>4.93 (2.47)</td>
<td>2.19 (2.17)</td>
<td>-4.27 (1.77)</td>
<td>-1.96 (2.10)</td>
<td>-1.06 (2.05)</td>
</tr>
<tr>
<td><strong>FR Number – DQ</strong></td>
<td>0.90 (2.41)</td>
<td>4.77 (2.20)</td>
<td>-4.90 (1.74)</td>
<td>0.04 (2.16)</td>
<td>-5.77 (1.88)</td>
</tr>
<tr>
<td><strong>UQ Benford – DQ</strong></td>
<td>-0.63 (2.27)</td>
<td>3.84 (2.05)</td>
<td>1.29 (1.73)</td>
<td>4.77 (2.08)</td>
<td>0.91 (1.87)</td>
</tr>
<tr>
<td><strong>CM Question – DQ</strong></td>
<td>12.18 (3.15)</td>
<td>9.28 (2.95)</td>
<td>11.88 (2.86)</td>
<td>4.70 (3.14)</td>
<td>4.60 (3.08)</td>
</tr>
<tr>
<td><strong>CM Number – DQ</strong></td>
<td>6.87 (2.99)</td>
<td>1.79 (2.72)</td>
<td>1.24 (2.52)</td>
<td>5.55 (2.99)</td>
<td>-1.38 (2.77)</td>
</tr>
</tbody>
</table>

**N**
5859
5847
5827
4318
4311
### Table 2.A.2: Comparison of experimental conditions on various measures

<table>
<thead>
<tr>
<th></th>
<th>Break-off (%)</th>
<th>Item nonresponse (%)</th>
<th>Answering time (seconds)</th>
<th>Trust in anonymity (%)</th>
<th>Disclosure risk (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct questioning</td>
<td>1.20 (0.34)</td>
<td>0.55 (0.21)</td>
<td>43.00 (0.73)</td>
<td>80.61 (1.26)</td>
<td>28.82 (1.45)</td>
</tr>
<tr>
<td>FR Wheel</td>
<td>3.27 (0.56)</td>
<td>1.84 (0.40)</td>
<td>188.00 (2.28)</td>
<td>69.22 (1.48)</td>
<td>22.93 (1.35)</td>
</tr>
<tr>
<td>FR Number</td>
<td>2.76 (0.51)</td>
<td>1.91 (0.40)</td>
<td>183.00 (2.44)</td>
<td>73.15 (1.41)</td>
<td>19.49 (1.26)</td>
</tr>
<tr>
<td>UQ Benford</td>
<td>2.00 (0.44)</td>
<td>0.98 (0.27)</td>
<td>165.00 (2.12)</td>
<td>73.37 (1.41)</td>
<td>20.94 (1.30)</td>
</tr>
<tr>
<td>CM Question</td>
<td>2.78 (0.52)</td>
<td>1.39 (0.32)</td>
<td>150.00 (1.87)</td>
<td>76.37 (1.36)</td>
<td>25.38 (1.39)</td>
</tr>
<tr>
<td>CM Number</td>
<td>3.39 (0.57)</td>
<td>2.30 (0.45)</td>
<td>190.00 (2.61)</td>
<td>76.65 (1.36)</td>
<td>20.00 (1.28)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Technique is cumbersome (%)</th>
<th>Applied technique correctly (%)</th>
<th>Technique protects (%)</th>
<th>Technique is reasonable (%)</th>
<th>Understood principle (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR Wheel</td>
<td>14.18 (1.12)</td>
<td>95.06 (0.70)</td>
<td>56.54 (1.59)</td>
<td>53.44 (1.60)</td>
<td>60.12 (1.57)</td>
</tr>
<tr>
<td>FR Number</td>
<td>12.99 (0.98)</td>
<td>92.41 (0.85)</td>
<td>67.35 (1.50)</td>
<td>59.28 (1.57)</td>
<td>66.16 (1.51)</td>
</tr>
<tr>
<td>UQ Benford</td>
<td>9.57 (0.94)</td>
<td>94.87 (0.71)</td>
<td>61.66 (1.56)</td>
<td>53.96 (1.60)</td>
<td>57.19 (1.59)</td>
</tr>
<tr>
<td>CM Question</td>
<td>8.59 (0.90)</td>
<td>97.03 (0.54)</td>
<td>67.42 (1.50)</td>
<td>59.90 (1.57)</td>
<td>62.22 (1.55)</td>
</tr>
<tr>
<td>CM Number</td>
<td>11.70 (1.03)</td>
<td>95.66 (0.66)</td>
<td>75.03 (1.39)</td>
<td>62.53 (1.56)</td>
<td>65.63 (1.53)</td>
</tr>
</tbody>
</table>

| N                   | 4867                        | 4865                            | 4862                    | 4862                      | 4865                     |

*Notes: Results for “Trust in anonymity” through “Understood principle” are based on dichotomized 5-point scales ("very much/likely" or "rather much/likely" versus "partly/rather unlikely/somewhat", "rather not/very unlikely/slightly", or "not at all/impossible/definitely not").*
Chapter 3

A New Randomizing Device for the RRT Using Benford’s Law: An Application in an Online Survey

Abstract The randomizing device is a crucial feature of any implementation of the Randomized Response Technique (RRT). Respondents’ cooperation and compliance with the RRT procedure hinge heavily on the device’s ease of use, its trustworthiness, and its availability. We introduce Benford RRT, a new randomizing device based on Benford’s law that uses a randomizing question and does not need any physical artifact. Therefore, it is particularly suitable for self-administered surveys and telephone surveys. A first application in an online survey on student cheating behavior shows that Benford RRT performed well and generated always similar or higher prevalence estimates of cheating than direct questioning. Additional analyses reveal that small changes in the probability \( p \) with which respondents are instructed to answer the sensitive question have no effect on prevalence estimates or respondent’s evaluation of privacy protection. This suggests that the perceived privacy protection of a particular RRT implementation is mainly driven by other design considerations than the mere statistical protection level.


We thank Ben Jann for his support in the design of this study as well as for valuable suggestions on improving the manuscript, and Claudia Jenny for proofreading.
3.1 Introduction

A crucial feature of any implementation of the Randomized Response Technique (RRT) is the randomizing device. It determines whether or not a particular respondent is required to answer a sensitive question and, consequently, protects respondents' privacy. Respondents' cooperation and compliance with the RRT procedure hinge heavily on the device's ease of use, its trustworthiness, and its availability. Dice, a box with colored balls, a spinner, or cards have frequently been used in face-to-face interviews. But these are difficult to use in paper-and-pencil, online or telephone surveys because no interviewer is present to provide them to respondents. More commonly available objects that can be used as randomizing device, such as coins, are preferable but might still be out of some respondents' immediate reach. This may lead to RRT break-offs or noncompliance and, as a consequence, invalid measurement. A solution to this problem of availability is to avoid physical randomizing devices and to use questions instead. However, such "randomizing questions" have so far rarely been used, as the range of suitable questions is very restricted.

In this chapter, we present a new randomizing device originally proposed in Diekmann (2012). It uses a "randomizing question" and comprises several desirable properties. Besides its ease of use and its applicability in all survey situations, it allows for increasing the statistical efficiency of the RRT without jeopardizing respondents' perceived privacy protection. For the latter, the method makes use of Benford’s law and takes advantage of respondents’ misperception of the properties of Benford-distributed numbers such as, in our example, house numbers. We show how this method can be implemented and we present results of a first large-scale empirical evaluation in an online survey on student cheating. Furthermore, we will discuss the important difference between respondents' objective privacy protection in RRT designs and their subjectively perceived privacy protection.

3.2 The Randomized Response Technique (RRT)

The Randomized Response Technique (RRT) is a well-known method to elicit more valid answers to sensitive questions in surveys (originally Warner 1965, for an overview see Fox and Tracy 1986 and Krumpal et al. 2015). It provides complete concealment of respondents’ answers by introducing a systematic random error, which inhibits any inference of admittance or non-admittance of sensitive behavior from an individual response. This is achieved by a randomizing device such as two dice. The randomizing device determines in the case of the unrelated-
question RRT variant (Horvitz, Shah, and Simmons 1967; Greenberg et al. 1969), which serves in the following as exemplary case, whether a particular respondent has to either answer a sensitive or a non-sensitive question. Respondents could, for instance, be instructed to throw two dice and answer the sensitive question “Have you ever cheated on your taxes?” if their outcome is 1 to 8 and to answer the non-sensitive question “Is your mother’s birthday in the months of January through June?” if their outcome is 9 to 12. As only the respondent knows the outcome of the dice throw, no one else is able to infer whether the response given is actually related to the sensitive behavior or not. Accordingly, respondents do not have to fear negative consequences of any kind by admitting a sensitive behavior and should feel free to answer truthfully.

3.2.1 Estimating the prevalence of sensitive behavior with the RRT

Even though individual responses are completely concealed, the prevalence of the sensitive behavior can be consistently estimated in the aggregate. The researcher simply takes into account that the observed “yes” responses are not only generated by respondents answering “yes” to the sensitive question but also by respondents answering “yes” to the non-sensitive question. Let $p$ be the probability that respondents are instructed to answer the sensitive question and $1 - p$ the probability for answering the non-sensitive question whose answer distribution $P(\text{yes} | \text{nonsens. quest.})$ is known. The share of observed “yes” answers is defined as

$$P(\text{yes observed}) = p * P(\text{yes} | \text{sens. quest.}) + (1 - p) * P(\text{yes} | \text{nonsens. quest.})$$

By rearranging the equation, we get the share of respondents answering “yes” to the sensitive question, and, hence, the prevalence of the sensitive behavior under the condition that respondents complied to the RRT instructions:

$$P(\text{yes} | \text{sens. quest.}) = \frac{P(\text{yes observed}) - (1 - p) * P(\text{yes} | \text{nonsens. quest.})}{p}$$

The variance of the RRT estimator is then given by (e.g., Fox & Tracy, 1986, p. 19):

$$\text{var}(P(\text{yes} | \text{sens. quest.})) = \frac{P(\text{yes observed}) * (1 - P(\text{yes observed}))}{n * p^2}$$

The variance is inversely related to $p^2$, hence the lower the probability that respondents have to answer the sensitive question, the higher the variance of the estimator of the sensitive behavior. Respondents' privacy protection comes at the cost of a lower statistical efficiency of the RRT estimator.
3.2.2 The RRT randomizing device and its requirements

The RRT randomizing device serves to introduce randomness into the answering process of survey respondents and, therefore, is the central part of any RRT implementation. The principal requirements a randomizing device has to meet are ease of use, trustworthiness, and availability. Ease of use means that respondents are able to carry out the randomization quickly and without too much effort. Throwing two dice, for instance, does not have to be explained and takes only seconds if dice are readily available.

Trustworthiness regarding the randomizing device means that respondents understand that the outcome of the randomization procedure is truly random and that they believe that the outcome is not detectable by somebody else. The first aspect of trustworthiness, understanding, is assured for well-known randomizing procedures such as throwing dice, flipping a coin, or drawing a card from a deck. Nevertheless, true randomness may be posed into question if uncommon or novel random devices are used, such as picking numbers on a screen or using digits of a phone number. Randomness may also be posed into question when the outcome distribution is susceptible to manipulation. This is the case with most “virtual” randomizing devices implemented in online surveys, such as digital coins, dice, or spinners (see Peeters, Lensvelt-Mulders, and Lasthuizen 2010; Coutts and Jann 2011 for implementations).

The second aspect of trustworthiness, confidence in the undetectability of the outcome, is often an issue when the RRT is used in interviewer-administered surveys. Respondents might suspect that the interviewer is somehow able to observe the outcome of the randomization procedure. Twenty percent of respondents instructed to draw colored chips from a box in an RRT survey indicated they believed the interviewer knew which chip they would draw — making the RRT pointless for those respondents (Wiseman, Moriarty, and Schafer 1975). A similar issue arises with virtual randomizing devices in online surveys whose outcome might be suspected of being traceable. Undetectability, furthermore, might be questioned when respondents’ answers to ‘randomizing questions’ are used in place of a physical randomizing device. A randomizing question, that is, a question which serves as randomizing device, may be asked if the distribution of a particular attribute in the surveyed population is known. For instance, the number of persons whose birthday falls on a particular month of the year (“If your birthday is between January and March, please answer the following question: … If your birthday is between April and December, please answer the following question: …”). However, responses to randomizing questions of that type are still detectable in principle if they refer to respondents themselves or their relatives and, thus, raise suspicion.
Availability, finally, means that the randomizing device should be within respondents’ reach during the survey. Availability is guaranteed if an interviewer is present to hand over the randomizing device or if the randomizing device is sent out together with a paper-and-pencil questionnaire. In online and telephone surveys, however, the use of a physical randomizing device is almost always problematic. Dice or cards, for instance, are rarely within respondents’ reach. Sending these devices to respondents in advance works in some situations (see Jong, Pieters, and Fox 2010 for an example). Yet, it is costly and still does not guarantee that respondents have the device actually at hand when they answer the survey. The same holds for more common devices such as coins or banknotes. Even though they are available to all respondents in principle, having to get up from the computer to get one’s wallet leads some respondents to skip the randomization procedure. The only safe strategy for self-administered and telephone surveys regarding availability is — in our view — to avoid any physical randomizing device and to use what we call a “randomizing question”.

Questions on birthdays or other known demographics have been used frequently as non-sensitive questions in the unrelated-question RRT design (Horvitz, Shah, and Simmons 1967). But they have been rarely used as randomizing device for the first step in the RRT procedure to determine whether the sensitive or the non-sensitive question has to be answered. In one of the few early RRT studies that applied such a randomizing question, Brown (1975, as cited in Fox and Tracy 1986, p. 61f) used a demographic question on respondents’ mothers’ dates of birth in order to determine whether a sensitive or a non-sensitive question had to be subsequently answered. Besides the apparent advantage of availability in all survey situations the use of a randomizing question comprises also some caveats. Detectability has already been mentioned. In addition, it is usually difficult to find one or more suitable randomizing questions as the set of possible questions with known response distribution in the surveyed population is usually very restricted.

### 3.2.3 Respondents’ objective and subjectively perceived privacy protection

The core rationale underlying the RRT is that respondents understand that their answers remain totally concealed and that thus admitting sensitive behavior bears no risk at all. Respondents’ privacy protection is supposed to lead to more truthful answers and hence to an increase in data validity. Because the deterministic link between individual survey response and admittance of a sensitive behavior is broken by introducing randomness to the answering process, respondents’ protection is guaranteed in all RRT designs. Nonetheless, a probabilistic link between individual response and sensitive behavior remains. The strength of the
probabilistic link depends on the particular RRT design and on the true prevalence of the sensitive behavior under question. The researcher directly influences it by defining the RRT design parameter \( p \), the probability with which respondents have to answer the sensitive question. A higher \( p \) increases the correlation between the individual response and the admittance of sensitive behavior. As a consequence, “respondents’ jeopardy” (Fox and Tracy 1986, 32), defined as \( P(\text{sens. behavior} | \text{‘yes’ answer}) \), the probability that a respondent giving a “yes” response actually admitted the sensitive behavior under question, increases.\(^1\)

However, the choice of \( p \) not only influences respondents’ jeopardy or — conversely — respondents’ privacy but also the variance of the RRT estimator as shown in the preceding section. From this fact originates the researcher’s dilemma in choosing an appropriate \( p \) for an RRT design. On the one hand, \( p \) should be low in order to provide a high level of privacy protection to respondents; on the other hand, \( p \) should be as high as possible in order to obtain an efficient estimator (see Lenzvelt-Mulders, Hox, and van der Heijden 2005 for statistical implications of the choice of RRT design parameters).

Yet, as Moriarty and Wiseman (1976) already pointed out, it is essential to distinguish between the objective \( p \) of an RRT design, and \( p \) and the privacy protection as perceived by respondents. Only the latter affects respondents’ trust as well as compliance and, as a consequence, the validity of measurements obtained through the RRT. Even though a correlation between the objective value of \( p \) and respondents’ perceived privacy protection may be expected, there is virtually no knowledge about this empirical relation. Studies on the effect of different values of \( p \) on respondents’ trust in the RRT, on perceived privacy protection and on data validity are almost nonexistent and the RRT literature gives no empirically grounded advice on which \( p \) to choose. A study of Soeken and Macready (1982) is the only exception known to us. They found a slight decrease of respondents’ perceived privacy protection with increasing \( p \) and a statistically significantly lower perceived protection for \( p = .91 \) compared to values of \( p \leq .84 \).

