Untersuchungen zur experimentellen Kontrolle sozialer Erwünschtheit

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Julia Meisters
aus Heinsberg

Düsseldorf, Dezember 2019

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Jochen Musch

2. Prof. Dr. Axel Buchner

Tag der mündlichen Prüfung: 04.02.2020

Danksagung

Ich danke Prof. Jochen Musch ganz herzlich für die Betreuung dieser Dissertation, für seinen fachlichen Rat und seine Denkanstöße. Außerdem danke ich Prof. Axel Buchner für die Übernahme der Zweitbegutachtung.

Darüber hinaus gilt mein ganz besonderer Dank Dr. Adrian Hoffmann, der mich auf dem gesamten Weg meiner Promotion begleitet und enorm unterstützt hat. Vielen Dank, dass du immer Zeit und ein offenes Ohr für meine Fragen und Anliegen hattest! Dein Feedback war mir immer eine sehr große Hilfe!

Außerdem danke ich meiner Familie, meinem Freund und allen meinen Kollegen und Freunden, die mich in den letzten Jahren auf jede erdenkliche Weise unterstützt haben. Insbesondere danke ich all denjenigen, die diese Dissertation kritisch gelesen haben.

Inhalt

Zusammenfassung

Soziale Erwünschtheit gefährdet die Validität direkter Selbstauskünfte zu sensiblen Themen. Befragte antworten häufig nicht ehrlich, sondern in Einklang mit sozialen Normen; dies führt zu einer Unterschätzung der Prävalenz sozial unerwünschter Merkmale (Paulhus, 1991; Tourangeau & Yan, 2007). Die Randomized-Response-Technik (RRT; Warner, 1965) ist eine indirekte Fragetechnik, die soziale Erwünschtheit in Selbstauskünften kontrollieren soll. Bisherige Studien zeigen, dass die RRT zwar validere Prävalenzschätzungen für sozial unerwünschte Merkmale ermöglicht als eine direkte Selbstauskunft, die wahre Prävalenz allerdings dennoch unterschätzt (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). In der vorliegenden Dissertation wurde die Validität, Sensitivität und Spezifität mehrerer RRT-Varianten untersucht und darüber hinaus geprüft, welche Faktoren die Validität von RRT-Schätzern verbessern. In Experiment 1 wurde getestet, ob Antwortsymmetrie und die dadurch gewährleistete Abwesenheit einer eindeutig selbstschützenden Antwortalternative die Validität von RRT-Schätzungen erhöht. In einer schwachen Validierung wurden Prävalenzschätzungen verglichen, die mit Hilfe des symmetrischen Crosswise-Modells (CWM; Yu, Tian, & Tang, 2008), des asymmetrischen Triangular-Modells (TRM; Yu et al., 2008) und einer direkten Selbstauskunft ermittelt wurden. Das symmetrische CWM führte zu signifikant höheren und daher mutmaßlich valideren Prävalenzschätzungen für ein sozial unerwünschtes Merkmal als eine direkte Selbstauskunft; das TRM hingegen führte zu Schätzungen, die sich nicht von denen aus einer direkten Selbstauskunft unterschieden. In Experiment 2 wurde eine erste schwache Validierung des Extended-Crosswise-Modells (ECWM; Heck, Hoffmann, & Moshagen, 2018) vorgenommen, welches eine Erweiterung des Crosswise-Modells ist. Im ECWM räumten mehr Befragte das sozial unerwünschte Merkmal ein als in einer direkten Selbstauskunft. Experiment 3 beschäftigte sich mit dem kürzlich in zwei Studien beobachteten Problem mangelnder Spezifität im CWM (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018). In einer starken Validierung, in welcher der individuelle Merkmalsstatus der Befragten bekannt war, konnten ausführliche Instruktionen und Verständnisfragen die Spezifität des CWM bei Befragten mit hohem Bildungsniveau erhöhen, bei Befragten mit niedrigem Bildungsniveau jedoch nicht. Sowohl eine direkte Selbstauskunft als auch das CWM ergaben eine Unterschätzung der bekannten Prävalenz eines sozial unerwünschten Merkmals; diese Unterschätzung war jedoch im CWM weniger stark ausgeprägt. In Experiment 4 wurde eine neue RRT-Variante vorgeschlagen und validiert – das Cheating-Detection-Triangular-

Modell (CDTRM). In diesem Modell wurden die leicht verständlichen Instruktionen des TRM mit einem Mechanismus zur Verweigererdetektion aus dem Cheating-Detection-Modell (CDM; Clark & Desharnais, 1998) kombiniert. Im Experiment zeigte sich, dass im TRM ohne diese Erweiterung eine wichtige Annahme des Modells verletzt war. Eine starke Validierung mit bekanntem individuellen Merkmalsstatus der Befragten zeigte, dass die Validität, Spezifität und Sensitivität des CDTRM, des CDM und einer direkten Selbstauskunft unterhalb des optimalen Niveaus lagen. Insgesamt führten alle diese Fragetechniken zu einer Unterschätzung der bekannten Prävalenz eines sozial unerwünschten Merkmals; diese Unterschätzung war jedoch im CDTRM und CDM weniger stark ausgeprägt als in der direkten Selbstauskunft. Zusammenfassend zeigte die vorliegende Arbeit, dass Validität, Sensitivität und Spezifität mehrerer RRT-Varianten zwar nicht perfekt waren, die RRT aber Prävalenzschätzungen ermöglichte, die näher an der bekannten Prävalenz sozial unerwünschter Merkmale lagen als Prävalenzschätzungen aus einer direkten Selbstauskunft. Faktoren, die die Validität von RRT-Umfragen erhöhten, waren Modelleigenschaften – wie Antwortsymmetrie oder ein Mechanismus zur Verweigererdetektion –, Eigenschaften des Kontexts – wie die Verwendung ausführlicher Instruktionen und Verständnisfragen – und Eigenschaften der Stichprobe – wie ein hohes Bildungsniveau der Befragten. Sowohl eine direkte Selbstauskunft als auch die RRT führten zu Unterschätzungen der bekannten Prävalenz sozial unerwünschter Merkmale; diese Unterschätzung fiel jedoch mit der RRT geringer aus als mit einer direkten Selbstauskunft.

