



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](#) →

This page was generated automatically upon download from the [ETH Zurich Research Collection](#). For more information please consult the [Terms of use](#).

Diss. ETH No. 23471

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 Grüşch, GR

accepted on the recommendation of

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

2016

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

Eine grosse Herausforderung bei Befragungen ist die valide Erfassung von sozial abweichenden oder sonst heiklen Verhaltensweisen und von stigmatisierenden Eigenschaften. Studien haben gezeigt, dass der Anteil der Befragten, die dazu falsche Angaben machen, beträchtlich sein kann. Dennoch basiert ein Grossteil der Forschung zu sozialen Normen und Devianz, Epidemiologie, politischen Einstellungen und Verhalten, und vielem mehr auf Selbstangaben in Umfragen. Diese Dissertation befasst sich mit der Validierung eines speziellen Verfahrens, welche Befragte dazu bringen soll, solche heiklen Fragen ehrlich und korrekt zu beantworten: die Randomized Response Technik (RRT, Warner 1965). Die RRT schützt die individuellen Antworten der Befragten durch Randomisierung, so dass von der gemachten Angabe eines Befragten nicht auf seine effektive Antwort auf die heikle Frage geschlossen werden kann. Weil Befragte so nichts zu befürchten haben, wird angenommen, dass sie ehrlich antworten. Inwiefern dies geschieht und die RRT tatsächlich validere Messungen ermöglicht, ist allerdings nicht abschliessend geklärt. Validierungsstudien liefern widersprüchliche Befunde, inwiefern die RRT ihr theoretisches Versprechen einlösen kann.

Meine Dissertation besteht aus mehreren Validierungsstudien, bei denen ich verschiedene RRT-Implementierungen für Online-Befragungen evaluiert habe. Die Ergebnisse sind, um es kurz zu fassen, niederschmetternd: Keine einzige evaluierte RRT-Implementierung generierte validere Daten als die normale direkte Befragung. Im Gegenteil, viele Implementierungen zeigten gar höhere Missklassifikations-Raten. Insbesondere produzierte eine Implementierung des vielversprechenden neuen Crosswise Modell RRT (Yu, Tian, and Tang 2008) beträchtlich viele falsch Positive, d.h. viele Respondenten wurden fälschlicherweise als Träger eines sensitiven Merkmals klassifiziert, obwohl sie dies in Realität gar nicht sind. Falsch Positive sind ein Typ von Missklassifikation, der in der Forschung zu heiklen Fragen und RRT bis anhin kaum Beachtung gefunden hat. Basierend auf diesen Validierungsergebnissen kann die RRT in ihren verschiedenen Varianten bis auf weiteres nicht zum generellen Einsatz empfohlen werden.

Zweitens kläre ich, was unterschiedliche Validierungsstrategien tatsächlich über eine bestimmte spezielle Fragetechnik aussagen können. Ich zeige, dass Validierungsstudien, welche die Möglichkeit von falsch Positiven nicht berücksichtigen, irreführend sein können. Dies wird exemplarisch an einer Implementierung des Crosswise Modells ersichtlich, welches in unseren Studien eine beträchtliche Rate von falsch Positiven produzierte. Ein gravierender Mangel, welcher von einer Reihe früherer Validierungsstudien übersehen wurde, da sie falsch Positive nicht berücksichtigten. Schlimmer noch, diese Studien interpretierten die resultierenden höheren Prävalenzschätzungen sensitiver Merkmale – mit mehr oder weniger Einschränkungen – als validere Schätzungen unter der sogenannten “More is better”-Annahme. Diese geht davon aus, dass soziale Erwünschtheit die einzige Quelle von Missklassifikation ist: Respondenten verneinen fälschlicherweise heikle Verhaltensweisen oder Merkmale (falsch Negative), aber bejahen solche niemals fälschlicherweise (falsch Positive). Konsequenterweise werden höhere Prävalenzen sozial unerwünschten Verhaltens automatisch dahingehend interpretiert, dass mehr Respondenten ehrlich antworten und die Daten valider sind. Wie das Auftreten von zahlreichen falsch Positiven beim Crosswise Modell zeigte, ist die “More is better”-Annahme nicht immer haltbar und das blinde Vertrauen auf sie deshalb eine grosse Schwäche eines Grossteils der bisherigen Forschung zu heiklen Fragen.

Drittens werden zwei neue Validierungsdesigns entwickelt und eingesetzt, welche diese Schwäche überwinden und eine aussagekräftige Validierung spezieller Fragetechniken erlauben – sei dies die RRT oder andere Methoden. Das erste Design ist eine experimentelle Validierung, bei der Selbstangaben zu Schummeln in einem incentivierten Würfelspiel auf Individual-Ebene geprüft werden. Das zweite Design ist eine vergleichende Validierung, welche eine Identifikation von systematischen falsch Positiven ermöglicht. Dies geschieht durch die Einführung eines Null-Prävalenz-Items, d.h. eines Items mit einer Prävalenz von (nahezu) null in der untersuchten Population. Beide Designs sind einfach einzusetzen und gut replizierbar, denn sie benötigen kein externes Validierungskriterium auf Individual-Ebene, welches oft nicht verfügbar ist. Damit sind sie nützliche Instrumente für zukünftige systematische Forschung zu heiklen Fragen.

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 genaueren 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 Respondenten dazu bringen soll, wahrheitsgetreu zu antworten. Im Anschluss realisierten wir, dass bisherige Validierungs-Strategien, inklusive unsere eigene erste 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 aufschlussreich bezüglich der Validität einzelner RRT-Implementierungen (eine Implementierung 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 praktiziert – unhaltbar 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 Chapter 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 einzusetzendes Validierungs-Design, das einfach replizierbar ist und für zukünftige Evaluationen von RRT-Implementierungen und anderer spezieller Fragetechniken zur Erhebung heikler Themen sehr nützlich sein dürfte.