### 3.3 Benford RRT: A new randomizing device using Benford’s law

In this section we present Benford RRT, a new randomizing device (originally suggested in Diekmann 2012), which fulfills the stated requirements of a good RRT randomizing device. At the core of Benford RRT lies a randomizing ques-

---

\(^1\) There are several other, more sophisticated measures of privacy protection for RRT designs suggested in the literature (for some early works see Lanke 1975; Greenberg et al. 1977).
tion on the first digit of an acquaintance’s address house number. First digits of house numbers follow, as we will show in the next section, a known distribution, namely the Benford distribution. This fact can be used to obtain a suitable randomizing device that is applicable in all circumstances. Furthermore, we show how Benford RRT increases the efficiency of the RRT estimator by exploiting the divergence between respondents’ objective privacy protection and their subjectively perceived privacy protection. This divergence is particularly high in the case of Benford RRT due to the “Benford Illusion”, the substantial misperception of the frequency of Benford distributed numbers.

3.3.1 Benford’s law of first digits

First digits of many real-life data follow a particular distribution with low digits (i.e., “1”) occurring more often than larger digits (i.e., “9”). This fact has been discovered and the distribution formalized by Newcomb (1881) and later Benford (1938). It is nowadays widely known as Benford’s law. Benford’s law states that the probabilities of first digits $d = 1, 2, \ldots, 9$ are

$$P(d) = \log_{10}(1 + 1/d)$$

First digits of the population of countries, the size of lakes, numbers in tax declarations or in newspaper articles, and many other data have all been shown to follow this distribution (e.g., Diekmann 2012).

In principle, all of these data sources could be used as a randomizing device for Benford RRT. The empirical fit to the Newcomb-Benford distribution should in any case be carefully tested, as the preconditions which produce Benford-distributed first digits might not be fulfilled. For instance, the first digits of numbers in the Bible do follow a Benford distribution, with the exception of the digit 7, which is overrepresented (Hünergähler 2007). Benford (1938) already hypothesized that first digits of house numbers follow a Benford distribution and found supporting evidence using the American Men of Science directory. Diekmann (2012) examined the same, using the Swiss telephone directory. Figure 3.1 shows that the empirical distribution of first digits of house numbers of Swiss addresses almost perfectly fits the Benford distribution, hence, obeys Benford’s law. In a subsequent test, respondents of a general population survey were asked to indicate the house number of an acquaintance. The distribution of the first digits of house numbers generated through this process were in line with the theoretical Benford distribution (Diekmann 2012). This gives empirical support to the assumption that first digits of house numbers of acquaintances generated by survey respondents approximately follow the Benford distribution.
Figure 3.1: Comparison of the empirical distribution of first digits of house numbers from the Swiss phone directory (TwixTel34, \( N \approx 3 \) million) with the Benford distribution. Numbers compiled by Stefan Wehrli.

### 3.3.2 Implementing Benford RRT

Benford RRT uses a question on the first digit of the house number of an acquaintance’s address as randomizing question. It can be implemented as follows (see also figure 3.2):

> Please think of an acquaintance of yours whose address you know. Now take the first digit of this person’s house number.

If this digit is 1 to 5, please answer the following question: Have you ever cheated on your taxes?

If this digit is 6 to 9, please answer this question: Is your mother’s birthday in the months of January through June?

In this example of an unrelated-question RRT design the first digit of the house number determines whether a respondent subsequently has to answer a sensitive or a non-sensitive question. \( p \) is defined as .78 by choosing the range of digits 1, 2, 3, 4, 5 leading to the sensitive question and 6, 7, 8, 9 to the non-sensitive question; but, of course, other values are possible.
3.3. Benford RRT: A new randomizing device using Benford’s law

As first digits of house numbers follow the Benford distribution, a question on the first digit of a randomly chosen address’s house number becomes a naturally occurring randomizing device with a known outcome distribution without the need for any physical artifact such as dice or coins. This makes it suitable for any survey situation.

### 3.3.3 The “Benford Illusion”

The use of a Benford question as randomizing device for the RRT bears the additional advantage that respondents usually underestimate the probability of the occurrence of Benford-distributed low digits because they typically assume a rather uniform distribution. Survey respondents, when explicitly asked, substantially underestimated the probability of the occurrence of the first digits 1 to 4 in house numbers by .09 with a subjective estimate of 0.61 (N=295, Diekmann, 2012). By making use of that misperception – the Benford Illusion – the trade-off between statistical efficiency and respondents’ perceived privacy protection in RRT designs is relaxed. A higher probability $p$ that respondents are instructed to answer the sensitive question may be chosen without provoking respondents’ privacy concerns because respondents’ subjectively perceived $p$ is substantially lower than the objective $p$.

The idea that a good randomizing device for the RRT should bear the property, that respondents perceive $p$ smaller than the objective $p$, was originally brought up by Moriarty and Wiseman (1976). They investigated respondents’ perception of $p$ for different randomizing devices and found that using two dice had the desired property. Respondents heavily underestimated the outcome probability of a throw of two dice being 4 to 10 by .13 with a median perceived probability of .70. A misperception bias that is similar in magnitude to the one of Benford
RRT. In this sense, Benford RRT can be seen as a substitute for the throw of two dice in interview situations where no interviewer is present to provide respondents with dice.

### 3.4 An application in a survey on student cheating

#### 3.4.1 Data and design

We implemented Benford RRT in an online student survey on exam cheating and plagiarism at two major Swiss universities (Hüglinger, Jann, and Diekmann 2014a). All students enrolled at the two institutions were contacted via their official university e-mail address in spring 2011. Out of a total of 19,410 students 6,494 finished the survey, resulting in a response rate of 33 percent (RR1, AAPOR 2011). Two hundred and one respondents who partially completed, i.e., reached the part of the questionnaire with the sensitive questions, are also included in the following analyses. Respondents who had neither sat an exam nor submitted a paper (386), or who had poor German language skills (230), as well as 67 respondents with incomplete data have been excluded, leaving us with a sample of 6,012 observations. The subsequent analyses are, furthermore, restricted to 1,001 respondents who were surveyed in direct questioning mode and the 994 surveyed using Benford RRT.

Survey respondents were asked five sensitive questions about their own cheating behavior, using either direct questioning, Benford RRT, or one of four other RRT variants, which will not be discussed here. Assignment to one of these sensitive question techniques was randomized. The wording of the sensitive questions was identical in all conditions. Benford RRT was implemented in an unrelated-question RRT design as presented in the preceding section. Half of the respondents were directed to the sensitive question with probability \( p = .70 \), the other half with \( p = .78 \). This allowed the investigation of whether a different \( p \) has any effect on respondents’ admittance of sensitive behavior or on their perceived privacy protection. The unrelated non-sensitive questions consisted of five questions on respondents’ mothers’ dates of birth, with answer distributions of
$P(\text{yes}|\text{sens. quest.}) = .5$ (see footnote\textsuperscript{2} for the question wording). Their order was randomized to offset any effects of a particular unrelated question.

### 3.4.2 Results

In order to evaluate Benford RRT, in the following section we compare prevalence estimates of respondents' admittance of sensitive behavior resulting from Benford RRT and from direct questioning (DQ). Assuming that respondents only falsely deny but never falsely admit a sensitive behavior, higher prevalence estimates are interpreted as a result of more respondents answering truthfully. According to this "more-is-better assumption", which is the basis of all comparative RRT studies (e.g., Lensvelt-Mulders et al. 2005), higher prevalence estimates of one method indicate its superior validity. Due to the experimental design, i.e., the fact that respondents were randomly assigned to either Benford RRT or direct questioning, differences in prevalence estimates can be interpreted as causal effects of the particular sensitive question technique. RRT point estimates and standard errors are calculated using a generalization of the formulae from the first section to the case where different values of $p$ and $P(\text{yes}|\text{sens. quest.})$ for subgroups of respondents are used. The procedure is implemented in the Stata ado program rrreg (Jann 2008).

Figure 10.3 presents comparisons of prevalence estimates for the five surveyed sensitive cheating behaviors between direct questioning (DQ) and Benford RRT (see also table 3.A.1 in the appendix). In the left panel, prevalence point estimates with 95% confidence intervals, specified by the lines on both sides of the point estimates, are depicted. Estimates range from 17.8 percentage of students admitting having copied in an exam to 1.5 percentages of students admitting partial paper plagiarism. Clearly discernible is the pattern of estimates resulting from Benford RRT being higher than the corresponding DQ estimates except for the first item, "copy in exam", where the Benford RRT estimate is marginally lower by .6 [-5.0; 3.9] percentage points. Note that confidence intervals for Benford RRT estimates are considerably larger than for DQ, which is due to the RRT's inherent lower statistical efficiency.

Differences between Benford RRT and direct questioning (DQ) estimates are portrayed in the right panel of figure 3.3. If confidence intervals do not include the

---

\textsuperscript{2} The wording of the unrelated questions was as follows (translated from German):

- Is your mother's birthday in the months of January through June?
- Is your mother's birthday in the first half of the month? (from 1st to 15th)
- Is your mother's birthday on an even-numbered day? (2nd, 4th, 6th, etc. of the month)
- Is your mother's birth year even-numbered? (Please, consider 0 as an even number.)
zero line, prevalence estimates between Benford RRT and DQ differ significantly at the 95% level. Results show a significant difference only for one out of the five sensitive items, namely "partial plagiarism", where the Benford RRT estimate is 4.9 [95% confidence bounds: .8; 9.0] percentage points higher than the DQ estimate. For the item "notes in exam" the Benford RRT estimate is 3.8 [-.2; 7.8] percentage points higher than the DQ estimate; but with a p-value of .06 the difference barely misses conventional significance level.

<table>
<thead>
<tr>
<th>Prevalence Estimates of Cheating</th>
<th>Difference Benford RRT – DQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>(in % of Students)</td>
<td></td>
</tr>
<tr>
<td>Copy in exam</td>
<td></td>
</tr>
<tr>
<td>DQ 17.8</td>
<td>6</td>
</tr>
<tr>
<td>Benford RRT 17.2</td>
<td></td>
</tr>
<tr>
<td>Notes in exam</td>
<td></td>
</tr>
<tr>
<td>DQ 9.1</td>
<td>3</td>
</tr>
<tr>
<td>Benford RRT 12.2</td>
<td></td>
</tr>
<tr>
<td>Drugs for exam</td>
<td></td>
</tr>
<tr>
<td>DQ 4.5</td>
<td>1</td>
</tr>
<tr>
<td>Benford RRT 11.4</td>
<td></td>
</tr>
<tr>
<td>Partial plagiarism</td>
<td></td>
</tr>
<tr>
<td>DQ 8.9</td>
<td>4</td>
</tr>
<tr>
<td>Benford RRT 7.8</td>
<td></td>
</tr>
<tr>
<td>Severe plagiarism</td>
<td></td>
</tr>
<tr>
<td>DQ 1.5</td>
<td></td>
</tr>
<tr>
<td>Benford RRT 2.4</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.3: Comparison of prevalence estimates of cheating between direct questioning (DQ) and Benford RRT. Lines indicate 95% CIs.\(^4\) \(N\) varies between the different items because questions on cheating in exams have been asked only of respondents who sat in at least one exam; questions on plagiarism only of respondents who have handed in a paper.

\(^4\) Wording of the sensitive questions (translated from German):

- Copy in exam: "In your studies, have you ever copied from other students during an exam?"
- Notes in exam: "In your studies, have you ever used illicit crib notes in an exam (including notes on mobile phones, calculators, or similar)?"
- Drugs for exam: "In your studies, have you ever used prescription drugs to enhance your performance in an exam?"
- Partial plagiarism: "In your studies, have you ever handed in a paper containing a passage deliberately taken from someone else's work without citing the original?"
- Severe plagiarism: "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?"
3.5. Conclusions

Further analysis showed that the survey break-off rate for Benford RRT was almost twice as high as for direct questioning but remained with 2.2\% of respondents within an acceptable level. Considering that answering the sensitive questions took respondents 175 seconds with Benford RRT and only 53 seconds with DQ, this is no surprise. Respondents’ self-stated trust in the survey’s anonymity and privacy protection measures was lower for Benford RRT (73\% do trust) than for DQ (81\%).\textsuperscript{5} The RRT procedure seems, in a first step, to intensify privacy concerns among respondents. However, the risk of disclosure, i.e., the risk that any respondents’ cheating behavior will be exposed because of the survey, is considered lower in the case of Benford RRT (79\% see no risk) compared to DQ (71\%).\textsuperscript{6} See table 3.A.2 in the appendix for detailed results.

Finally, we compared prevalence estimates and respondents’ perceived privacy protection for Benford RRT designs with different levels of privacy protection, i.e., with different values of $p$, the probability with which respondents are instructed to answer the sensitive question. Using $p = .70$ and $p = .78$, results showed no significant differences in prevalence estimates and no discernible pattern of one RRT design performing systematically differently from the other (see detailed results in figure 3.A.1 and table 3.A.3 as well as 3.A.4 in the appendix). Furthermore, respondents’ assessment of anonymity and privacy protection as well as risk of disclosure did not differ between the two conditions. Choosing $p = .78$ instead of $p = .70$ had clearly no effect on prevalence estimates or respondents’ perception of privacy. Yet, the choice of $p$ affects statistical efficiency. Therefore, $p = .78$ is the preferred choice for an implementation of the Benford RRT. Possibly, even a higher $p$ than $p = .78$ could be chosen without affecting respondents’ privacy and data validity.

3.5 Conclusions

In this chapter we have introduced Benford RRT, a new randomizing device for the RRT based on Benford’s law, and we have presented results of an empirical evaluation of the method. The new randomizing device uses a randomizing

\textsuperscript{5} Wording of the question: “Please be honest. How much do you trust our measures for guaranteeing survey participants’ anonymity and privacy protection?” Response categories “very much” and “quite a bit” have been coded as respondent does trust; response categories “partly”, “rather not” and “not at all” as respondent does not trust.

\textsuperscript{6} Wording of the question: “How likely do you deem it possible, that it can be traced back whether a particular respondent has carried out one of the surveyed sensitive behaviors (copying in exam, crib notes, plagiarism, etc.)?” Response categories “impossible” and “very unlikely” have been coded as respondent sees no risk; response categories “rather unlikely”, “rather likely” and “very likely” as respondent sees a risk.
question and does not need any physical artifact. Therefore, it is particularly suitable for self-administered surveys and telephone surveys. In addition, it allows for increasing the statistical efficiency of the RRT, without jeopardizing respondents’ perceived privacy protection, by taking advantage of the Benford Illusion, namely, respondents’ misperception of Benford-distributed first digits.

Benford RRT performed well in our online survey on student cheating behavior. In one out of five items it generated a higher, and under the more-is-better assumption a more valid, estimate of sensitive behavior than direct questioning. A second item estimate was substantially higher but with $p = .06$ missed conventional significance level. No Benford RRT estimate was substantially lower than the DQ estimates, and all Benford RRT estimates were positive and meaningful. In contrast to other RRT online implementations (see for instance Coutts et al. 2011; Jann, Jerke, and Krumpal 2012; Coutts and Jann 2011; Peeters 2005), the problem of severely negatively biased or even negative estimates did not arise in our implementation of Benford RRT. It should be noted, however, that a new RRT variant, the Crosswise Method (Yu, Tian, and Tang 2008), which was also implemented in our study, performed even better than Benford RRT and seems to be another well performing, promising method to survey sensitive questions (see Höglinger, Jann, and Dickmann 2014b).

Results also showed that an increase of the probability $p$ with which respondents are instructed to answer the sensitive question by $.08$ to $p = .78$ had no effect on estimates nor on respondents’ perceived privacy protection. Thus, it is safe to choose $p$ as high as $p = .78$ when implementing Benford RRT. Future studies should address in more detail how far $p$ can be increased without endangering data validity. It remains unclear, though, whether a de- or increase of $p$ within a reasonable range has no effect on respondents perceived privacy protection in other RRT designs or whether this is somehow related to the Benford Illusion, a special property of Benford RRT. Results suggest, in any case, that respondents’ perception of privacy protection is mainly driven by other design considerations than the mere choice of $p$.