Abstract

Socially desirable responding threatens the validity of direct self-reports on sensitive issues. Respondents often do not answer honestly, but rather in line with social norms; this leads to an underestimation of the prevalence of socially undesirable attributes (Paulhus, 1991; Tourangeau & Yan, 2007). The *Randomized-Response-Technique* (RRT; Warner, 1965) is an indirect questioning technique that aims at controlling socially desirable responding in self-reports. Previous studies show that the RRT provides more valid prevalence estimates for socially undesirable attributes than direct self-reports, but still underestimates the true prevalence (Lensvelt-Mulders et al., 2005). In the present thesis, the validity, sensitivity and specificity of several RRT variants were assessed, and the influence of several factors potentially improving the validity of RRT estimates was examined. Experiment 1 investigated whether response symmetry, that is, the absence of a clearly self-protecting response alternative, increases the validity of RRT estimates. In a weak validation study, prevalence estimates obtained via the symmetrical Crosswise Model (CWM; Yu et al., 2008), the asymmetrical Triangular Model (TRM; Yu et al., 2008) and direct self-reports were compared. The symmetrical CWM provided significantly higher and thus presumably more valid prevalence estimates for a socially undesirable attribute than direct self-reports, whereas the asymmetrical TRM provided estimates that did not differ from estimates obtained via direct self-reports. In experiment 2, a first weak validation of the Extended Crosswise Model (ECWM; Heck et al., 2018) was conducted. More respondents admitted to the socially undesirable attribute when surveyed via the ECWM than via direct self-reports. Experiment 3 focused on the problem of suboptimal specificity of CWM estimates that was recently observed in two studies (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018). In a strong validation study, in which the status of individual respondents with respect to the sensitive attribute was known, detailed instructions combined with comprehension questions improved the specificity for higher-educated respondents, but not for lower-educated respondents. Overall, prevalence estimates obtained via both the CWM and direct self-reports underestimated the known prevalence of a socially undesirable attribute; this underestimation was however less pronounced in the CWM. In experiment 4, a new RRT variant was presented and validated—the Cheating Detection Triangular Model (CDTRM). This model combines the easy-to-understand instructions of the TRM with the cheating detection mechanism from the Cheating Detection Model (CDM; Clark & Desharnais, 1998). The experiment showed that for the TRM without this

extension an important model assumption was violated. A strong validation, in which the individual status of respondents was known, showed that the validity, specificity and sensitivity of the CDTRM, the CDM, and of direct self-reports were below the optimal level. Each of the questioning techniques underestimated the known prevalence of a socially undesirable attribute; this underestimation was however less pronounced for the CDTRM and the CDM than for direct self-reports. Overall, the current work shows that the validity, sensitivity and specificity of several RRT variants were not perfect; however, prevalence estimates obtained via RRT were more valid than prevalence estimates obtained via direct self-reports. Factors increasing the validity of RRT include model characteristics—such as response symmetry or a cheating detection mechanism—, context characteristics—such as detailed instructions and comprehension questions—, and sample characteristics—such as a high level of education of the respondents. Both direct self-reports and the RRT provided underestimates of the known prevalence of socially undesirable attributes; this underestimation was however less pronounced for the RRT than for direct self-reports.

1 Einleitung

Selbstauskünfte von Befragten stellen für Forschende in der Psychologie und den Sozialwissenschaften eine weit verbreitete Datenquelle dar (Rasinski, Visser, Zagatsky, & Rickett, 2005; Schwarz, 1999). Die Validität solcher Selbstauskünfte hängt dabei in entscheidendem Maße von der Ehrlichkeit der Befragten ab (Rasinski et al., 2005). Insbesondere wenn Fragen zu sensiblen Themen wie Drogenkonsum oder Fremdenfeindlichkeit gestellt werden, antworten nicht alle Befragten ehrlich. Beispielsweise aus Angst vor Strafen oder sozialer Missbilligung antworten einige der Befragten stattdessen so, dass ihre Antwort sozialen Normen entspricht (Paulhus, 1991; Tourangeau & Yan, 2007). Sozial erwünschtes Antwortverhalten führt dazu, dass die Prävalenz sozial unerwünschter Merkmale unterschätzt wird und die Prävalenz sozial erwünschter Merkmale überschätzt wird, und gefährdet somit die Validität von Selbstauskünften (Paulhus, 1991; Tourangeau & Yan, 2007).

Um den Einfluss sozial erwünschten Antwortverhaltens zu reduzieren, wurden indirekte Fragetechniken wie die *Randomized-Response-Technik* (RRT; Warner, 1965) entwickelt. Die RRT garantiert mit Hilfe einer Zufallsverschlüsselung die Vertraulichkeit individueller Antworten und soll es so Befragten ermöglichen, ehrlichere Antworten auf sensible Fragen zu geben. Daher sollte die RRT zu valideren Prävalenzschätzungen für sensible Merkmale führen als eine direkte Selbstauskunft. Um die Validität der RRT zu untersuchen, können verschiedene Arten von Validierungsstudien durchgeführt werden: In sogenannten *schwachen* Validierungen wird die Prävalenzschätzung, die aus einer direkten Selbstauskunft resultiert, mit der Prävalenzschätzung verglichen, die mit Hilfe einer RRT gewonnen wird. Dabei werden höhere Schätzungen bei sozial unerwünschten und niedrigere Schätzungen bei sozial erwünschten Merkmalen als valider angesehen, da diese vermutlich weniger durch soziale Erwünschtheit verzerrt sind. Bei schwachen Validierungen bleibt jedoch ungeklärt, wie gut die resultierenden Prävalenzschätzungen die wahre Prävalenz des sensiblen Merkmals abbilden und ob es sich um Über- oder Unterschätzungen handelt (Moshagen, Hilbig, Erdfelder, & Moritz, 2014; Umesh & Peterson, 1991). Wünschenswerter sind daher sogenannte *starke* Validierungsstudien, in denen ein starkes Außenkriterium zur Verfügung steht. Bei diesen Studien ist entweder die Prävalenz des sensiblen Merkmals auf Stichprobenebene bekannt oder besser noch, der wahre Merkmalsstatus auf individueller Ebene der Befragten. In beiden Fällen kann die bekannte Prävalenz mit den resultierenden Prävalenzschätzungen aus der Befragung verglichen werden (Lensvelt-Mulders et al., 2005; Moshagen et al., 2014;