Whether the increase in more truthful answers achieved through Benford RRT justifies the additional burden put on respondents and the need for bigger sample sizes in order to compensate for the RRT’s lower statistical efficiency depends on two things: the sensitivity of the topic surveyed and whether a sizeable respondents’ sample is actually available. If an implementation of the RRT is considered, however, Benford RRT seems to be a well-performing RRT variant that is easily implemented not only, but particularly, in survey situations where no interviewer is present.
3.A Appendix

Table 3.A.1: Comparison of prevalence estimates of cheating between direct questioning (DQ) and Benford RRT.

<table>
<thead>
<tr>
<th>Item</th>
<th>Sensitive Question Condition</th>
<th>N</th>
<th>Point Estimate</th>
<th>SE</th>
<th>95% Confidence Interval</th>
<th>p-Value Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy</td>
<td>DQ</td>
<td>978</td>
<td>17.8</td>
<td>1.2</td>
<td>15.4 - 20.2</td>
<td>0.796</td>
</tr>
<tr>
<td>in exam</td>
<td>Benford RRT</td>
<td>974</td>
<td>17.2</td>
<td>1.9</td>
<td>13.5 - 21.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>RRT - DQ</em></td>
<td></td>
<td>-0.6</td>
<td>2.3</td>
<td>-5.0 - 3.9</td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td>DQ</td>
<td>978</td>
<td>9.1</td>
<td>0.9</td>
<td>7.3 - 10.9</td>
<td></td>
</tr>
<tr>
<td>in exam</td>
<td>Benford RRT</td>
<td>970</td>
<td>12.9</td>
<td>1.8</td>
<td>9.3 - 16.5</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td><em>RRT - DQ</em></td>
<td></td>
<td>3.8</td>
<td>2.0</td>
<td>-0.2 - 7.8</td>
<td></td>
</tr>
<tr>
<td>Drugs</td>
<td>DQ</td>
<td>975</td>
<td>3.4</td>
<td>0.6</td>
<td>2.2 - 4.5</td>
<td></td>
</tr>
<tr>
<td>for exam</td>
<td>Benford RRT</td>
<td>964</td>
<td>4.5</td>
<td>1.6</td>
<td>1.3 - 7.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>RRT - DQ</em></td>
<td></td>
<td>1.1</td>
<td>1.7</td>
<td>-2.3 - 4.5</td>
<td>0.532</td>
</tr>
<tr>
<td>Partial</td>
<td>DQ</td>
<td>722</td>
<td>2.9</td>
<td>0.6</td>
<td>1.7 - 4.1</td>
<td></td>
</tr>
<tr>
<td>plagiarism</td>
<td>Benford RRT</td>
<td>718</td>
<td>7.8</td>
<td>2.0</td>
<td>3.9 - 11.7</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td><em>RRT - DQ</em></td>
<td></td>
<td>4.9</td>
<td>2.1</td>
<td>0.8 - 9.0</td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>DQ</td>
<td>724</td>
<td>1.5</td>
<td>0.5</td>
<td>0.6 - 2.4</td>
<td></td>
</tr>
<tr>
<td>plagiarism</td>
<td>Benford RRT</td>
<td>717</td>
<td>2.4</td>
<td>1.8</td>
<td>-1.2 - 5.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>RRT - DQ</em></td>
<td></td>
<td>0.8</td>
<td>1.9</td>
<td>-2.8 - 4.5</td>
<td>0.651</td>
</tr>
</tbody>
</table>

Table 3.A.2: Comparison of break-off rate, response time and respondents’ evaluation of the sensitive question technique between direct questioning (DQ) and Benford RRT. SE in parentheses.

<table>
<thead>
<tr>
<th>Sensitive Question Condition</th>
<th>Break-off</th>
<th>Response Time</th>
<th>Trust in Protection</th>
<th>No Disclosure Risk</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ</td>
<td>1.2</td>
<td>53</td>
<td>80.7</td>
<td>71.1</td>
<td>1001</td>
</tr>
<tr>
<td>(0.3)</td>
<td>(1.5)</td>
<td>(1.3)</td>
<td>(1.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benford RRT</td>
<td>2.2</td>
<td>175</td>
<td>73.3</td>
<td>79.2</td>
<td>994</td>
</tr>
<tr>
<td>(0.5)</td>
<td>(2.2)</td>
<td>(1.4)</td>
<td>(1.3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Break-off: % who did not complete survey after reaching the sensitive questions.
Response Time: Av. time (seconds) to answer the sensitive questions (highest 2.5 percentiles excluded). Trust in Protection: % who trust in anonymity and privacy protection measures. No Disclosure Risk: % who think there is no disclosure risk.
### Prevalence Estimates of Cheating (in % of Students)

<table>
<thead>
<tr>
<th>Cheating Type</th>
<th>p = .70</th>
<th>p = .76</th>
<th>Difference [p = .76] - [p = .70]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy in exam</td>
<td>4.3</td>
<td>9.2</td>
<td>5.0 (N=880)</td>
</tr>
<tr>
<td>Notes in exam</td>
<td>3.9</td>
<td>7.9</td>
<td>5.0 (N=484)</td>
</tr>
<tr>
<td>Drugs for exam</td>
<td>3.9</td>
<td>5.1</td>
<td>5.0 (N=484)</td>
</tr>
<tr>
<td>Partial plagiarism</td>
<td>3.5</td>
<td>6.8</td>
<td>4.0 (N=382)</td>
</tr>
<tr>
<td>Severe plagiarism</td>
<td>3.5</td>
<td>1.7</td>
<td>6.0 (N=304)</td>
</tr>
</tbody>
</table>

Figure 3.A.1: Comparison of Benford RRT prevalence estimates of cheating between designs with differing probability p with which respondents are instructed to answer the sensitive question. Lines indicate 95%-CI.
Table 3.A.3: Comparison of Benford RRT prevalence estimates of cheating between designs with differing probability $p$ with which respondents are instructed to answer the sensitive question.

<table>
<thead>
<tr>
<th>Item</th>
<th>RRT Parameter $p$</th>
<th>N</th>
<th>Point Estimate</th>
<th>SE</th>
<th>95% Confidence Interval</th>
<th>p-Value Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy in exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = .70$</td>
<td>490</td>
<td>14.7</td>
<td>2.8</td>
<td>9.2</td>
<td>20.2</td>
<td></td>
</tr>
<tr>
<td>$p = .78$</td>
<td>484</td>
<td>19.7</td>
<td>2.6</td>
<td>14.7</td>
<td>24.8</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>5.0</td>
<td>3.8</td>
<td>-2.5</td>
<td>12.5</td>
<td></td>
<td>0.188</td>
</tr>
<tr>
<td>Notes in exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = .70$</td>
<td>487</td>
<td>13.4</td>
<td>2.8</td>
<td>7.9</td>
<td>18.9</td>
<td></td>
</tr>
<tr>
<td>$p = .78$</td>
<td>483</td>
<td>12.4</td>
<td>2.4</td>
<td>7.7</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>-1.1</td>
<td>3.7</td>
<td>-8.2</td>
<td>6.1</td>
<td></td>
<td>0.771</td>
</tr>
<tr>
<td>Drugs for exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = .70$</td>
<td>484</td>
<td>3.9</td>
<td>2.5</td>
<td>-1.0</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td>$p = .78$</td>
<td>480</td>
<td>5.1</td>
<td>2.1</td>
<td>0.9</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>1.2</td>
<td>3.3</td>
<td>-5.2</td>
<td>7.6</td>
<td></td>
<td>0.717</td>
</tr>
<tr>
<td>Partial plagiarism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = .70$</td>
<td>362</td>
<td>9.3</td>
<td>3.1</td>
<td>3.2</td>
<td>15.4</td>
<td></td>
</tr>
<tr>
<td>$p = .78$</td>
<td>356</td>
<td>6.3</td>
<td>2.5</td>
<td>1.4</td>
<td>11.3</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>-2.9</td>
<td>4.0</td>
<td>-10.8</td>
<td>4.9</td>
<td></td>
<td>0.458</td>
</tr>
<tr>
<td>Severe plagiarism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = .70$</td>
<td>361</td>
<td>3.5</td>
<td>2.9</td>
<td>-2.2</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>$p = .78$</td>
<td>356</td>
<td>1.2</td>
<td>2.2</td>
<td>-3.1</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>-2.2</td>
<td>3.6</td>
<td>-9.3</td>
<td>4.9</td>
<td></td>
<td>0.538</td>
</tr>
</tbody>
</table>

Table 3.A.4: Comparison of break-off rate, response time and respondents’ evaluation of the sensitive question technique with differing probability $p$ with which respondents are instructed to answer the sensitive question. SE in parentheses.

<table>
<thead>
<tr>
<th>RRT Parameter $p$</th>
<th>Break-off Response Time</th>
<th>Trust in Protection</th>
<th>No Disclosure Risk</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = .70$</td>
<td>2.4</td>
<td>(0.7)</td>
<td>(3.4)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>$p = .78$</td>
<td>2.0</td>
<td>(0.6)</td>
<td>(2.7)</td>
<td>(2.0)</td>
</tr>
</tbody>
</table>

Notes: Break-off: % who did not complete survey after reaching the sensitive questions. Response Time: Av. time (seconds) to answer the sensitive questions (highest 2.5 percentiles excluded). Trust in Protection: % who trust in anonymity and privacy protection measures. No Disclosure Risk: % who think there is no disclosure risk.
Chapter 4

More Is Not Always Better: An Experimental Individual-Level Validation of the RRT and the Crosswise Model

Abstract Social desirability and the fear of sanctions can deter survey respondents from responding truthfully to sensitive questions. Self-reports on norm breaking behavior such as shoplifting, non-voting, or tax evasion may therefore be subject to considerable misreporting. To mitigate such misreporting, various indirect techniques for asking sensitive questions, such as the randomized response technique (RRT), have been proposed in the literature. In our study, we evaluate the viability of several variants of the RRT, including the recently proposed crosswise-model RRT, by comparing respondents' self-reports on cheating in dice games to actual cheating behavior, thereby distinguishing between false negatives (underreporting) and false positives (overreporting). The study has been implemented as an online survey on Amazon Mechanical Turk (N = 6,505). Our results indicate that the forced-response RRT and the unrelated-question RRT, as implemented in our survey, fail to reduce the level of misreporting compared to conventional direct questioning. For the crosswise-model RRT, we do observe a reduction of false negatives (that is, an increase in the proportion of cheaters who admit having cheated). At the same time, however, there is an increase in false positives (that is, an increase in non-cheaters who falsely admit having cheated). Overall, our findings suggest that none of the implemented sensitive questions techniques substantially outperforms direct questioning. Furthermore, our study demonstrates the importance of distinguishing false negatives and false positives when evaluating the validity of sensitive question techniques.

4.1 Introduction

Surveying sensitive topics such as deviant behavior, stigmatizing traits, or controversial attitudes poses serious challenges to survey research. First, respondents’ data need to be carefully protected, particularly for sensitive themes like illegal behavior or politically repressed opinions. Second, even with good data protection, respondents might be tempted to misreport on sensitive questions or refuse to answer, for example, due to embarrassment or due to fear of negative sanctions (Tourangeau and Yan 2007). To avoid biased or incomplete measurement, survey researchers therefore have to find questioning procedures that maximize respondents’ willingness to provide truthful answers.

Various approaches to address this issue have been pursued in previous research, but the results on the success and the failure of the different questioning strategies appear inconsistent and highly dependent on implementation details, the research question, or the studied population (Krumpal and Näher 2012). Most promising results can be found with respect to survey mode and, in particular, to whether an interviewer is present or not. For example, Kreuter, Presser, and Tourangeau (2008) compared CATI (Computer Assisted Telephone Interviewing), IVR (Interactive Voice Response), and online mode in a study on poor (and potentially embarrassing) academic performance among university alumni, where the respondents’ answers could be validated against the university’s grade records. The level of misreporting (false denial of poor performance) was highest in CATI mode, where an interviewer was present. However, also in the more anonymous IVR and online modes, misreporting remained high.

4.1.1 The randomized response technique

Other approaches try to mitigate misreporting and non-response by employing so-called indirect questioning techniques, one of which is the randomized response technique (RRT; originated by Warner 1965). The basic idea of the RRT is to protect respondents through random misclassification so that a given answer does not reveal the true answer to the sensitive question. Ideally the anonymity induced by the misclassification makes respondents more comfortable providing truthful answers. For example, in the forced-response variant of the RRT (Boruch 1971) a randomizing device such as a coin flip determines whether a respondent is instructed to provide a truthful answer to a sensitive yes/no question or simply

We thank Andreas Diekmann for his support and advice, Philip Tscheimer for his help with programming the survey, Debra Hevenstone for language editing, and Stefan Wehali and the ETH Decision Science Laboratory (DeScIL) for posting the survey on Amazon Mechanical Turk.
respond with "yes" (or "no"), irrespective of the true answer. Therefore, as long as only the respondent knows the outcome of the randomizing device, a given answer does not reveal the true answer to the sensitive question; the given answer could also just be a surrogate response due to the randomizing device.

Despite the theoretical appeal of the RRT, it remains questionable whether respondents understand the procedure, trust that their anonymity is protected, and are more inclined to provide a truthful answer (when instructed to do so). Furthermore, due to lack of understanding, respondents might fail to comply with the RRT instructions even if they are asked to provide an answer that is unrelated to the sensitive question (Edgell, Himmelfarb, and Duchan 1982; Edgell, Duchan, and Himmelfarb 1992; Böckenholt, Barlas, and van der Heijden 2009). A meta-analysis by Lensvelt-Mulders et al. (2005), mostly covering face-to-face and paper-and-pencil RRT studies published between 1965 and 2000, concludes that, on average, the RRT yields more valid results than direct questioning, but the variability in results is high. Furthermore, findings from a number of newer studies on the application of the RRT in online mode are not very promising (Coukts and Jann 2011; Höglinger, Jann, and Diekmann 2014b; Holbrook and Kroscnick 2010; Ostapczuk and Musch 2011; Peeters 2005).

### 4.1.2 The crosswise-model RRT

Recently, a variant of the RRT, the "crosswise model," proposed by Yu, Tian, and Tang (2008), has received growing attention. Several studies report that the crosswise-model RRT consistently produces higher prevalence estimates of sensitive behaviors than direct questioning (Corbacho et al. 2016; Hoffmann et al. 2015; Hoffmann and Musch 2015; Höglinger, Jann, and Diekmann 2014b; Jann, Jerke, and Krumpal 2012; Korndörfer, Krumpal, and Schmukle 2014; Kundt 2014; Kundt, Misch, and Nerré 2014; Shamsipour et al. 2014). The crosswise-model RRT works by presenting two yes/no questions to the respondent, a sensitive question and an unrelated non-sensitive question, and then asking whether the answers to both questions are the same (both "yes" or both "no") or whether the two answers are different (one "yes," one "no"). The advantages of the crosswise-model RRT over alternative RRT variants, it is argued, are that the instructions are easy to understand, the response options are obviously ambiguous with respect to the sensitive question (i.e. there is no clear self-protective answering strategy), and no respondents are forced to give "false" answers.
4.1.3 Validation of sensitive question techniques

As mentioned above, results from studies evaluating indirect questioning techniques are often inconclusive. One reason for the variability in the findings is that the studies employ different validation strategies. By far the most frequent approach is to use the results from direct questioning as a baseline, to which the results from one or several indirect questioning techniques are compared. We use the term *comparative validation study* to refer to studies employing such an approach. The argument is that if the question is sensitive, respondents will tend to underreport when asked to answer the question directly. An indirect questioning technique that successfully reduces underreporting should therefore yield higher estimates than direct questioning (likewise, if the problem is over-reporting, such as in questions on voter turnout, a successful indirect technique should yield lower estimates than direct questioning). Hence, comparative validation studies rely on the so-called more-is-better (less-is-better) assumption (Lensvelt-Mulders et al. 2005); an indirect questioning technique is considered more valid if it produces higher (lower) prevalence estimates than direct questioning. More generally, if comparing multiple indirect techniques, the technique producing the highest (lowest) estimate is judged to be the most valid.