Umesh & Peterson, 1991). Häufig ist es jedoch entweder nicht möglich oder aber mit hohem Aufwand verbunden, an sensible Daten zu Befragten zu gelangen; daher sind starke Validierungen selten (Lensvelt-Mulders et al., 2005; Umesh & Peterson, 1991). Eine Meta-Analyse über 32 schwache Validierungen ergab, dass die RRT zu höheren und daher nach einem schwachen Validierungskriterium zu potentiell valideren Prävalenzschätzungen für sozial unerwünschte Merkmale führte als eine direkte Selbstauskunft (Lensvelt-Mulders et al., 2005). Eine Meta-Analyse über 6 starke Validierungen konnte jedoch zeigen, dass auch mit Hilfe der RRT die bekannte Prävalenz sozial unerwünschter Merkmale noch unterschätzt wurde (Lensvelt-Mulders et al., 2005). Ein Problem der RRT ist, dass Befragte manchmal die relativ komplexen Instruktionen nicht verstehen oder der Zufallsverschlüsselung nicht vertrauen (Landsheer, van der Heijden, & van Gils, 1999). Dies kann dazu führen, dass Befragte die Instruktionen – versehentlich oder willentlich – missachten, was die Validität der RRT gefährdet (Edgell, Duchan, & Himmelfarb, 1992; Edgell, Himmelfarb, & Duchan, 1982). Um dem Problem einer solchen Instruktionsverweigerung entgegenzuwirken, wurden zum einen sogenannte *Non-Randomized-Response-Techniken* (NRRT; Tian & Tang, 2014) vorgeschlagen, welche durch besonders leicht verständliche Instruktionen gekennzeichnet sind. Hierzu zählen beispielsweise das *Triangular-Modell* (TRM; Yu et al., 2008), das *Crosswise-Modell* (CWM; Yu et al., 2008) und das *Extended-Crosswise-Modell* (ECWM; Heck et al., 2018). Zum anderen wurden Varianten der RRT entwickelt, die eine Verweigererdetektion beinhalten, wie beispielsweise das *Cheating-Detection-Modell* (CDM; Clark & Desharnais, 1998). In diesem Modell wird neben dem Anteil der ehrlichen Merkmalsträger auch der Anteil der Befragten geschätzt, die die Instruktionen nicht befolgen.

In der vorliegenden Dissertation wurde in vier Validierungsstudien geprüft, ob die zuvor genannten RRT-Varianten zu valideren und daher vermutlich zu weniger durch soziale Erwünschtheit verzerrten Prävalenzschätzungen führen, als eine direkte Selbstauskunft. Darüber hinaus wurde untersucht, welche Faktoren die Validität der RRT beeinflussen. Zunächst wurden in einer ersten schwachen Validierung Prävalenzschätzungen verglichen, die mit Hilfe des TRM, des CWM sowie mit einer direkten Selbstauskunft ermittelt wurden. Außerdem wurden die drei genannten Fragetechniken eingesetzt, um Prävalenzschätzungen für ein nichtsensibles Kontrollmerkmal zu gewinnen; diese Schätzungen wurden dann im Sinne einer starken Validierung mit der bekannten Prävalenz des Kontrollmerkmals verglichen (Experiment 1). In einer weiteren Studie wurde das ECWM, eine Erweiterung des CWM, erstmalig im Vergleich zu einer direkten Selbstauskunft schwach validiert (Experiment 2). In einer

dritten Studie wurde für das CWM mit Hilfe einer starken Validierung, in welcher der Merkmalsstatus individueller Befragter bekannt war, untersucht, inwiefern ausführliche Instruktionen und Verständnisfragen helfen können, die Validität zu verbessern (Experiment 3). Schließlich wurde in einer weiteren starken Validierung mit Daten auf individueller Ebene die Validität des im Rahmen dieser Arbeit vorgeschlagenen *Cheating-Detection-Triangular-Modells* untersucht. Dieses Modell kombiniert die leicht verständlichen Instruktionen des TRM mit dem Messmodell des CDM (Experiment 4).

In den nachfolgenden Abschnitten der vorliegenden Dissertation wird zunächst die RRT vorgestellt. Alle im Rahmen der vorliegenden Arbeit untersuchten Varianten der RRT werden kurz beschrieben und es erfolgt ein Überblick über den aktuellen Forschungsstand (Kapitel 2). Anschließend werden die Forschungsfragen der vorliegenden Dissertation abgeleitet und begründet (Kapitel 3) und die hierzu durchgeführten empirischen Studien zusammengefasst (Kapitel 4). Schließlich folgt eine Einordnung und Diskussion der Ergebnisse (Kapitel 5). Die einzelnen Manuskripte finden sich im Anhang.