The more-is-better assumption is often legitimate. In many cases it is reasonable to assume that respondents avoid socially undesirable answers and thus underreport on sensitive questions. However, sometimes, social desirability might differ between subpopulations, a well-known example being the number of sexual partners as reported by men and women (Smith 1992; Tourangeau and Smith 1996). Therefore, the more-is-better assumption can sometimes be challenged on the ground that social desirability bias points in different directions depending on the subpopulation. Furthermore, even if the more-is-better assumption is justified, a higher estimate from an indirect questioning technique does not necessarily imply that the technique produces more valid measurements than direct questioning. If it is true that direct questioning yields underestimation, then higher estimates by an indirect technique is a necessary condition, but not a sufficient condition. The more-is-better assumption assumes that the increase in estimates is due to more truthful answers. However, given the complexity of the instructions of most RRT implementations it may also be simply due to the respondents’ inability to correctly apply the procedure. That is, the increase in estimates might be due to non-compliance with the RRT instructions (e.g., due to problems with the randomizing device, misunderstanding of instructions, or unwillingness to follow the instructions) rather than more truthful answering. Overall, we conclude that comparative validation studies can only provide weak support for the validity of sensitive question techniques (for similar arguments see: Lensvelt-Mulders et
At least some of the shortcomings of comparative validation studies can be overcome by what we call aggregate-level validation studies. In such studies, the true population prevalence of the sensitive trait or behavior is known from an external and reliable source or can be determined based on theoretical reasoning. For example, in studies of voter turnout, true aggregate turnout is known from administrative records (for recent examples see Rosenfeld, Imai, and Shapiro 2015 and Moshagen, Musch, and Erdfelder 2012). If the true value is known, then overestimation and underestimation by different questioning techniques can be observed directly without having to resort to direct questioning as a baseline, which is a clear improvement over comparative validation studies. Yet, also such aggregate-level validation studies might be inconclusive. First, true values might differ from the assumed value, perhaps because the study focuses on a special subpopulation or because there is sample selection bias (e.g., due to nonresponse). Second, and more importantly, a close match between the prevalence estimate from a particular sensitive questioning technique and the true value does not necessarily imply that the technique produces valid measurements at the individual level. As argued above, different mechanisms might affect the prevalence estimate, not all of which are consistent with more truthful answering. In other words, apart from possible sample selection bias, the aggregate-level validation approach rests on the assumption that socially desirable responding is the only misreporting mechanism.

A useful distinction in this context is between false negatives (or true positives) and false positives (or true negatives). The goal of sensitive question techniques is to reduce the number of false negatives, that is, the number of respondents who deny the sensitive question even though it does apply. However, a sensitive question technique might also increase the number of false positives, that is, the number of respondents who agree with the sensitive question even though it does not apply. Comparing overall prevalence estimates from the technique with either direct questioning or a known “true” prevalence, does not allow one to distinguish between a reduction in false negatives and an increase in false positives, both of which will increase the estimated total prevalence. To be able to disentangle the two effects, validation data at the individual level is required. Hence, we argue that individual-level validation studies are necessary to be able to evaluate the degree to which a technique does, in fact, produce valid measurements.

Despite their clear advantage over the comparative approach, individual-level validation studies are very rare. Reviewing RRT studies from over 35 years, Lensvelt-Mulders et al. (2005) counted just six published individual-level vali-
dation studies dealing with sensitive topics such as convictions, arrests, welfare fraud, or failing university courses. We are aware of five additional studies published since (Hoffmann et al. 2015; John et al. 2013; Kirchner 2015; Moshagen et al. 2014; Wolter and Preisendorfer 2013). The available validation studies provide valuable insights, but they do not explicitly focus on disentangling false negatives and false positives. Moreover, some of the studies use a sample that only includes respondents who possess the sensitive trait or engaged in the sensitive activity, so that, by design, only false negatives can be studied. In sum, we believe that additional individual-level validation studies are necessary to disentangle the different response mechanisms and to examine the possibility of false positives in these types of survey techniques. Such studies are the only way to conclusively assess the performance of different sensitive question techniques.

4.1.4 Our study

The goal of our study is to evaluate the validity of different variants of the RRT using a validation design that does not rely on the more-is-better assumption and that allows separate analysis of false negative and false positives. To achieve this we conducted an online survey on Amazon Mechanical Turk, in which the respondents had the opportunity to play one of two dice games. Respondents were given monetary incentives to cheat in these games. After playing the games, respondents were asked about whether they cheated, using direct questioning, forced-response RRT, unrelated-question RRT, or the crosswise-model RRT. For the first game the proportion of cheaters can be estimated based on the laws of chance, for the second game cheating is observable. Comparing the cheating behavior at the aggregate and individual levels with the results from the cheating question reveals the degree to which the questioning techniques are successful in eliciting truthful answers.

4.2 Data and Methods

Study participants were recruited via the online platform Amazon Mechanical Turk (AMT). AMT is an online crowdsourcing marketplace where “requesters” can post tasks (called “Human Intelligence Tasks” or HITs) that can then be completed by “workers” in exchange for money. HITs are announced with a short description of the task and the corresponding payment. AMT is suitable for any task that can be easily outsourced online to an anonymous workforce and is increasingly used to recruit participants for scientific surveys and experiments (Horton, Rand, and Zeckhauser 2011; Mason and Suri 2012; Ipeirotis 2010). On Novem-
ber 5, 2013, we posted a HIT that asked for filling out a scientific survey on “Mood and Personality” for a base payment of $1 and the prospect of winning an additional $2 bonus payment. The HIT was closed on December 5, 2013, when our quotas per experimental condition were fulfilled. Workers who accepted our HIT received an access link to the survey. After having completed the survey, they received payment. Participation was restricted to US residents because one of the sensitive questions was on voting in the US presidential elections. To identify untrustworthy participants, we employed a screening question from Berinsky, Margolis, and Sances (2014), which was passed by 97% of the respondents. The median time required to complete the survey was 6.7 minutes. Details on study and screenshots of the questionnaire are available in the survey documentation (Höglinger and Jann 2015).

A total of 6,505 participants were recruited, of which 6,473 completed the survey at least up to the part containing the sensitive questions. Only the latter are included in our analysis. Furthermore, we exclude 205 participants who did not pass the screening question, 115 participants who did not roll the die in the dice game (or for whom the result of the roll was not recorded due to technical problems), and 1 participant who won in the roll-a-six game but did not claim his legitimate bonus payment.1 The final sample size for our analysis is $N = 6,152$. As displayed in table 4.1, the sample has an even gender distribution and the majority of respondents are under 35 (mean age 32). Respondents are relatively well educated, with 88 percent having attended at least some college. About two thirds are employed or self-employed. A large majority of respondents completed the survey at home and most respondents had extensive experience with “scientific studies such as surveys or experiments on MTurk” (wording from the questionnaire; the median number of previous MTurk studies is 50).

4.2.1 The dice games

Participants were randomly assigned to one of two dice games in which they could win a $2 bonus payment: the prediction game or the roll-a-six game. The games were inspired by Greene and Paxton (2009) and Fischbacher and Föllmi-Heusi (2013) (also see Fischbacher and Heusi 2008 as well as Suri, Goldstein, and Mason 2011). In both games, participants used a digital online die embedded in the questionnaire that could be “rolled” by clicking on a button. Roll outcomes

1 There were 516 winners in the roll-a-six game. The fact that only one of them did not claim the bonus payment indicates that the proportion respondents who falsely deny having won is negligible. To simplify the analysis we exclude this observation and assume the proportion to be zero (also in the prediction game, where winners cannot be identified at the individual level).
Table 4.1: Descriptive statistics of the sample

<table>
<thead>
<tr>
<th></th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>49.9</td>
</tr>
<tr>
<td>female</td>
<td>50.1</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>18 – 24</td>
<td>24.3</td>
</tr>
<tr>
<td>25 – 29</td>
<td>27.0</td>
</tr>
<tr>
<td>30 – 34</td>
<td>18.5</td>
</tr>
<tr>
<td>35 – 39</td>
<td>10.7</td>
</tr>
<tr>
<td>40 – 49</td>
<td>10.1</td>
</tr>
<tr>
<td>50 or older</td>
<td>9.3</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>college degree</td>
<td>54.0</td>
</tr>
<tr>
<td>some college</td>
<td>34.2</td>
</tr>
<tr>
<td>high school or other</td>
<td>11.8</td>
</tr>
<tr>
<td><strong>Labor market status</strong></td>
<td></td>
</tr>
<tr>
<td>employed</td>
<td>54.1</td>
</tr>
<tr>
<td>self-employed</td>
<td>12.7</td>
</tr>
<tr>
<td>unemployed</td>
<td>11.3</td>
</tr>
<tr>
<td>student</td>
<td>13.0</td>
</tr>
<tr>
<td>other</td>
<td>8.9</td>
</tr>
<tr>
<td><strong>Current location</strong></td>
<td></td>
</tr>
<tr>
<td>at home</td>
<td>85.4</td>
</tr>
<tr>
<td>at work</td>
<td>9.9</td>
</tr>
<tr>
<td>other</td>
<td>4.7</td>
</tr>
<tr>
<td><strong>Prior MTurk studies</strong></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>6.8</td>
</tr>
<tr>
<td>1 – 9</td>
<td>19.3</td>
</tr>
<tr>
<td>10 – 99</td>
<td>32.9</td>
</tr>
<tr>
<td>100 – 999</td>
<td>30.2</td>
</tr>
<tr>
<td>1000 or more</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Notes: Labor market status recoded from multiple response data (prioritizing categories in the order as listed in the table). N = 6,152

were randomized and followed a uniform distribution. The die could be rolled several times, but as explained to the respondents, only the first roll counted.

In the prediction game participants had to correctly predict the outcome of a die roll to win the $2 bonus payment. On a first screen, the rules of the game and the conditions under which a participant would win the bonus payment were explained. On the second screen, participants were asked to make their prediction (in private) and memorize it. On the third screen they were instructed to roll the die, inspect the result, and then indicate whether their prediction was correct or not. Because the prediction was made in private, cheating could not be detected. Since the probability of winning was one sixth, however, the proportion
of cheating respondents can be estimated at the aggregate level (assuming that all respondents whose prediction was correct do claim the bonus payment).

In the roll-a-six game participants had to roll a six in order to win the $2 bonus payment. Respondents were again presented a first screen on which the game was explained. On the second screen they were instructed to roll the die and then indicate whether the result was a six or not. As in the prediction game, cheating was easily possible as the bonus payment was determined solely on the basis of the respondent’s answer and not on the actual outcome of the roll. Furthermore, estimation of the proportion of cheaters is again possible at the aggregate level as the theoretical probability of winning was one sixth. In contrast to the prediction game, however, also the identification of individual cheaters is possible since the outcomes of the die roll were recorded. Although respondents were not told that the outcomes would be tracked, it was clear that this was possible. Therefore, the proportion of cheaters can be expected to be lower in the roll-a-six game than in the prediction game. Likewise, when asked about whether they cheated, cheating respondents in the roll-a-six game may be expected to provide more truthful answers than cheating respondents in the prediction game.

4.2.2 The sensitive question techniques

In the second part of the questionnaire, respondents were asked four “sensitive” questions, the last of which being about whether they gave an honest answer in the dice game (see table 4.2).

<table>
<thead>
<tr>
<th>Item</th>
<th>Wording</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shoplifting</td>
<td>“Have you ever intentionally taken something from a store without paying for it?”</td>
</tr>
<tr>
<td>Tax evasion</td>
<td>“Have you ever provided misleading or incorrect information on your tax return?”</td>
</tr>
<tr>
<td>Non-voting*</td>
<td>“Did you vote in the 2012 US presidential election?”</td>
</tr>
<tr>
<td>Cheating in the prediction game*</td>
<td>“In the $2 dice task at the beginning of this survey: Did you honestly report whether your prediction of the dice roll was right?”</td>
</tr>
<tr>
<td>Cheating in the roll-a-six game*</td>
<td>“In the $2 dice game at the beginning of this survey: Did you honestly report whether you actually rolled a 6?”</td>
</tr>
</tbody>
</table>

* Reverse coded for the purpose of analysis.
To evaluate different sensitive question techniques, respondents were randomly assigned to one of four conditions: direct questioning (DQ), the crosswise-model RRT (CM), the unrelated-question RRT (UQ), or the forced-response RRT (FR). Table 4.3 reports the number of observations per sensitive question technique and dice game variant. Respondents were unevenly distributed across conditions in order to counterbalance the different statistical efficiencies of the procedures.\footnote{Item-nonresponse was negligible, below 1% for all sensitive questions in all experimental conditions. We therefore refrain from reporting results on item-nonresponse in the analyses below.}

Table 4.3: Number of observations by dice game variant and sensitive question technique

<table>
<thead>
<tr>
<th></th>
<th>Prediction game</th>
<th>Roll-a-six game</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct questioning (DQ)</td>
<td>387</td>
<td>382</td>
</tr>
<tr>
<td>Crosswise-model RRT (CM)</td>
<td>1168</td>
<td>1145</td>
</tr>
<tr>
<td>Unrelated-question RRT (UQ)</td>
<td>760</td>
<td>780</td>
</tr>
<tr>
<td>Forced-response RRT (FR)</td>
<td>759</td>
<td>771</td>
</tr>
</tbody>
</table>

Direct questioning (DQ) was included as a benchmark for the evaluation of the different sensitive question techniques. The sensitive questions were introduced by a screen announcing some sensitive questions, stating the importance of honest answers for the success of the study, providing privacy assurance, and telling the respondents that their answers to the sensitive questions would not affect their payment or the HIT approval (this introductory screen was identical for all conditions). After that, the four sensitive questions followed on four separate screens.

For the crosswise-model RRT (CM) we used an implementation as proposed in Jann, Jerke, and Krumpal (2012) and Höglinger, Jann, and Diekmann (2014b). Respondents were asked two questions: A sensitive question and an unrelated non-sensitive question. Respondents then had to indicate whether their answers to the two questions were the same (both "no" or both "yes") or different (one "yes," one "no") without reporting the individual answers. The unrelated questions, which were randomly paired with the sensitive questions for each respondent, asked about the birthday (in January or February, between the 1\textsuperscript{st} and the 6\textsuperscript{th} of the month) of the respondent’s mother or father. Between the introductory screen and the screen with the first sensitive question, an additional screen was
displayed explaining the questioning technique and how it protects anonymity (similar screens were also displayed for the other sensitive question techniques).

For the unrelated-question RRT (UQ) we used an implementation as proposed by Diekmann (2012). Respondents were asked to think of an acquaintance and use the first digit of this person’s house number as their personal random number. If their random digit was 1, 2, 3, 4, or 5, respondents then had to answer the subsequent sensitive questions; otherwise they had to answer the subsequent unrelated non-sensitive questions. Diekmann (2012) provides evidence that first digits of house numbers follow “Benford’s Law”. Accordingly, the probability of 1, 2, 3, 4, or 5 (i.e., of having to answer the sensitive questions) is 0.778.\(^3\) The unrelated questions were randomly paired with the sensitive questions for each respondent and asked about the birthday of the respondent’s mother (in January–June, in an even-numbered month, in the first half of the month, on an even-numbered day, in an even-numbered year).

For the forced-response RRT (FR) we used an implementation as proposed by Höglinger, Jann, and Diekmann (2014b). Respondents were presented twelve fields on the screen, numbered from one to twelve. They were told to privately choose a field and memorize their choice (without clicking on the field). 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 of their choice. Possible instructions were “Answer question”, “Directly tick yes”, or “Directly tick no”. The instructions were randomized across fields.

4.2.3 Data analysis

The RRT leads to data misclassification so that adjusted methods for data analysis are required. Let \(Y^*\) be the (unobserved) answer to the sensitive question (\(Y^* = 1\) if the answer is “yes”, \(Y^* = 0\) else) and \(Y\) be the observed response (\(Y = 1\) if the response is “yes” in case of DQ, UQ and FR or “the same” in case of CM; \(Y = 0\) else).\(^4\) For direct questioning, \(Y = Y^*\). The RRT procedures, however, introduce

---

\(^3\) To evaluate whether Benford’s Law holds, we included a question on the first digit of an acquaintance’s address for a subsample of respondents in a different experimental condition. The proportion of respondents reporting a 1, 2, 3, 4, or 5 was 0.784 (95% confidence interval: 0.763 to 0.804). Similar tests were included for all unrelated questions used in CM and UQ. Since deviations between the theoretical values (assuming an even distribution of birthdays) and the estimated proportions were only small, we focus on results based on the theoretical values in the analyses below.

\(^4\) Throughout this discussion we assume that “yes” is the sensitive answer, although some of the sensitive questions in our study were framed differently (for example, we asked respondents whether played honestly in the dice game, not whether they cheated). For the purpose of analysis, all data was appropriately recoded.
misclassification so that \( Y \neq Y^* \). In general, in a misclassification setting, the relation between \( Y \) and \( Y^* \) can be described as

\[
\Pr(Y = 1) = \Pr(Y = 1|Y^* = 1) \Pr(Y^* = 1) + \Pr(Y = 1|Y^* = 0) \Pr(Y^* = 0)
\]

Solving for \( \Pr(Y^* = 1) \) yields

\[
\Pr(Y^* = 1) = \lambda(\Pr(Y = 1)) = \frac{\Pr(Y = 1) - p_{10}}{p_{11} - p_{10}}
\]

with \( p_{11} = \Pr(Y = 1|Y^* = 1) \) and \( p_{10} = \Pr(Y = 1|Y^* = 0) \). In the RRT, \( p_{11} \) and \( p_{10} \) are known by design. Hence, we can estimate \( \Pr(Y^* = 1) \) by inserting a sample estimate for \( \Pr(Y = 1) \) (i.e. the sample mean \( \bar{Y} \)) into the above formula. Furthermore, since \( \Pr(Y^* = 1) \) is a linear transformation of \( \Pr(Y = 1) \) and, in general, \( V(ax + b) = a^2 V(x) \) (see, e.g., Mood et al. 1974: 179), the sampling variance of estimator \( \Pr(Y^* = 1) \) is given as

\[
V(\Pr(Y^* = 1)) = \frac{1}{(p_{11} - p_{10})^2} V(\Pr(Y = 1))
\]

where \( V(\Pr(Y = 1)) \) can be estimated from the data using standard techniques (e.g. as \( \bar{Y}(1-\bar{Y})/(n-1) \) where \( n \) is the sample size). For direct questioning, there is no misclassification, so that \( p_{11} = 1 \) and \( p_{10} = 0 \) and hence

\[
\lambda(\Pr(Y = 1)) = \Pr(Y = 1)
\]

For the CM, let \( p_Z \) be the known probability that the answer to the non-sensitive question is “yes”. Then \( p_{11} = p_Z \) and \( p_{10} = 1 - p_Z \). Hence,

\[
\lambda(\Pr(Y = 1)) = \frac{\Pr(Y = 1) + p_Z - 1}{2p_Z - 1}
\]

For UQ, again let \( p_Z \) be the known probability that the answer to the non-sensitive question is “yes.” Furthermore, let \( p_U \) be the probability that the respondent is instructed to answer the non-sensitive question instead of the sensitive question. We then have \( p_{11} = 1 - p_U(1 - p_Z) \) and \( p_{10} = p_U p_Z \), so that

\[
\lambda(\Pr(Y = 1)) = \frac{\Pr(Y = 1) - p_U p_Z}{1 - p_U}
\]

Finally, for FR, let \( p_{yes} \) and \( p_{no} \) be the probabilities of an unconditional “yes” or “no” answer, respectively. Then \( p_{11} = 1 - p_{no} \) and \( p_{10} = p_{yes} \), so that

\[
\lambda(\Pr(Y = 1)) = \frac{\Pr(Y = 1) - p_{yes}}{1 - p_{yes} - p_{no}}
\]
The above formulas can be used to obtain prevalence estimates for the sensitive behaviors. Employing the more-is-better assumption or comparing the estimates to the aggregate cheating rates in the dice games, we can then decide which of the techniques works best. The formulas, however, assume that respondents comply with the instructions so that, for example, no false positives occur (apart from false positives induced by design). If this assumption is violated, then the overall estimates can be misleading. To evaluate the degree to which the techniques produce valid results, we therefore perform separate analyses for those who cheated in the dice game and for those who did not cheat. What we are interested in is the true positive rate (TPR) that is, the proportion of cheaters who admit having cheated, and the false positive rate (FPR), that is, the proportion of non-cheaters who falsely "admit" having cheated. Furthermore, as overall measure of validity, we are interested in the correct classification rate (CCR).

For the roll-a-six game, these analyses are straightforward since cheating is observed at the individual level. Let \( X^* = 1 \) if the respondent rolled a six and \( X^* = 0 \) else. Furthermore, let \( X = 1 \) if the respondent claimed having rolled a six and \( X = 0 \) else. A respondent is identified as a cheater (false winner) if \( X = 1 \) even though \( X^* = 0 \). Non-cheaters are given if \( X = X^* \), that is, if \( X = X^* = 1 \) (true winner) or \( X = X^* = 0 \) (true loser). For sake of simplicity, assume that "reverse" cheating (\( X = 0 \) even though \( X^* = 1 \)) is nonexistent, that is, assume that there are no respondents who did roll a six but then did not claim the bonus payment (false losers). The true positive rate is then given as

\[
TPR = \Pr(Y^* = 1|X \neq X^*) = \lambda(Pr(Y = 1|X \neq X^*))
\]

and the false positive rate is given as

\[
FPR = \Pr(Y^* = 1|X = X^*) = \lambda(Pr(Y = 1|X = X^*))
\]

Furthermore, the correct classification rate is

\[
CCR = TPR \cdot Pr(X \neq X^*) + (1 - FPR) Pr(X = X^*)
\]

Since \( X^* \) is observed in the roll-a-six game, all of the above quantities can be readily estimated from the data. In the prediction game, however, \( X^* \) is unobserved (in the prediction game, \( X^* \) denotes whether the respondent's prediction was correct or not, \( X \) denotes whether the respondent claimed that the prediction was correct). Again, assume that all respondents whose predictions were correct did claim the bonus payment (no false losers), that is, that the combination \( X^* = 1 \) and \( X = 0 \) does not exist (as mentioned above, only one of 516 winners in the roll-a-six game did not claim the bonus payment; it appears highly plausible
to assume that the proportion of false losers is negligible also in the prediction game. We know from the design of the game that $\Pr(X^* = 1) = \frac{1}{6}$, so that the proportion of cheaters, given that there are no false losers, is equal to

$$\Pr(X \neq X^*) = \Pr(X = 1) - \Pr(X^* = 1) = \Pr(X = 1) - \frac{1}{6}$$

Furthermore, the false positive rate of true losers is given as

$$\Pr(Y^* = 1|X = X^* = 0) = \lambda(\Pr(Y = 1|X = 0))$$

The overall false positive rate or the true positive rate, however, cannot be identified without further assumptions. In general, the true positive rate can be written as

$$\Pr(Y^* = 1|X \neq X^*) = \frac{\Pr(Y^* = 1 \cap X \neq X^*)}{\Pr(X \neq X^*)} = \frac{\Pr(Y^* = 1 \cap X \neq X^*)}{\Pr(X = 1) - \frac{1}{6}}$$

To identify $\Pr(Y^* = 1 \cap X \neq X^*)$ we need to make an assumption about the false positive rate of winners. The most reasonable assumption, in our opinion, is that the false positive rate of winners is equal to the false positive rate of true losers. Both types of respondents were honest in the prediction game and we do not see much reason why they should differ in their response behavior when asked about whether they were honest or not.\(^5\) That is, we assume $\Pr(Y^* = 1|X = X^* = 1) = \Pr(Y^* = 1|X = X^* = 0)$ and, hence, $\Pr(Y^* = 1|X = X^*) = \Pr(Y^* = 1|X = X^* = 0)$, so that the overall false positive rate can be written as

$$\text{FPR} = \Pr(Y^* = 1|X = X^*) = \lambda(\Pr(Y = 1|X = 0))$$

For the derivation of the true positive rate note that

$$\Pr(Y^* = 1 \cap X \neq X^*) = \Pr(Y^* = 1 \cap X = 1) - \Pr(Y^* = 1 \cap X = X^* = 1)$$

(again given that there are no false losers). Since

$$\Pr(Y^* = 1 \cap X = 1) = \Pr(X = 1) \Pr(Y^* = 1|X = 1)$$

$$= \Pr(X = 1) \lambda(\Pr(Y = 1|X = 1))$$

\(^5\) Note, however, that the composition of the two groups is somewhat different. Among the winners there are potential cheaters, that is, respondents who would have cheated should they not have won, as well as non-cheaters. The group of true losers only contains non-cheaters. Differential assumptions about the response behavior of potential cheaters and non-cheaters could be made, but would not fundamentally change our results.
and, from results from above,
\[ \Pr(Y^* = 1 \land X = X^* = 1) = \Pr(X = X^* = 1) \Pr(Y^* = 1|X = X^*) \]
\[ = \Pr(X^* = 1) \Pr(Y^* = 1|X = X^*) \]
\[ = \frac{1}{2} \lambda (\Pr(Y = 1|X = 0)) \]

the true positive rate is given as
\[ \text{TPR} = \Pr(Y^* = 1|X \neq X^*) = \frac{\Pr(X = 1) \lambda (\Pr(Y = 1|X = 1)) - \frac{1}{2} \lambda (\Pr(Y = 1|X = 0))}{\Pr(X = 1) - \frac{1}{2}} \]

Furthermore, the correct classification rate is
\[ \text{CCR} = \text{TPR} \cdot (\Pr(X = 1) - \frac{1}{2}) + (1 - \text{FPR})(\Pr(X = 0) + \frac{1}{2}) \]

4.3 Results

4.3.1 Comparative validation

We first report results as in a standard comparative validation study, using the more-is-better assumption. Figure 4.1 displays the point estimates for the sensitive behaviors from the different sensitive question techniques, as well as the differences in the estimates between direct questioning (DQ) and the indirect techniques (also see table 4.A.1 in the appendix). For shoplifting, estimates from all three indirect techniques are significantly higher than the estimate from direct questioning. The highest estimate was obtained by the unrelated-question RRT (UQ). Also for tax evasion, all three techniques significantly outperformed direct questioning, with the crosswise-model RRT (CM) producing the highest estimate. CM also produced the highest estimates for the remaining three items, although the difference to direct questioning is not significant for the non-voting item. The unrelated-question RRT (UQ) and the forced-response RRT (FR) did not produce significantly higher estimates than direct questioning for these three items. Moreover, for cheating in the roll-a-six game, the estimate from FR is significantly lower than the estimate from DQ. From these results we would conclude that the CM clearly performed best of all techniques; it produced the highest estimates for four of the five items and produced significantly higher estimates than direct questioning for the four of the five items. The difference between CM and the other techniques is particularly pronounced for the two cheating items. While cheating rates were 5% or less according to the other techniques, they were about
15% according to CM. The results for UQ and FR are mixed. They outperformed direct questioning for the first two items, but not for the remaining three. For the last item, FR even produced a slightly negative estimate, indicating substantial non-compliance with the RRT instructions.\footnote{A negative estimate is possible if a substantial proportion of respondents deviate from the instructions determined by the randomizing device. This seems to be a common problem with the forced-response RRT (see, e.g., Covts and Jann 2011).}

![Figure 4.1: Comparative validation of sensitive question techniques (point estimates and 95% confidence intervals)](image)

### 4.3.2 Aggregate-level validation

As illustrated above, were we to conduct a comparative validation study based on the more-is-better assumption, we would find that the crosswise-model RRT is the most valid technique. However, the more-is-better assumption is a strong assumption that might be violated. In the second step, we therefore compare the prevalence estimates from the various techniques to the true prevalence of the sensitive behaviors at the aggregate level. We can conduct such an analysis for the
two items on cheating in the dice games. Figure 4.2 displays the true rates as well as the various estimates including 95% confidence intervals (left panel). In the right panel of the figure, the differences between the true rates and the estimates are shown. For the prediction game, all questioning techniques performed poorly. DQ, UQ and FR all produced estimates below 5% although the true cheating rate was around 25%. The CM comes closest to the true cheating rate with an estimate of a bit more than 15%, but still underestimates the true rate by about 11 percentage points. For the roll-a-six game, we see that DQ and UQ both produced accurate estimates of a cheating rate of about 5%. As expected, cheating was substantially less prevalent in the roll-a-six game than in the prediction game, due to the design of the game (the roll-a-six game provided less incentive for cheating than the prediction game because it was obvious that cheating could potentially be detected). FR significantly underestimated the cheating rate. For CM, on the other hand, an overestimation by about 8 percentage points occurred. Hence, while for the prediction game the more-is-better assumption seems to be valid in the sense that the highest estimate comes closest to the true value, the assumption fails for the roll-a-six game. Respondents did not substantially underreport their cheating behavior in the roll-a-six game when asked directly, probably because it was obvious that such misreporting could be detected. One could argue that cheating in the roll-a-six game is therefore not a good test case for evaluating sensitive question techniques; there is no bias that could be improved on by the techniques. On the other hand, we would expect that a valid sensitive question technique produces unbiased results also if the question is, in fact, not sensitive. A positive bias such as observed for the CM should not occur.

4.3.3 Individual-level validation

Overall, the results from the aggregate-validation are ambiguous. For the first item, cheating in the prediction game, the crosswise-model RRT (CM) is the clear winner. If we had exclusively looked at the prediction game, we would have again concluded that CM is the most valid technique. However, cheating in the roll-a-six game indicates that there might be a problem with the CM. In the third step of our analysis we therefore evaluate the accuracy of the measurements obtained by the different questioning approaches at the individual level. Figure 4.3 displays the true and false positive rates of the different techniques for the prediction game and the roll-a-six game. Direct questioning had a true positive rate (TPR) of only 10% in the prediction game, that is, only 10% of respondents who cheated in

---

Confidence intervals are also reported for the true cheating rates even though in the roll-a-six game the sample cheating rate can be determined exactly. The confidence intervals reflect the variability in the cheating rates one could expect were one to repeat the experiment.
the prediction game admitted having done so when asked directly. FR did not manage to improve the TPR and UQ slightly increased the TPR to 15%. The CM, on the other hand, was considerably more successful in eliciting truthful answers from cheaters in the prediction game, with a true positive rate of almost 30% (although still being far from 100%). Yet, the CM also had a substantial false positive rate (FPR) of about 10%. That is to say, about 10% of respondents who did not cheat in the prediction game accidently admitted having cheated when using the CM. Due to the (relatively) high TPR and the positive FPR the estimate of the cheating rate from the CM came closest to the true cheating rate at the aggregate level (as seen above). However, the correct classification rate (CCR) of the CM was, in fact, worst of all techniques (since about 75% of the respondents did not cheat, the positive FPR has a strong influence on the CCR). The UQ and FR did not have the problem of false positives, but did also not really improve on the TPR compared to DQ, so that these techniques did not reach a better CCR than DQ as well. Overall, for the prediction game, we can therefore conclude that the unrelated-question RRT (UQ) and the forced-response
RRT (FR) did not manage to produce more accurate measurements than direct questioning, and that the crosswise-model RRT (CM), although seemingly more valid than direct questioning at aggregate level, fared worst in terms of correct classification at the individual level due to the occurrence of false positives.

![Prediction game](image1)

![Roll-a-six game](image2)

**Figure 4.3:** Individual-level validation of sensitive question techniques (point estimates and 95% confidence intervals). Negative false positive rates were set to zero for the computation of the correct classification rate.

For the roll-a-six game (right panel in figure 4.3) we obtain a similar picture. Also here the CM was affected by a substantial amount of false positives (to a similar degree as in the prediction game) and, again, although not severely affected by false positives, the UQ and FR did not perform better than direct questioning. For true positives, the ranking of the techniques changed in that direct questioning now performed best, with a true positive rate of about 70%.
That the true positive rates for the indirect techniques were lower in this case than for direct questioning might be due to the fact that the RRT, although meant to provide an opportunity to be honest without the risk of disclosure, also provides respondents the possibility to be dishonest without the risk of disclosure. Because it was obvious in the roll-a-six game that a dishonest answer about whether a respondent cheated or not could potentially be identified, some of the respondents who would have felt compelled to answer truthfully in direct questioning might have misused the RRT as a protection mechanism to answer untruthfully without risk of detection. To summarize the results for the roll-a-six game: none of the indirect techniques managed to improve the true positive rate compared to direct questioning and the CM was affected by a substantial amount of false positives, so that similar to the prediction game, the correct classification was best for direct questioning and worst for the CM.