2 Randomized-Response-Technik (RRT)

Die RRT ist eine indirekte Fragetechnik, die mit Hilfe einer Randomisierungsprozedur eine Zufallsverschlüsselung individueller Antworten vornimmt. Die so garantierte Vertraulichkeit individueller Antworten soll Befragte dazu motivieren, ehrlichere Antworten auf sensible Fragen zu geben als in direkten Selbstauskünften. In der ursprünglich von Warner (1965) vorgeschlagenen RRT werden den Befragten zwei gegensätzliche Aussagen präsentiert: Eine sensible Aussage A (beispielsweise „Ich habe schon einmal Kokain konsumiert“) und deren Verneinung B („Ich habe noch nie Kokain konsumiert“). Die zur Verfügung stehenden Antwortoptionen lauten „Ich stimme zu“ versus „Ich stimme nicht zu“. Eine Randomisierungsprozedur, beispielsweise in Form eines Würfelwurfs, entscheidet darüber, auf welche der beiden Aussagen Befragte antworten sollen. So könnten Befragte instruiert werden, auf Aussage A zu antworten, wenn sie eine 1 oder 2 würfeln (Randomisierungswahrscheinlichkeit p = 1/3), und auf Aussage B zu antworten, wenn sie eine 3, 4, 5 oder 6 würfeln (p = 2/3). Die Randomisierungsprozedur findet im Geheimen statt, sodass weder Dritte noch Versuchsleiter wissen, auf welche der beiden Aussagen geantwortet wurde. Somit bleibt vertraulich, ob Befragte der sensiblen Aussage zugestimmt haben und daher Merkmalsträger des sensiblen Merkmals sind. Da die Randomisierungswahrscheinlichkeit p bekannt ist, kann auf Stichprobenebene ein Maximum-Likelihood-Schätzer für die Prävalenz π des sensiblen Merkmals anhand folgender Formel bestimmt werden (Warner, 1965):

$$\hat{\pi} = \frac{p - 1 + \frac{n'}{n}}{2p - 1} \quad , \quad p \neq \frac{1}{2} \quad . \tag{1}$$

Hierbei entspricht n der Stichprobengröße und n' der absoluten Anzahl der „Ich stimme zu“-Antworten. Die Varianz des Schätzers kann wie folgt bestimmt werden (Warner, 1965; Yu et al., 2008):

$$var(\hat{\pi}) = \frac{\hat{\pi}(1 - \hat{\pi})}{n} + \frac{p\,(1 - p)}{n(2p - 1)^2} \quad . \tag{2}$$

Die Varianz der RRT ist gegenüber der Varianz einer direkten Selbstauskunft deutlich erhöht (Ulrich, Schröter, Striegel, & Simon, 2012), obwohl einige Varianten der RRT entwickelt wurden, deren Varianz geringer ist als die des ursprünglichen Warner-Modells (siehe beispielsweise Boruch, 1971; Dawes & Moore, 1980; Mangat, 1994; eine Übersicht über weitere RRT-Varianten findet sich in Chaudhuri & Christofides, 2013). Außerdem ist die Bearbeitungszeit einer RRT-Frage deutlich höher als die Bearbeitungszeit einer direkten

Selbstauskunft, denn Befragte müssen zunächst instruiert werden und gegebenenfalls ein Randomisierungsinstrument benutzen. Daher ist der Einsatz der RRT nur dann sinnvoll und gerechtfertigt, wenn die RRT auch tatsächlich zu valideren Schätzungen für die Prävalenz sensibler Merkmale führen kann als eine direkte Selbstauskunft (Lensvelt-Mulders et al., 2005). Eine Meta-Analyse über 32 schwache Validierungen zeigte, dass die RRT für sozial unerwünschte Merkmale höhere und daher mutmaßlich validere Prävalenzschätzungen ermittelte als eine direkte Selbstauskunft (Lensvelt-Mulders et al., 2005). Eine Meta-Analyse über 6 starke Validierungen zeigte jedoch, dass auch RRT-Schätzer die bekannte Prävalenz sozial unerwünschter Merkmale noch unterschätzten (Lensvelt-Mulders et al., 2005). In einigen Studien ermittelte die RRT allerdings Prävalenzschätzungen, die aufgrund von Modellpassungsproblemen unplausiblerweise unter 0% oder über 100% lagen (Holbrook & Krosnick, 2010; John, Loewenstein, Acquisti, & Vosgerau, 2018). Diese problematischen Befunde zur Validität der RRT könnten dadurch erklärbar sein, dass ein Teil der Befragten die Randomisierungsprozedur nicht wie instruiert befolgt (Clark & Desharnais, 1998; Edgell et al., 1982; Holbrook & Krosnick, 2010; John et al., 2018). Zudem existieren Hinweise darauf, dass manche Befragte die relativ komplexen Instruktionen der RRT nicht verstehen oder der Randomisierungsprozedur nicht vertrauen (Landsheer et al., 1999). Beide Aspekte bilden jedoch eine Grundvoraussetzung für die Validität der RRT (Landsheer et al., 1999). Um dem Problem der versehentlichen oder willentlichen Missachtung der RRT-Instruktionen entgegenzuwirken und so die Validität der RRT zu steigern, wurden zum einen Varianten der RRT mit vereinfachten Instruktionen entwickelt (*Non-Randomized-Response-Techniken*; Tian & Tang, 2014). Hierzu zählen das *Triangular-Modell* (TRM; Yu et al., 2008), das *Crosswise-Modell* (CWM; Yu et al., 2008) und das *Extended-Crosswise-Modell* (ECWM; Heck et al., 2018). Diese Modelle wurden im Rahmen der vorliegenden Dissertation validiert und sollen daher im Folgenden näher erläutert werden (Kapitel 2.1). Zum anderen wurden Varianten der RRT entwickelt, die eine Schätzung des Anteils der Instruktionsverweigerer ermöglichen. Hierzu zählt das *Cheating-Detection-Modell* (CDM; Clark & Desharnais, 1998), welches ebenfalls in der vorliegenden Arbeit validiert wurde und daher nachfolgend erklärt wird (Kapitel 2.2). Darüber hinaus wurde im Rahmen der vorliegenden Dissertation mit dem *Cheating-Detection-Triangular-Modell* (CDTRM) eine Kombination der leicht verständlichen Instruktionen des TRM mit dem Mechanismus zur Verweigererdetektion des CDM vorgeschlagen und validiert. Das CDTRM wird in Kapitel 2.3 erläutert.

2.1 Non-Randomized-Response-Techniken (NRRT)

Die sogenannten *Non-Randomized-Response-Techniken* (NRRT; Tian & Tang, 2014) bilden eine Modellklasse der RRT, für die kein externes Randomisierungsinstrument wie ein Würfel benötigt wird; NRRT sollen daher einfacher verständlich und leichter anwendbar sein als andere Varianten der RRT (Yu et al., 2008). Zu den NRRT zählen das *Triangular-Modell* (TRM; Yu et al., 2008), das *Crosswise-Modell* (CWM; Yu et al., 2008) sowie eine Weiterentwicklung des CWM, das *Extended-Crosswise-Modell* (Heck et al., 2018). Im Folgenden werden diese drei Modelle sowie der jeweilige aktuelle Forschungsstand dazu vorgestellt.

2.1.1 Triangular-Modell (TRM)

Im TRM (Yu et al., 2008) werden den Befragten zwei Aussagen gleichzeitig präsentiert: Eine sensible Aussage („Ich habe schon einmal Kokain konsumiert“) und eine nicht-sensible Aussage, die als Randomisierungsvariable mit der Randomisierungswahrscheinlichkeit p dient (beispielsweise „Ich bin im November oder Dezember geboren“; $p = .158$ nach offiziellen Geburtsstatistiken; Pötzsch, 2012). Befragte sollen nun angeben, ob sie „*mindestens einer* der beiden Aussagen (egal welcher)“ oder „*keiner* der beiden Aussagen“ zustimmen. Abbildung 1 stellt das TRM als Baummodell dar.

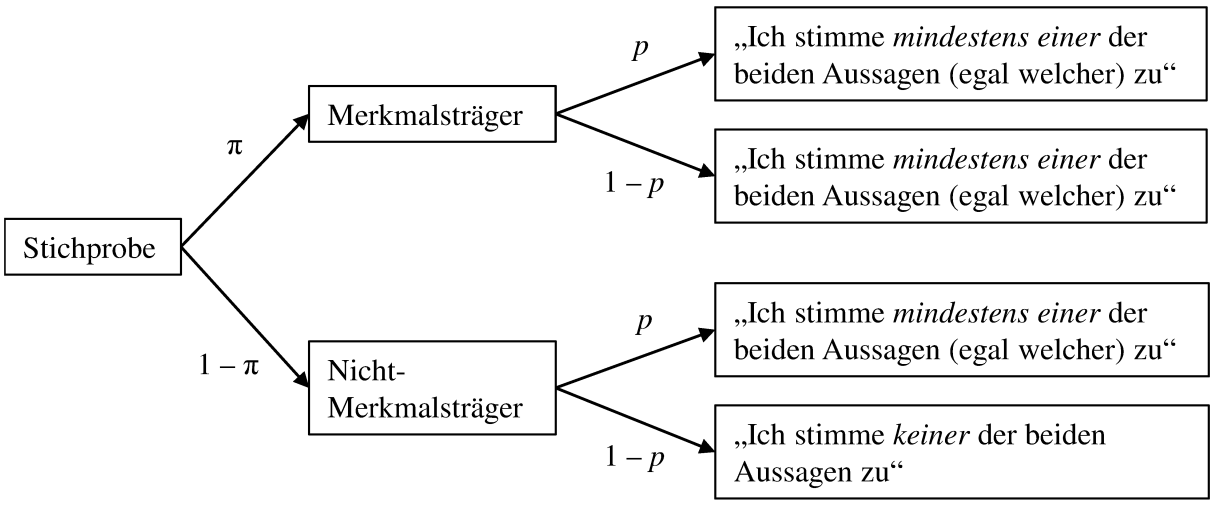


Abbildung 1: Baummodell des Triangular-Modells (Yu et al., 2008). Der Parameter π bezeichnet die unbekannte Prävalenz des sensiblen Merkmals, der Parameter p bezeichnet die Randomisierungswahrscheinlichkeit.

Maximum-Likelihood-Schätzer für die Prävalenz π des sensiblen Merkmals können anhand der folgenden Formel bestimmt werden (Yu et al., 2008):

$$\hat{\pi} = 1 - \frac{\frac{n'}{n}}{1-p} \quad . \tag{3}$$

Hierbei entspricht *n* der Stichprobengröße und *n′* der absoluten Anzahl der Befragten, welche die Antwortoption „Ich stimme *keiner* der beiden Aussagen zu“ gewählt haben.

Die Antwort „Ich stimme *mindestens einer* der beiden Aussagen (egal welcher) zu“ kann im TRM entweder von Merkmalsträgern stammen oder von Nicht-Merkmalsträgern, die aufgrund der Randomisierungsprozedur angehalten waren, diese Antwort zu wählen. Daher wird Merkmalsträgern im TRM vollste Vertraulichkeit ihrer Antworten garantiert. Im Gegensatz dazu kann die Antwort „Ich stimme *keiner* der beiden Aussagen zu“ nur von Nicht-Merkmalsträgern stammen. Insofern stellt diese Antwortoption eine *selbstschützende* Antwortoption dar, die auch unehrliche Merkmalsträger auswählen könnten, um eindeutig auszuschließen, dass sie für einen Merkmalsträger gehalten werden. Daher wird das TRM als *asymmetrisches* Modell bezeichnet. Es ist im TRM nicht feststellbar, wie viele Befragte diese selbstschützende Antwortoption wählen statt die Instruktionen zu befolgen; die Asymmetrie der Antwortoptionen kann daher potentiell die Validität des TRM gefährden.

Bislang wurden lediglich zwei schwache Validierungsstudien zum TRM durchgeführt, in denen sensible Merkmale wie das Plagiieren in studentischen Abschlussarbeiten, die Inanspruchnahme psychologischer Beratung und der Gebrauch illegaler Drogen oder verschreibungspflichtiger Medikamente zur mentalen Leistungssteigerung untersucht wurden (Erdmann, 2019; Jerke & Krumpal, 2013). Für keines dieser Merkmale führte das TRM zu Prävalenzschätzungen, die sich signifikant von denen aus einer direkten Selbstauskunft unterschieden; somit liegen bislang keine Hinweise auf eine im Vergleich zu einer direkten Selbstauskunft erhöhte Validität des TRM vor.