Our conclusion from the individual-level validation is that direct questioning, in fact, produced the most accurate measurements for both sensitive items. That is, from these results we have to conclude that direct questioning is the most valid technique. None of the tested indirect questioning techniques yielded an improvement over direct questioning. Keeping in mind that indirect techniques sacrifice statistical efficiency (and hence require larger sample sizes than direct questioning) we cannot recommend their general application (unless guaranteeing full privacy protection to respondents by misclassifying their answers is an important goal of a study). We also show that the CM is particularly problematic as it is affected by false positives. For example, the occurrence of false positives is the reason for why the CM overestimated the cheating rate in the roll-a-six game; it is also the reason why, at the aggregate level, the CM came seemingly closest to the true cheating rate in the prediction game. That the false positive rates of the CM were similar for both games indicates that there was a certain fraction of respondents in our sample who were unable or unwilling to apply the CM procedure correctly. How large this fraction is might depend on the specific population under study. It is clear, however, that the presence of such noncompliance has strong effects on the estimates obtained by the CM. We suspect that the false positives are the reason for why the CM performed so well in many previous studies that used a comparative design without the possibility for individual-level validation. False positives inflate the CM estimates and, from a more-is-better perspective, make it look like the CM provides more valid estimates than other techniques.

---

8 The possibility of such a paradoxical effect of indirect questioning techniques is also mentioned by Wolter and Preisendörfer (2013). Lelkes et al. (2012) found similar adverse effects of complete anonymity on truthful reporting.
4.4 Conclusions

In order to evaluate the validity of survey respondents' self-reports based on various sensitive question techniques we carried out an online experiment in which respondents' self-reported rates of cheating were compared to true cheating rates. Participants played one of two incentivized dice games in which they could cheat, that is, in which they could illegitimately claim a bonus payment. After the game, participants were asked whether they cheated using either direct questioning or one of several RRT implementations. The resulting self-reports were then validated against the actual rate of cheating in the dice game. Unlike most other evaluation studies of indirect questioning techniques, our study relies on a true validation criterion and detects misreporting at the individual level.

Results reveal that all tested questioning techniques suffer sizeable misclassification in the direction of the socially desirable answer. Among the different techniques only between 9% and 28% of all cheaters could be correctly classified as cheaters in the first variant of the dice game (prediction game). In the second variant of the dice game (roll-a-six game) between 41% and 71% of cheaters could be correctly classified. The large difference in the true positive rate between the two games suggests that the sensitivity of an item and – possibly, whether answers are potentially verifiable or not – has an important effect on respondents' decision whether to misreport or not. Although, at least for the prediction game, some of the evaluated indirect questioning techniques yielded higher true positive rates than direct questioning, none of the techniques produced overall more valid measurements than direct questioning. The reason is that the indirect techniques tend to produce poor results for respondents who do not possess the sensitive trait (i.e. who did not cheat). In particular, a substantial false positive rate was observed for the crosswise-model RRT (CM), that is, for the subsample of non-cheaters, the CM erroneously yielded cheating rates of about 11% or 12%. Furthermore, the forced-response RRT (FR) yielded negative cheating rates in the subsample of non-cheaters, which indicates that some of the respondents did not comply to the RRT instructions and answered “no” even though the procedure instructed them to answer “yes.” The unrelated-question RRT (UQ) had the least problems with respect to misclassification in the subsample of non-cheaters, but it did also not substantially reduce the amount of misclassification in the subsample of cheaters.

An important insight of our study is that the findings would have been quite different had there not been the possibility for individual-level validation. False positives in the CM inflated the prevalence estimates so that the CM consistently yielded higher prevalence of sensitive behaviors than direct questioning. Hence, employing the more-is-better assumption, the CM seemed superior. As illustrated
by the first sensitive item in our study for which validation was possible (cheating in the prediction game), comparing prevalence estimates from indirect questioning techniques to the true prevalence rate at the aggregate level, although certainly an improvement over the more-is-better assumption, can still be misleading. The CM provided a prevalence estimate that came closest to the true prevalence. Hence, one could again conclude that the CM has superior validity. The analysis at the individual level, however, revealed that this is a false conclusion. The CM came close to the true prevalence primarily because it misclassified some of the non-cheating respondents as cheaters. That is, our study not only shows that the CM might not be as promising as suggested by previous studies (Corbacho et al. 2016; Hoffmann et al. 2015; Hoffmann and Musch 2015; Höglinger, Jann, and Diekmann 2014b; Jann, Jerke, and Krumpal 2012; Korndörfer, Krumpal, and Schmukle 2014; Kundt 2014; Kundt, Misch, and Nerré 2014; Shamsipour et al. 2014), it also points to a general weakness in past research on sensitive question techniques. Because complicated misreporting patterns are possible, we must be very cautious when interpreting results from comparative evaluation studies employing the more-is-better assumption, from validation studies that rely on aggregated prevalence validation, or from one-sided validation studies in which the sensitive trait or behavior applies to all or none of the respondents. We argue that an integral evaluation of the performance of a sensitive questioning technique is only possible if answers can be validated at the individual level so that false negatives and false positives can be disentangled.

Of course, our study also has limitations. For example, we cannot answer why a substantial share of non-cheaters misreported in the CM. It is noteworthy that such misreporting did not occur with direct questioning. As such, we would speculate the cause might have to do with confusion rather than carelessness. It would be worthwhile to conduct further research on the CM to identify the design feature that causes this type of misreporting and to evaluate possible modifications to address the problem. Furthermore, our study uses two very specific items, cheating in the prediction game and cheating in the roll-a-six game, to evaluate the sensitive question techniques and, in addition, has been conducted in a special setting and in a special population (a survey on Amazon Mechanical Turk). Whether our results can be generalized to other sensitive questions, and to other populations and settings remains questionable. Further research should therefore investigate whether our results can be replicated in other contexts. Finally, we only evaluated three specific variants of the randomized response technique. Although the results of our study are discouraging for all three variants, there might be alternative designs or implementations that are more successful. Future research should focus on evaluating such alternatives. Using a research design that
allows individual-level validation of respondents' answers, however, would be crucial for such research to be meaningful.
## 4.A Appendix

The data and documentation of the survey and the analysis scripts are provided in the online supplement at [https://ideas.repec.org/p/bss/wpaper/17.html](https://ideas.repec.org/p/bss/wpaper/17.html) and [https://ideas.repec.org/p/bss/wpaper/18.html](https://ideas.repec.org/p/bss/wpaper/18.html).

<table>
<thead>
<tr>
<th>Method</th>
<th>Shoplifting (N = 6136)</th>
<th>Tax evasion (N = 6136)</th>
<th>Non-voting (N = 6131)</th>
<th>Cheating in the prediction game (N = 3065)</th>
<th>Cheating in the roll-a-six game (N = 3070)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct questioning (DQ)</td>
<td>40.23</td>
<td>10.03</td>
<td>34.46</td>
<td>2.33</td>
<td>3.94</td>
</tr>
<tr>
<td></td>
<td>(1.77)</td>
<td>(1.08)</td>
<td>(1.72)</td>
<td>(0.77)</td>
<td>(1.00)</td>
</tr>
<tr>
<td>Crosswise-model RRT (CM)</td>
<td>46.42</td>
<td>19.52</td>
<td>38.11</td>
<td>15.41</td>
<td>14.34</td>
</tr>
<tr>
<td></td>
<td>(1.62)</td>
<td>(1.50)</td>
<td>(1.61)</td>
<td>(2.05)</td>
<td>(2.06)</td>
</tr>
<tr>
<td>Unrelated-question RRT (UQ)</td>
<td>54.53</td>
<td>17.63</td>
<td>34.74</td>
<td>3.74</td>
<td>5.23</td>
</tr>
<tr>
<td></td>
<td>(1.64)</td>
<td>(1.42)</td>
<td>(1.60)</td>
<td>(1.63)</td>
<td>(1.66)</td>
</tr>
<tr>
<td>Forced-response RRT (FR)</td>
<td>49.22</td>
<td>14.30</td>
<td>32.51</td>
<td>0.85</td>
<td>-1.94</td>
</tr>
<tr>
<td></td>
<td>(1.71)</td>
<td>(1.52)</td>
<td>(1.68)</td>
<td>(1.83)</td>
<td>(1.73)</td>
</tr>
</tbody>
</table>

**Differences:**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CM – DQ</td>
<td>6.18</td>
<td>9.50</td>
<td>3.64</td>
<td>13.08</td>
<td>10.40</td>
</tr>
<tr>
<td></td>
<td>(2.40)</td>
<td>(1.85)</td>
<td>(2.35)</td>
<td>(2.19)</td>
<td>(2.29)</td>
</tr>
<tr>
<td>UQ – DQ</td>
<td>14.30</td>
<td>7.60</td>
<td>0.28</td>
<td>1.41</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td>(2.41)</td>
<td>(1.78)</td>
<td>(2.34)</td>
<td>(1.80)</td>
<td>(1.94)</td>
</tr>
<tr>
<td>FR – DQ</td>
<td>8.99</td>
<td>4.27</td>
<td>-1.95</td>
<td>-1.47</td>
<td>-5.87</td>
</tr>
<tr>
<td></td>
<td>(2.46)</td>
<td>(1.87)</td>
<td>(2.40)</td>
<td>(1.99)</td>
<td>(2.00)</td>
</tr>
</tbody>
</table>
Table 4.A.2: Cheating rates in the prediction game and the roll-a-six game as displayed in figure 4.2 (standard errors in parentheses)

<table>
<thead>
<tr>
<th>Method</th>
<th>Prediction game (N = 3065)</th>
<th>Roll-a-six game (N = 3070)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>observed estimated difference</td>
<td>observed estimated difference</td>
</tr>
<tr>
<td>Direct questioning (DQ)</td>
<td>23.64 (2.50) 2.33 (0.77) -21.32 (2.47)</td>
<td>4.46 (1.06) 3.94 (1.00) -0.52 (0.74)</td>
</tr>
<tr>
<td>Crosswise-model (CM)</td>
<td>26.63 (1.45) 15.41 (2.05) -11.22 (2.42)</td>
<td>6.04 (0.71) 14.34 (2.06) 8.30 (2.08)</td>
</tr>
<tr>
<td>RRT (UQ)</td>
<td>26.13 (1.80) 3.74 (1.63) -22.40 (2.30)</td>
<td>5.01 (0.78) 5.23 (1.66) 0.21 (1.65)</td>
</tr>
<tr>
<td>Forced-response (FR)</td>
<td>26.53 (1.80) 0.85 (1.83) -25.68 (2.48)</td>
<td>5.20 (0.80) -1.94 (1.73) -7.14 (1.74)</td>
</tr>
</tbody>
</table>

Table 4.A.3: Individual-level validation results in the prediction game and the roll-a-six game as displayed in figure 4.3 (standard errors in parentheses)

<table>
<thead>
<tr>
<th>Method</th>
<th>Prediction game (N = 3065)</th>
<th>Roll-a-six game (N = 3070)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR  FPR  CCR</td>
<td>TPR  FPR  CCR</td>
</tr>
<tr>
<td>Direct questioning (DQ)</td>
<td>9.84 (3.22) 0.00 (0.00) 78.68 (2.47)</td>
<td>70.59 (11.39) 0.82 (0.47) 97.90 (0.75)</td>
</tr>
<tr>
<td>Crosswise-model (CM)</td>
<td>28.36 (5.52) 10.71 (2.64) 73.07 (2.93)</td>
<td>52.93 (9.46) 11.86 (2.08) 86.01 (2.05)</td>
</tr>
<tr>
<td>RRT (UQ)</td>
<td>14.40 (4.70) -0.18 (1.94) 77.74 (2.05)</td>
<td>54.77 (10.39) 2.61 (1.60) 95.25 (1.64)</td>
</tr>
<tr>
<td>Forced-response (FR)</td>
<td>8.93 (5.05) -2.07 (2.31) 75.84 (2.18)</td>
<td>41.11 (10.66) -4.30 (1.69) 96.94 (0.73)</td>
</tr>
</tbody>
</table>

Notes: TPR = true positive rate, FPR = false positive rate, CCR = correct classification rate (negative false positive rates were set to zero for the computation of CCR)
Chapter 5

False Positives Undermine the Crosswise-Model RRT: An Enhanced Comparative Validation Design for Sensitive Question Research

Abstract Validly measuring sensitive issues such as norm-violations or stigmatizing traits through self-reports in surveys is often problematic. Special sensitive question techniques like the Randomized Response Technique (RRT, Warner 1965) and, among its variants, the recent crosswise-model RRT (Yu, Tian, and Tong 2008) should generate more honest answers by providing full response privacy. Different types of validation studies have examined whether particular 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 or behavior, higher prevalence estimates or estimates closer to a known population value have typically been interpreted as more valid estimates. If false positives occur, however, conclusions drawn under this assumption might be misleading. We present an easy-to-apply comparative validation design that is able to detect systematic false positives without the need for an individual-level validation criterion – which is often unavailable. Results from its application in a survey on “Organ donation and health” (N = 1,686) showed that a crosswise-model RRT implementation produced false positives to a non-ignorable extent. This serious defect was not revealed by several previous valida-

This chapter is an edited version of Höglinger, Marc and Andreas Diekmann. 2016. “False Positives Undermine the Crosswise-Model RRT: An Enhanced Comparative Validation Design for Sensitive Question Research.” Unpublished working paper. We thank Murray Bales for proofreading the manuscript.
5.1 Introduction

Measurements of sensitive issues such as norm-violations or stigmatizing traits through self-reports in surveys are often not reliable. Validation studies show that a considerable share of respondents falsely deny sensitive behavior when asked about it in surveys (e.g., Höglinger and Jann 2016; Locander, Sudman, and Bradburn 1976; Preisendörfer and Wolter 2014). In the best case, sensitive behavior is simply underestimated using such biased data while, in the worst case, conclusions about correlates and causes of the sensitive behavior in question are plain wrong. Despite this serious flaw, research in deviance, epidemiology, political science, and many other areas relies heavily on self-report data. Finding ways to validly measure sensitive items is, therefore, very important. However, surveying sensitive topics poses not only a measurement problem. For some highly sensitive issues, for example illegal activities, or—to use a health-related example—HIV infection in particular contexts, respondents might need special protection beyond what the usual survey confidentiality and privacy measures can provide to absolutely prevent sensitive information being leaked during and after the surveying process.

5.1.1 The Randomized Response Technique

Special sensitive question techniques such as the Randomized Response Technique (RRT, Warner 1965) and, among its several variants, the recently proposed crosswise-model RRT (Yu, Tian, and Tang 2008) are supposed to provide a solution to both problems mentioned. Using some randomization procedure, such as dice, that introduces noise into the response process, this technique grants respondents full response privacy. Full response privacy means there is no possibility to infer from a single respondent’s response their actual answer to a sensitive question. In turn, respondents are supposed to answer more honestly and the validity of self-reports should increase. 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. The literature is indeed full of examples of RRT applications that did not work as well as expected (e.g., Coutts and Jann 2011; Holbrook and Krosnick 2010; Peeters 2005; Wolter and Preisendörfer 2013). Moreover, Höglinger, Jann, and Diekmann (2014b)
showed that minor differences in details of RRT implementations lead to quite
diverse prevalence estimates. Therefore, carefully evaluating whether particular
implementations actually improve data validity is crucial before they are used in
surveys.\footnote{This also holds for RRT cheating-detection models that are intended to correct for respondents' non-compliance and have been repeatedly claimed to improve data validity. However, they rely on strong assumptions about the potential type of non-compliance (Clark and Desharnais 1998; Moshagen, Musch, and Erdfelder 2012; Moshagen et al. 2010; van den Hout, Böckenholt, and van der Heijden 2010).}

\subsection*{5.1.2 Comparative RRT validation studies}

The vast majority of RRT evaluations are what we call in the following \textit{comparative validation studies}\footnote{The typology for the different validation strategies is taken from Höglinger and Jann (2016). For other in-depth discussions of validation strategies, see Umesh and Peterson (1991) or Moshagen et al. (2014).}. Prevalence estimates of various 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).\footnote{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 such
as blood donation (less-is-better applies then). The more-is-better (less-is-better)
assumption is plausible for items that are unequivocally judged as socially un-
derirable (desirable), and where underreporting (overreporting) 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 com-
pletely opposite way by a different subpopulation (Percy et al. 2005; Smith 1992).
Hence, the direction of a potential social desirability bias might differ between
groups.

Moreover, some respondents actually might falsely admit sensitive behav-
or, i.e. they respond as if they possess a sensitive trait although they do not.
We call this type of misreporting false positives. While false positives are quite
unlikely for direct questioning (albeit not impossible), their occurrence is much
more likely 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 DQ mode to answer carelessly, including falsely giving a
socially undesirable response. If false positives occur, however, a higher prev-
lence estimate of a socially undesirable trait resulting from an RRT application might not be the result of more valid data. The more-is-better assumption is no longer tenable as soon as false positives might occur and conclusions regarding the validity of a particular technique relying on it are possibly wrong.

An often-cited meta-analysis (Lensvelt-Mulders et al. 2005) concluded on the basis of 32 comparative and six individual-level validation studies that “randomized response designs result in more valid data”. Many new comparative studies have been carried out since then, with only some authors acknowledging the more-is-better assumption might be critical and results should be interpreted with care (e.g. Krumpal 2012; Moshagen et al. 2010; Ostapczuk et al. 2009; Ostapczuk, Musch, and Moshagen 2011; St. John et al. 2010). Results regarding the validity of the RRT have been mixed, with some comparative studies reporting serious problems such as lower prevalence estimates than direct questioning, unrealistically high prevalence estimates, or negative estimates (Coutts and Jann 2011; Höglinger, Jann, and Diekmann 2014b; Holbrook and Krosnick 2010). However, as these studies relied crucially on the more-is-better assumption, the results must be interpreted with great caution.

5.1.3 Aggregate and individual-level validation studies

Aggregate-level validation studies compare estimated prevalence estimates to a known aggregate criteria such as official voting turnout rates (recent examples are Rosenfeld, Imai, and Shapiro 2015; Moshagen, Musch, and Erdfelder 2012). They are preferable to comparative validations because they do not need the direct questioning estimate as a benchmark. However, 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. Again, using aggregate-level validation usually does not allow a conclusion to be drawn about a sensitive question technique’s validity.

Individual-level validations, finally, 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 to draw conclusions regarding the validity of the self-reports. Without doubt, individual-level validation studies are preferable to comparative and aggregate-level validations. However, the topics’ range for individual-level validations is extremely restricted. For many areas or items of interest they are impossible to carry out. Often, it is a unique opportunity that gives researchers access to sensitive individual record data that can be used as a validation criterion for such a study. As a consequence, individual-level validations are rare, usually deal with special populations, and often cannot be
5.1. Introduction

They are thus not used systematically for methods research. Since 2000 we indeed know of only a few RRT individual-level validation studies that have been published (Hoffmann et al. 2015; Höglinger and Jann 2016; John et al. 2013; Kirchner 2015; Moshagen et al. 2014; van der Heijden et al. 2000; Wolter and Preisendörfer 2013). Moreover, most of these had severe restrictions. Some 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). Hence, their conclusions regarding the general validity of the evaluated techniques, nonetheless, implicitly rely on the more-is-better assumption. Others used designs that allowed for testing for false positives in principle, but did not make use of this opportunity (Hoffmann et al. 2015; Kirchner 2015). This indicates a profound lack of awareness of the potential occurrence of false positives in sensitive question research.

5.1.4 The seemingly promising crosswise-model RRT variant

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. In the crosswise-model, respondents are asked two questions simultaneously, a sensitive one (e.g., “Have you been tested HIV positive?”), 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 not known, 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. Regression analysis with individual-level covariates is possible as it is for all RRT variants (Fox and Tracy 1986, for the crosswise-model in particular Jann, Jerke, and Krumpal 2012). 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 that distrustful respondents unconditionally choose the “no” response as a self-protective strategy irrespective of the RRT instructions or their true answer (Coutts et al. 2011).

---

4 An exception are some recently proposed experimental validation designs (Hoffmann et al. 2015; Höglinger and Jann 2016; Moshagen et al. 2014).
5.1.5 More-is-better untenable for the crosswise-model

The crosswise-model has elicited higher prevalence estimates of sensitive behavior or attitudes than direct questioning in a series of comparative validation studies (Hoffmann and Musch 2015; Höglinger, Jann, and Diekmann 2014b; Jann, Jerke, and Krumpal 2012; Korndörfer, Krumpal, and Schmukle 2014; Kundt 2014; Kundt, Misch, and Nerré 2014; Shamsipour et al. 2014) and in one individual-level validation study not considering false positives (Hoffmann et al. 2015). Relying on the more-is-better assumption, this has typically been interpreted as more valid estimates – albeit some authors called for caution before drawing a final conclusion about the crosswise-model’s validity. And indeed, recently, Walzenbach and Hinz (2014) found unrealistically high prevalence estimates of socially desirable behavior, suggesting the crosswise-model might inflate prevalence estimates. Finally, in an individual-level validation study Höglinger and Jann (2016) found that the crosswise-model produced considerable false positive rates of 11% and 12%. To validate sensitive question techniques, they let respondents play dice games in which they could cheat for money. After the game, respondents were surveyed as to whether they had played honestly or not, and the resulting self-reports validated with actual cheating. The study was carried out as a survey on Amazon Mechanical Turk and, besides the special study population involved, the dice games that induced cheating in respondents produced a very special setting in which the sensitive question techniques were validated. It is therefore desirable to complement and corroborate this finding with other studies. However, even though we do not definitely know whether the crosswise-model regularly produces false positives or only in some implementations and in particular circumstances, the fact that false positives occurred implies that blind reliance on the more-is-better assumption is definitely untenable.

5.1.6 False positives — a blind spot in past sensitive question research

False positives might also occur in other RRT variants\(^5\), and even with other sensitive question techniques such as the item count technique or list experiment (for a recent application and a review, see Blair, Imai, and Lyall 2014), forgiving wording or other question format changes. But validation studies have so far largely neglected this possibility. We think devoting more effort to detecting potential false positives produced by sensitive question techniques is strongly advisable. One reason for this apparent blind spot in sensitive question research

\(^5\) Höglinger and Jann (2016), however, found no false positives for two forced response and one unrelated question RRT implementation. Also in direct questioning no false positives occurred.
is the difficulty of carrying out individual-level validation studies. The fact they are, in addition, typically hard to replicate due to the often unique opportunity of having access to individual validation data is a serious obstacle to forming incremental knowledge and innovation in sensitive question research. The need for easy-to-implement validation designs that can be systematically used for sensitive question research arises from this.

5.1.7 This study: detecting false positives with an enhanced comparative validation design

We propose as an alternative to the standard comparative and aggregate-level validation studies a comparative design which is able to detect systematic false positives. Hence, it allows for a test of the crucial more-is-better assumption 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. This idea was inspired by the over-claiming method where self-enhancing individuals claim knowledge of non-existent foils (Phillips and Clancy 1972; Paulhus et al. 2003) and a comparative crosswise-model validation using a low-prevalence item (Walzenbach and Hinz 2014).

We present the results of an application of the zero-prevalence comparative validation in a survey on “Organ donation and health” (N = 1,685), where respondents were asked about their willingness to donate organs after death, whether they had ever donated blood and whether they drink excessively. As zero-prevalence items served questions on having received a donor organ and on having suffered from Chagas disease, a disease with nearly zero prevalence in Germany where the study was carried out. The sensitive question technique validated was an implementation of the crosswise-model for which we had evidence from a previous individual-level validation (Höglinger and Jann 2016) that systematic false positives occurred. The goal of the present study was twofold: Replicating the finding that the crosswise-model produced false positives and assessing whether the suggested enhanced comparative validation design is able to detect false positives. As will be shown, the suggested design worked as expected. Our results showing that an application of the crosswise-model generated false positives is in line with the previous individual-level validation study. This is a seemingly persistent serious defect that, however, several comparative validation
studies could not reveal because they did not consider false positives and whose conclusions on the crosswise-model’s validity are likely flawed. In addition, we used a non-sensitive question on respondents’ educational achievement for an individual-level validation that corroborated the results from the zero-prevalence comparative validation.

5.2 Data and methods

5.2.1 General design and respondents’ characteristics

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. Full documentation including screen shots of the questionnaire is available in the online supplement. After excluding one respondent who assessed his language skills (in German) as “rather poor”, 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 4 minutes 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. The educational background was quite above-average with 76% having accomplished the general or subject-specific university entrance qualification (about equivalent to a High School diploma).


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

8 We additionally performed most analyses excluding the 47 respondents who 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.


Table 5.1: Sensitive questions

<table>
<thead>
<tr>
<th>Item</th>
<th>Wording</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never donated blood</td>
<td>“Have you ever donated blood?”</td>
</tr>
<tr>
<td>Unwilling to donate organs</td>
<td>“Are you willing to donate your organs or tissues after death?”</td>
</tr>
<tr>
<td>Excessive drinking</td>
<td>“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.)?”</td>
</tr>
<tr>
<td>Received a donated organ</td>
<td>“Have you ever received a donated organ (kidney, heart, part of a lung or liver, pancreas)?”</td>
</tr>
<tr>
<td>Suffered from Chagas disease</td>
<td>“Have you ever suffered from Chagas disease (Trypanosomiasis)?”</td>
</tr>
</tbody>
</table>

* Reverse coded for the purpose of analysis

5.2.2 The sensitive question techniques implemented

To validate the sensitive question techniques we asked respondents a series of five items with varying degrees of sensitivity. Table 5.1 lists these items which were presented in random order: 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. The last two items are the zero-prevalence items to test for systematic false positives. Both “ever received a donated organ”\(^9\) and “ever suffered from Chagas disease (Trypanosomiasis)”\(^10\) have a close to zero prevalence in the German population. We deliberately chose zero-prevalence items that suited the survey topic and had near-zero prevalence without being completely impossible so that they appeared meaningful to respondents.

One-third of the respondents were randomly assigned to the direct questioning (DQ) version of the sensitive questions (figure 5.1), and two-thirds to the crosswise-model variant (CM). The unbalanced assignment partly counterbal-

---

\(^9\) 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.

\(^10\) 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. Strassen et al. (2014) estimate an incidence rate for Germany of between 0.0001% and 0.0004%. 
ances the lower statistical efficiency of the crosswise-model RRT. The sensitive
questions were preceded by a screen announcing some sensitive questions, stat-
ing the importance of honest answers for the success of the study and providing
some privacy assurance.

Figure 5.1: Screen shot of the direct questioning implementation (translated from
German)

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
5.2). Respondents then had to indicate whether their answers to the two ques-
tions were identical (both “No”, or both “Yes”) or different (one “Yes”, the other
“No”). The CM procedure was carefully introduced to respondents. On the first
screen, we outlined the procedure and briefly explained how the technique pro-
tects individual answers. In addition, respondents were referred for further infor-
mation 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 accomplished the “Abitur”. Then, the five sensitive items followed.

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 un-
related 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. 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.

### 5.2.3 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”. $p^{yes,a}$ is the known probability of a “yes” answer to the unrelated question. The probability of the response “identical” then is

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

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

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

For the direct questioning data, we set $p^{yes,a}$ 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-square 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 \texttt{irreg}
(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 2014b 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).

5.3 Results

5.3.1 Sensitivity of the items

To assess the sensitivity of the five surveyed items, we asked participants towards the end of the survey to rate how touchy answering them might be to some respondents. Most items were not assessed as particularly sensitive by the majority of respondents (see table 5.2). 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.

<table>
<thead>
<tr>
<th>Sensitive item</th>
<th>Respondents assessing an item as “quite touchy” or “very touchy”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never donated blood</td>
<td>2%</td>
</tr>
<tr>
<td>Unwilling to donate organs after death</td>
<td>23%</td>
</tr>
<tr>
<td>Excessive drinking last two weeks</td>
<td>43%</td>
</tr>
<tr>
<td>Received a donated organ</td>
<td>11%</td>
</tr>
<tr>
<td>Suffered from Chagas disease</td>
<td>15%</td>
</tr>
</tbody>
</table>

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
5.3.2 Comparative validation of the sensitive question techniques

We now turn to the comparative validation of the sensitive question techniques. Figure 5.3 shows a comparison of self-report estimates of the sensitive items for direct questioning (DQ) and the crosswise-model (CM) (also see table 5.A.1 in the appendix). 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.

![Graph showing prevalence estimates and differences between DQ and CM](image)

Figure 5.3: 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)

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 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 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 be non-zero. The CM’s inaccurate prevalence estimates are clearly due to a false positive bias caused by this special sensitive question technique. This corroborates findings from a previous individual-level validation study (Höglinger and Jann 2016) and demonstrates the zero-prevalence comparative validation was able to detect systematic false positives. In addition, our results show that 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. Quite on the contrary, taking the finding from the two zero-prevalence items into account, it is most likely that both differences are caused at least to a considerable degree by the same systematic false-positive bias inherent in the CM implemented.

### 5.3.3 Individual-level validation

As a complementary individual-level validation of the sensitive question techniques, we used a barely sensitive question on whether respondents had accomplished 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 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 accomplished the ‘Abitur?’” with the response options “yes” and “no”. In the registration form, respondents had to select their educational.

---

11 None out of 548 respondents indicated having received a donated organ in the DQ condition, two out of 547 respondents indicated having suffered from Chagas disease.
achievement from among several categories. Third, respondents had registered for the panel up to five years prior to our survey and so it is possible that a few had accomplished the "Abitur" in the meantime and had not updated the corresponding panel information. However, 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 5.4, also see table 5.A.2 in the appendix) show that the prevalence estimates of respondents not having accomplished 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 accomplished the "Abitur" is neither barely sensitive nor ambiguous. Yet looking at results of the individual-level validation (middle and lower panel) tells a very different story. The false negative rate, i.e., the share of respondents misclassified as having accomplished the "Abitur" even though they did not is 9% in DQ and 29% for the CM. Accordingly, there actually is considerable misclassification, and substantially more in the CM relative to DQ. The false positive rate, i.e., the percentage of respondents incorrectly classified as not having accomplished the "Abitur" even though they did, is not significantly different from zero in the DQ condition but a considerable 7% in the CM. Note that the CM's high false negative and high false positive rates level each other out, resulting in an accurate aggregate prevalence estimate.

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.

12 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 selecting it, restricting the validation to respondents who unequivocally indicated having accomplished the "Abitur" or not.
5.3.4 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 next 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”. Hence, it is hard to imagine why respondents would deliberately give a false “yes” answer.

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 direct ques-
tioning 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\(^\text{13}\), 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 5.5). Randomly answering always produces more false positives than negatives for a prevalence that in reality is below 0.5, which is typical for sensitive items.\(^\text{14}\) Hence, in principle it could explain the overestimation bias found in our study as well as the consistently higher estimates from previous validations.

![Graphs](image)

Figure 5.5: 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 (half the sample) 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”

\(^{13}\) 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.

\(^{14}\) 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.
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 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 5.5). 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. A comparison of the so elicited “yes” prevalence with the theoretically expected prevalence showed a good match in general (see table 5.A.3 in the appendix). 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 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

15 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.
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 might likely be 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 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 5.3). The largest, albeit not significant effect (-4 percentage points, $p = 0.108$) was found for the introduction of a “don’t know” response option. All other manipulations such as reversing the 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 5.4). 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 posi-

---

16 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 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 5.3: Effects of CM implementation details on false positive rate (bivariate regression coefficients, standard errors in parentheses)

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

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

Tive responses. But, somehow in contrast to this finding, being among the 10% fastest respondents in answering the five sensitive items was 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 accomplished 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
Table 5.4: Bivariate associations between respondents’ behavior and personal characteristics and false positive rate (bivariate regression coefficients)

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

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

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.
5.4 Discussion and conclusion

We introduced an enhanced comparative sensitive question validation design that detects false positives and thereby allows for testing the more-is-better assumption on which comparative validations rely. Our zero-prevalence comparative validation 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. If the estimates of the zero-prevalence item are accurately zero, no systematic false positives occurred, and the more-is-better assumption is warranted on much firmer grounds. Yet, if the estimates are non-zero, false positives occurred and the more-is-better assumption is untenable for the technique under investigation. Augmented by zero-prevalence items, comparative validation studies are a useful tool for methodological research, even if false positives might occur – a possibility that definitely should not be a priori ruled out.