In der vorliegenden Dissertation wurden Prävalenzschätzungen des *asymmetrischen* TRM mit Schätzungen aus einer direkten Selbstauskunft und erstmalig auch mit Schätzungen des verwandten, jedoch *symmetrischen* CWM verglichen. Das TRM wurde außerdem erstmalig in einer starken Validierung mit der bekannten Prävalenz eines nicht-sensiblen Kontrollmerkmals verglichen (Experiment 1). Darüber hinaus wurde die Kombination der leicht verständlichen Instruktionen des TRM mit dem Messmodell des CDM zum *Cheating-Detection-Triangular-Modell* (CDTRM) vorgeschlagen. Mit Hilfe einer starken Validierung, in welcher der wahre Merkmalsstatus individueller Befragter bekannt war, wurde untersucht, ob das

CDTRM validere Prävalenzschätzungen ergibt als das TRM in seiner ursprünglichen Formulierung ohne Verweigererdetektion (Experiment 4).

2.1.2 Crosswise-Modell (CWM)

Im CWM werden den Befragten genau wie im TRM zwei Aussagen gleichzeitig präsentiert: Eine Aussage zu einem sensiblen Merkmal von unbekannter Prävalenz sowie eine nichtsensible Aussage, die als Randomisierungsvariable dient. Befragte im CWM sollen dann angeben, ob sie „*beiden* oder *keiner* der beiden Aussagen" zustimmen oder ob sie „*genau einer* der beiden Aussagen (egal welcher)" zustimmen. Beide Antwortalternativen können im CWM sowohl von Merkmalsträgern als auch von Nicht-Merkmalsträgern stammen. Daher bietet das CWM im Gegensatz zum TRM keine selbstschützende Antwortalternative und wird als *symmetrisches* Modell bezeichnet. Abbildung 2 stellt das CWM als Baummodell dar.

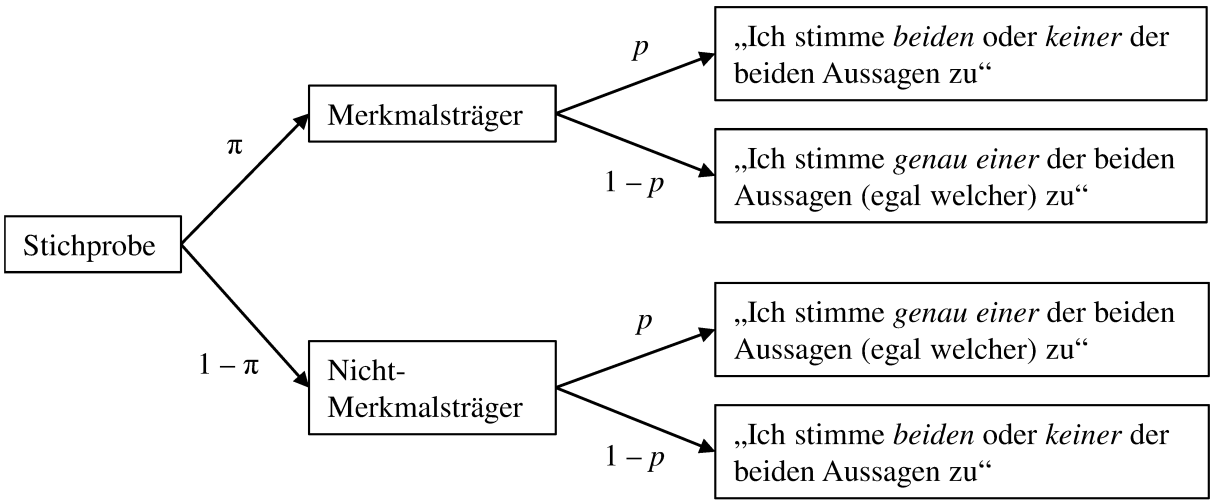


Abbildung 2: Baummodell des Crosswise-Modells (Yu et al., 2008). Der Parameter π bezeichnet die unbekannte Prävalenz des sensiblen Merkmals, der Parameter *p* bezeichnet die Randomisierungswahrscheinlichkeit.

Maximum-Likelihood-Schätzer für die Prävalenz π des sensiblen Merkmals im CWM sind identisch zu den Schätzern aus dem ursprünglichen RRT-Modell nach Warner (1965) und können folglich anhand von Gleichung (1) bestimmt werden (Ulrich et al., 2012; Yu et al., 2008). Hierbei entspricht jedoch n' der absoluten Anzahl der Antworten „Ich stimme *beiden* oder *keiner* der beiden Aussagen zu".

Das CWM wurde bereits in zahlreichen Studien untersucht. In vielen schwachen Validierungen führte das CWM dabei für sozial unerwünschte Merkmale zu signifikant höheren und insofern potentiell valideren Prävalenzschätzungen als eine direkte Selbstauskunft. Dies betraf Merkmale wie beispielsweise Xenophobie (Hoffmann & Musch, 2016), Vorurteile gegenüber weiblichen Führungskräften (Hoffmann & Musch, 2019), die Intention die Partei *Alternative für Deutschland* zu wählen (Waubert de Puiseau, Hoffmann, & Musch, 2017), Plagiieren in studentischen Abschlussarbeiten (Jann, Jerke, & Krumpal, 2012), Steuerhinterziehung (Korndörfer, Krumpal, & Schmukle, 2014; Kundt, Misch, & Nerré, 2017), illegales Doping unter Bodybuildern (Nakhaee, Pakravan, & Nakhaee, 2013) sowie Misstrauen im Vertrauensspiel (Thielmann, Heck, & Hilbig, 2016). Zudem zeigte sich in einer starken Validierung, dass das CWM die bekannte Prävalenz eines experimentell induzierten sozial unerwünschten Merkmals korrekt zu schätzen vermochte, während eine direkte Selbstauskunft die bekannte Prävalenz des Merkmals signifikant unterschätzte (Hoffmann, Diedenhofen, Verschuere, & Musch, 2015). Außerdem erwies sich das CWM als die verständlichste mehrerer indirekter Fragetechniken (Hoffmann, Waubert de Puiseau, Schmidt, & Musch, 2017). In einer weiteren Studie zum CWM wurde die Hälfte der Befragten instruiert, ehrlich zu antworten, während die andere Hälfte eine „Fake-good"-Instruktion erhielt. In dieser Studie erwiesen sich Prävalenzschätzer, die mit Hilfe des CWM ermittelt wurden, als signifikant weniger durch „faking good" verfälschbar als Prävalenzschätzer aus einer direkten Selbstauskunft (Hoffmann, Meisters, & Musch, 2019).