Validating an application of the recently proposed crosswise-model RRT (CM) with the suggested validation design we found that the CM produced false positives to a non-ignorable extent. For the two zero-prevalence items we found false positive rates of 8% (received a donated organ) and 5% (suffered from Chagas disease). This result confirms the finding in Höglinger and Jann (2016). The comparative validation with a zero-prevalence item proved capable of detecting false positives, which is otherwise only possible using individual-level validation studies that are often difficult or impossible to carry out. In addition, an individual-level validation using a non-sensitive question corroborated that the CM produced false positives. Previous validation studies appraised the crosswise-model for its easy applicability and seemingly more valid results (Hoffmann and Musch 2015; Hoffmann et al. 2015; Höglinger, Jann, and Diekmann 2014b; Jann, Jerke, and Krumpal 2012; Korndörfer, Krumpal, and Schmukle 2014; Krumpal 2012; Kndt 2014; Kndt, Misch, and Nerré 2014; Shamsipour et al. 2014). However, none of these considered false positives. Our results strongly suggest that the crosswise-model as implemented in those studies in reality does not, as previously suggested, produce more valid data than DQ.

Further, the validation design used allowed us to analyze potential causes and correlates of false positives. Yet the results showed that false positives were not clearly related to any of the CM implementation details we experimentally manipulated. However, by excluding responses elicited using some potentially problematic unrelated questions that might not have produced the expected “yes” answer prevalence, false positives could be reduced considerably for one item (Chagas disease). Looking at respondents’ behavior and personal characteristics, false
positives were positively associated with speeding through the crosswise-model explanation screen and, inexplicably to us, with socially desirable responding as measured by the Crown-Marlowe scale. Still, each of these factors can 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. Some of the causes of false positives might be circumvented or alleviated by improvements in details of the crosswise-model implementation. Possibly, a different randomizing device with a more unequivocal outcome, for instance, a spinner as used in Corbacho et al. (2016) or a "number-picking" table as used in Höglinger, Jann, and Diekmann (2014b), are less prone to bias than the unrelated questions we used. Most conveniently, our validation design allows for testing such implementation 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 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. The results presented in previous validation studies that did not consider the possibility of false positives would be much more credible had they also implemented a zero-prevalence item that would have revealed possible systematic false positives.

To conclude, the main lesson from this study is, in our view, 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, not even when using an aggregate-level validation. Consequently, this paper would have ended up once more supporting the seemingly superior validity of the crosswise-model – putting more or less emphasis on the limitation due to the reliance on the more-is-better assumption. 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 (Höglinger and Jann 2016; Hoffmann et al. 2015) provide useful tools for overcoming this blind spot in future studies.
5.A Appendix

The full survey documentation, data, and analysis scripts including extended analyses are available as online supplement at http://www.socio.ethz.ch/forschung/organspende/supplement.html.

Table 5.A.1: Comparative validation of sensitive question techniques as displayed in figure 5.3 (standard errors in parentheses)

<table>
<thead>
<tr>
<th>Levels</th>
<th>Never donated blood</th>
<th>Unwilling to donate organs</th>
<th>Excessive drinking</th>
<th>Received a donated organ</th>
<th>Suffered from Chagas disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct questioning (DQ)</td>
<td>48.82 (2.14)</td>
<td>22.01 (1.82)</td>
<td>20.58 (1.73)</td>
<td>0.00 (0)</td>
<td>0.37 (0.26)</td>
</tr>
<tr>
<td>Crosswise-model (CM)</td>
<td>51.58 (2.33)</td>
<td>27.30 (2.23)</td>
<td>32.71 (2.28)</td>
<td>7.60 (1.95)</td>
<td>4.77 (1.91)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference</th>
<th>CM – DQ</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CM – DQ</td>
<td>2.76 (3.16)</td>
<td>5.29 (2.88)</td>
<td>12.13 (2.86)</td>
<td>7.60 (1.95)</td>
</tr>
</tbody>
</table>

| N           | 1669 | 1641 | 1672 | 1669 | 1669 |

Table 5.A.2: Aggregate and individual-level validation as displayed in figure 5.4 (standard errors in parentheses)

<table>
<thead>
<tr>
<th>Aggregate prevalence</th>
<th>False negative rate</th>
<th>False positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct questioning (DQ)</td>
<td>23.94 (2.02)</td>
<td>9.48 (2.73)</td>
</tr>
<tr>
<td>Crosswise-model (CM)</td>
<td>23.01 (2.43)</td>
<td>29.29 (5.03)</td>
</tr>
</tbody>
</table>

| Difference CM – DQ | -0.93 (3.16) | 19.81 (5.72) | 6.74 (2.54) |

Notes: N = 1,361. Aggregated validation values are 25.76 for DQ, and 24.97 for CM
Table 5.A.3: Comparison of the elicited and theoretical "yes"-prevalence to unrelated questions used in the CM (standard errors in parentheses)

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

Notes: *N from 250 to 268 per question. *p < 0.05
Chapter 6

Summary and Conclusions

The studies collected in this dissertation assessed different RRT implementations using various validation designs: a comparative validation, an experimental individual-level validation, and a comparative validation enhanced by a zero-prevalence item. Before I recapitulate the conclusions from each of the studies, I briefly summarize the three main outcomes of my work.

First, none of the evaluated RRT implementations succeeded in producing more valid data than direct questioning. All sensitive question techniques, including direct questioning, showed considerable shares of misreporting for at least some sensitive items. Hence, my conclusion regarding sensitive questions and the RRT is relatively pessimistic. Getting truthful answers to sensitive questions from respondents is very difficult, and the RRT, for now at least, also offers no proper solution to this. Even worse, the recent crosswise-model (Yu, Tian, and Tang 2008), a promising RRT variant that was supposed to finally achieve this goal, did not work as expected and produced sizeable false positive rates of between 5% and 12% – a misreporting type largely overlooked in previous studies. In sum, the RRT in its various variants cannot be recommended without first further clarifying which variant actually works in which implementation and in which context. The conclusion drawn 25 years ago by Umesh and Peterson (1991) in their meta-analysis of RRT research still holds today: “Contrary to common beliefs (and claims), the validity of the RRM [the Randomized Response Method, MH] does not appear to be very good”. (121)

Second, conclusions of earlier RRT validations that did not consider the possibility of false positives and blindly relied on the more-is-better assumption must be questioned. As the results showed for the crosswise-model RRT, the more-is-better assumption is not always warranted. While direct questioning and RRT variants other than the crosswise-model were less or not affected by false positives, this possibility cannot be excluded a priori. Because complicated misre-
porting patterns are possible for special sensitive question techniques, one must be very cautious when interpreting results from comparative evaluation studies, from validation studies that rely on an aggregated prevalence validation, or from one-sided validation studies in which the sensitive trait applies to all or none of the respondents. This is exemplified in the many recent studies that did not consider false positives and interpreted the crosswise-model's relatively higher prevalence estimates of sensitive behavior as more valid estimates. Their conclusions are seriously challenged by my finding that the most common crosswise-model implementation produced a sizeable share of false positives which led to inflated prevalence estimates. False positives might also occur with sensitive question strategies other than the RRT, such as the item count technique or list experiment (e.g. Blair, Imai, and Lyall 2014), “forgiving wording” (Peter and Valkenburg 2011), or other question format changes. Studies have so far largely neglected this possibility. But a real sensitive question technique validation is only possible if false negatives and false positives are considered.

Third, I presented two designs to validate special sensitive question techniques (be they the RRT or others) that overcome the mentioned weakness of most previous validations. The first design was an experimental validation where self-reports about cheating in an incentivized dice game can be validated on an individual level. The second was a comparative validation that is able to detect systematic false positives thanks to the introduction of one or more zero-prevalence items. If it can be shown that no systematic false positives occurred for the zero-prevalence item, a comparison of the sensitive question techniques under the more-is-better assumption is warranted on much firmer grounds. The advantage of the latter strategy over individual-level validations is that it is easily applicable in any survey and with any population of interest. One reason for the disregard of false positives in sensitive question research is the difficulty of carrying out individual-level validation studies. That they are, in addition, typically hard to replicate due to the often unique opportunities to access individual validation data poses a serious obstacle to the formation of incremental knowledge and innovation. The presented designs represent two valuable tools for overcoming this obstacle and might help in finally shedding light on a blind spot in earlier sensitive question research: the possibility of false positives.

The first study in chapter 2, a comparative validation, showed that different RRT implementations, even if only differing in details such as the randomizing device, produced quite diverse results. Accordingly, the results of one particular RRT implementation might not be generalizable to other implementations of the same RRT variant, less to the RRT in general. Of course, it is bad news that the RRT is very sensitive to details of the implementation because it is preferable for measurement instruments to be robust to such alterations. But that also means
RRT implementations can be advanced with appropriate design improvements. Comparing the prevalence estimates of the different methods, the forced-response RRT implementations were found not to yield higher prevalence estimates than direct questioning and even negative estimates in some cases. The latter might be caused by respondents’ noncompliance with the RRT procedure, in particular by respondents who answer “no” despite being instructed to give an automatic “yes” answer. The fact that, even by using a randomizing device that is tailored to the online mode and by putting a lot of effort into the development and pretesting of particular implementations, the forced-response RRT does not yield higher estimates of sensitive behavior than direct questioning – which would be a necessary condition for more valid answers – is a bitter pill. Regarding the crosswise-model RRT, results for the unrelated question variant were promising at first sight because it produced consistently higher prevalence estimates than direct questioning. However, the second crosswise-model variant that employed an explicit randomizing device instead of unrelated questions fared differently and produced higher estimates for only two out of the five items. Thus, also for the crosswise-model the details of the implementation had a considerable effect on the estimates. As pointed out in chapter 2, comparative validations depend crucially on the more-is-better assumption. Hence, the results from chapter 2 are in no way conclusive. We took this limitation seriously and, in the aftermath of the study, proceeded to design a validation study that allowed for a much more meaningful assessment of RRT implementations (chapter 4).

The only RRT implementation that seemingly worked better and produced similar estimates to direct questioning for three items and higher estimates for two was the Benford RRT, which was explored in greater detail in chapter 3. This implementation used unrelated questions as a randomizing device. In addition, it made use of the “Benford illusion”, respondents’ misperception of Benford-distributed first digits, to increase the statistical efficiency of the RRT without jeopardizing respondents’ perceived privacy protection (Diekmann 2012). Besides producing reasonable estimates, it was not affected by the problem of negative estimates, a recurrent problem of forced-response RRT implementations. Regarding the success of the Benford illusion, the results were not conclusive. Relative to the other evaluated implementations, respondents gave the Benford RRT a lower rating with regard to the perceived protection, reasonableness, and understanding of the special technique. However, no effect of a change in objective privacy was found on respondents’ perceived privacy protection nor on the prevalence estimates of sensitive behavior. The fact that respondents’ perceived privacy protection is not affected by an, albeit small, change in objective privacy suggests that respondents’ perceived privacy protection is mostly driven by design
details other than the mere choice of $p$, the probability with which respondents are instructed to answer the sensitive question.

In an attempt to overcome the weakness of the first study (and of most other previous RRT validations), for the second study in chapter 4 I developed an experimental design where respondents' self-reports about cheating in a dice game could be validated on an individual level. Hence, false negatives as well as false positives could be identified. The results revealed that all evaluated questioning techniques suffered from a sizeable misclassification. Only a small share of all cheaters could be correctly classified as such. None of the evaluated special techniques yielded higher true positive rates nor more valid overall estimates than direct questioning. Interestingly, direct questioning fared considerably better in classifying cheaters in one of the two dice games where cheaters were potentially (and quite obviously) verifiable. Hence, unprotected answering in the direct questioning mode coupled with the potential verifiability of answers might lead to more honest answering — because respondents might fear that their lying will be discovered (also see Freisendorfer and Wolter 2014 for this argument). The allegedly superior crosswise-model implementation performed the worst. Its higher prevalence estimates of sensitive behavior, previously interpreted as more valid estimates in comparative and aggregate-level validations, turned out to be the result of a considerable number of false positives. These false positives inflated the aggregated prevalence estimates. Hence, the crosswise-model is likely not as promising as suggested in a series of earlier studies (including my own, see chapter 2; Corbacho et al. 2016; Hoffmann and Musch 2015; Hoffmann et al. 2015; Jann, Jerke, and Krumpal 2012; Korndörfer, Krumpal, and Schmukle 2014; Krumpal 2012; Kunt 2014; Kundt, Misch, and Neré 2014; Shamsipour et al. 2014). Perhaps the most important insight of chapter 4 is that the most common validation strategies in sensitive question research, comparative or aggregate validations can lead to false conclusions — and very likely have done so in the case of the crosswise-model. Our findings from this study would have been very different had we not considered false positives in addition to false negatives.

The study in chapter 5 uses an enhanced comparative validation design that is able to detect false positives and, in this way, allows for testing the more-is-better assumption on which comparative validations necessarily rely. This is achieved by introducing one or more (near) zero-prevalence items among the sensitive items surveyed. Because systematic false positives are detected, the more-is-better assumption rests on much firmer grounds than in standard comparative validation studies. Most importantly, the zero-prevalence comparative validation does not need an individual-level validation criterion, which is often unavailable. The zero-prevalence comparative validation cannot replace individual-level validations that, without doubt, are preferable for many reasons. But it does represent
a useful complement because individual-level validations cannot be performed with many survey topics and populations of interest. The past has shown that the difficulty of carrying out individual-level validations led to their infrequent and unsystematic use for sensitive question research. The presented design, in contrast, is easily applicable. It was able to replicate the finding that the unrelated question crosswise-model implementation produced considerable false positives. Further, the design allows for analyzing effects on and covariates of false positives. In the study, however, I could not identify a factor that clearly caused the false positives. Yet it seems likely that different sources of error such as arbitrary, i.e. random, answering or biased unrelated question outcomes jointly produce a sizeable amount of false positives. Further research must clarify whether RRT design improvements such as a more unequivocal randomizing device or a better explanation of the procedure are able to reduce this type of misclassification.

To conclude, the main contribution of this dissertation lies in the critical application of different validation strategies for sensitive question techniques, the critical discussion of their weaknesses, and the development of novel validation designs that overcome the limitations of most previous validations. What I did not achieve is to develop an RRT implementation that really works and can be recommended to survey practitioners. This remains a goal for future research. I believe the RRT might still have potential and that we should not be too harsh in judging this particular method. I am convinced that if we look at other methods with the same scrutiny we will find similar problems. Methods research looks for problems where others do not bother or where others do not even think there might be problems (like the false positives in this specific area). Therefore, results of methods research are often frustrating at first. They show how things go wrong – and often there is no immediate fix to offer. But, with further effort we will certainly find solutions, whether they actually successfully motivate respondents to give accurate answers, or we at least manage to properly adjust our analyses for misreporting or RRT non-compliance – like RRT cheating detection models are supposed to do (e.g. Clark and Desharnais 1998; Moshagen, Musch, and Erfelder 2012; Moshagen et al. 2010; van den Hout, Böckenholt, and van der Heijden 2010).1

Despite these challenges, surveys and self-reports will remain an invaluable tool in the social scientist’s toolbox and only constant and ongoing efforts to ensure valid measurement will allow us to use them to gain true insights into our substantive research problems. The experimental individual-level validation and the enhanced comparative validation presented in chapters 4 and 5 both provide useful instruments for the ongoing research into the development of sensitive

1 A conclusive validation of these techniques is still outstanding.
question techniques that will one day help to improve the validity of self-report data.
References


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).


References


References


Curriculum vitae

Marc Höglinger

Date of birth      May 16, 1979
Place of origin   Grüssch (GR), Switzerland
Citizenship       Swiss
Marital status    living together, two children

Academic and professional positions

2010 – present    Research assistant and doctoral candidate, ETH Zurich, Chair of Sociology
2010 – present    Lecturer in research methods and statistics, Kalaidos University of Applied Sciences, Zurich, and Caroum Campus, Department of Health Sciences, Zurich (freelance)
2007 – 2009       Research assistant and doctoral candidate, University of Bern, Chair of Sociology (50%)
2006 – 2010       Scientific collaborator and lecturer, Kalaidos University of Applied Sciences, Zurich
2005 – 2006       Scientific project collaborator Swisscom Innovations, Economic & Social Aspects division, Bern (50%)
2000 – 2006       Social care assistant in a housing center for refugees (50%)

Education and training

2007    lic. phil. (M.A.) in Sociology, with minors in Economics and Social & Economic History, University of Zurich
2008    Essex Summer School in Social Science Data Analysis, University of Essex, UK
2009    ICPSR Summer Program in Quantitative Methods of Social Research, University of Michigan, US