Die Befunde zweier starker Validierungen haben jedoch zuletzt die Validität des CWM in Frage gestellt, da gezeigt werden konnte, dass das CWM über eine suboptimale Spezifität von weniger als 100% verfügt (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018). Die Spezifität des CWM gibt den Anteil der Nicht-Merkmalsträger an, die auch als solche identifiziert werden. In einer starken Validierung wurde die Lebenszeitprävalenz zweier Merkmale untersucht, die in der Population vermutlich mit einer Wahrscheinlichkeit von nahe 0% auftreten. Demzufolge sollte die Rate der Nicht-Merkmalsträger in der Stichprobe bei nahezu 100% liegen. Für beide Merkmale schätzte eine direkte Selbstauskunft den Anteil der Nicht-Merkmalsträger wie erwartet auf 100%. Im CWM betrug dieser Anteil, und damit die Spezifität, jedoch nur 92% beziehungsweise 95% und lag damit signifikant unter 100% (Höglinger & Diekmann, 2017). In einer starken Validierung mit bekanntem individuellen Merkmalsstatus sollten Studienteilnehmer eines von zwei Online-Würfelspielen spielen. Dabei wurde Betrugsverhalten als sensibles Merkmal experimentell induziert und anschlie-

ßend als starkes Außenkriterium verwendet. Das CWM zeigte dabei eine Spezifität zwischen 88% und 92%; die Spezifität einer direkten Selbstauskunft und zweier weiterer RRT-Varianten unterschied sich hingegen nicht signifikant von 100%. Die Sensitivität aller Verfahren, das heißt der Anteil der Merkmalsträger, die auch als solche identifiziert wurden, lag signifikant unter 100% (Höglinger & Jann, 2018).

Bislang ist noch unklar, warum die Spezifität der RRT nicht immer optimal ist. Fraglich ist insbesondere, ob das Problem der ungenügenden Spezifität nur in bestimmten Kontexten und nur für bestimmte RRT-Varianten auftritt oder über verschiedene Kontexte und über mehrere RRT-Varianten hinweg beobachtet werden kann. Höglinger und Diekmann (2017) vermuten, dass die Spezifität des CWM mit besserer Implementierung steigen könnte. In der vorliegenden Dissertation wurde mit Hilfe zweier starker Validierungen mit bekanntem individuellen Merkmalsstatus der Befragten untersucht, ob ein mangelndes Instruktionsverständnis ursächlich für die suboptimale Spezifität im CWM ist und daher ausführliche Instruktionen und Verständnisfragen die Spezifität im CWM steigern können (Experiment 3). Darüber hinaus wurde auch die Spezifität des CDM und des neu vorgeschlagenen CDTRM untersucht (Experiment 4).

2.1.3 Extended-Crosswise-Modell (ECWM)

Das ECWM (Heck et al., 2018) ist eine kürzlich vorgeschlagene Weiterentwicklung des CWM. Im ursprünglichen CWM wird implizit angenommen, dass alle Befragten die Instruktionen befolgen. Zwar gibt es im CWM keine selbstschützende Antwortoption, es wäre jedoch denkbar, dass Befragte systematisch eine der beiden Antwortoptionen bevorzugen, beispielsweise weil sie diese für weniger verfänglich halten. Dies wäre im CWM nicht detektierbar, würde jedoch die resultierenden Prävalenzschätzer systematisch verzerren und damit die Validität der Schätzung gefährden. Das ECWM soll eine solche systematische Präferenz einer Antwortoption detektierbar machen.

Im ECWM erhalten Befragte die Instruktionen des CWM und werden randomisiert einer von zwei unabhängigen Gruppen zugeteilt. Die sensible Aussage ist in beiden Gruppen identisch, die Randomisierungswahrscheinlichkeiten p_1 und p_2 beider Gruppen sind jedoch komplementär ($p_2 = 1 - p_1$). Im ECWM können zwei voneinander unabhängige Antworthäufigkeiten in den beiden Gruppen beobachtet werden; daher verfügt das ECWM über einen Freiheitsgrad und ist testbar. Da die sensible Aussage in beiden Gruppen identisch ist, sollten sich die Prävalenzschätzungen π_1 und π_2 für das sensible Merkmal, die in beiden Gruppen

resultieren, nur zufällig voneinander unterscheiden. Falls sich π_1 und π_2 signifikant voneinander unterscheiden, zeigt dies eine systematische Präferenz für eine der beiden Antwortoptionen an; in diesem Fall sind die Prävalenzschätzungen nicht vertrauenswürdig (Heck et al., 2018). Falls π_1 und π_2 sich nicht voneinander unterscheiden, liegt keine systematische Präferenz für eine der beiden Antwortoptionen vor und π_1 und π_2 können in einen übergeordneten Prävalenzschätzer π integriert werden, ohne gegenüber dem CWM an Effizienz zu verlieren (Heck et al., 2018). Das ECWM kann zwar eine systematische Präferenz der Befragten für eine der beiden Antwortoptionen aufdecken, jedoch keine anderen Formen der Instruktionsverweigerung, wie beispielsweise eine zufällige Auswahl der Antwortoptionen (Heck et al., 2018). Das ECWM wurde bereits in einer Studie angewandt (Heck et al., 2018). Da es sich jedoch um ein vergleichsweise neues Modell handelt, ist das ECWM bislang noch nicht im Vergleich zu einer direkten Selbstauskunft evaluiert worden. Diese Forschungslücke wurde in Experiment 2 der vorliegenden Dissertation geschlossen.

2.2 Randomized-Response-Technik mit Verweigererdetektion: Cheating-Detection-Modell (CDM)

Das CDM (Clark & Desharnais, 1998) ist eine Variante der RRT, die es ermöglicht, neben dem Anteil der Merkmalsträger in einer Stichprobe auch den Anteil der Befragten zu schätzen, die die Instruktionen missachten (Instruktionsverweigerer). Im CDM wird allen Befragten eine sensible Aussage dargeboten (beispielsweise „Ich habe schon einmal Kokain genommen“); die zugehörigen Antwortoptionen lauten „Ich stimme zu“ versus „Ich stimme *nicht* zu“. Mit Hilfe einer Randomisierungsprozedur werden Befragte mit der Wahrscheinlichkeit p (beispielsweise $p = 1/3$) instruiert, unabhängig von ihrem Merkmalsstatus mit „Ich stimme zu“ zu antworten und mit der Gegenwahrscheinlichkeit $1 - p$ dazu aufgefordert, wahrheitsgemäß zu der Aussage Stellung zu nehmen. Das CDM bietet ebenso wie das TRM eine klare selbstschützende Antwortoption („Ich stimme *nicht* zu“) und ist daher ein *asymmetrisches* Modell. Im CDM wird jedoch im Gegensatz zum TRM explizit berücksichtigt, dass Befragte die Instruktionen missachten und die selbstschützende Antwortoption auswählen können. Die Grundannahme des CDM ist, dass die Stichprobe in drei einander nicht überlappende Gruppen von Befragten unterteilt werden kann: Ehrliche Merkmalsträger (π), ehrliche Nicht-Merkmalsträger (β) und Instruktionsverweigerer, welche die selbstschützende Antwortoption wählen ($\gamma = 1 - \pi - \beta$). Über den Merkmalsstatus der Instruktionsverweigerer macht

das CDM keine Annahme; es ist daher möglich, dass alle, nur ein Teil oder keiner der Instruktionsverweigerer Merkmalsträger sind. Folglich kann die Prävalenz des sensiblen Merkmals nur innerhalb des Intervalls von π (unter der Annahme, dass kein Instruktionsverweigerer Merkmalsträger ist) bis $\pi + \gamma$ (unter der Annahme, dass alle Instruktionsverweigerer Merkmalsträger sind) geschätzt werden. Damit das CDM identifizierbar ist, muss es auf zwei unabhängige Stichproben angewandt werden, bei denen unterschiedliche Randomisierungswahrscheinlichkeiten p_1 und p_2 zum Einsatz kommen. Abbildung 3 stellt das CDM als Baummodell für eine der beiden Gruppen von Befragten dar.

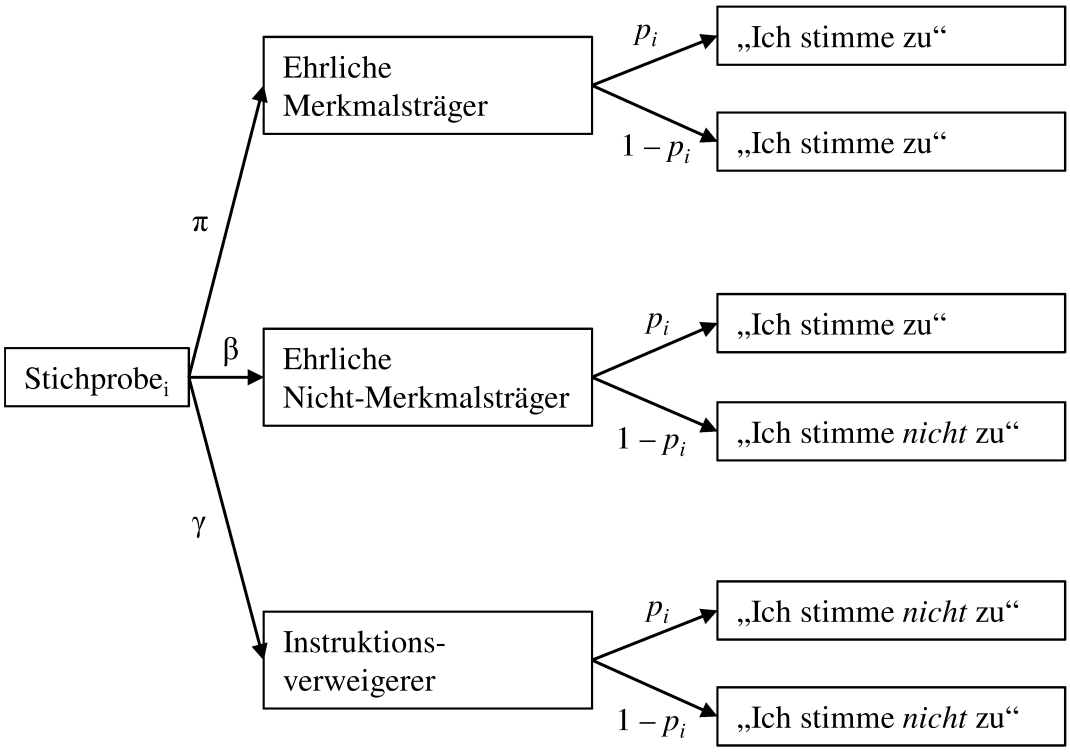


Abbildung 3: Baummodell des Cheating-Detection-Modells (Clark & Desharnais, 1998). Der Parameter π bezeichnet den Anteil der ehrlichen Merkmalsträger des sensiblen Merkmals, β bezeichnet den Anteil der ehrlichen Nicht-Merkmalsträger, γ den Anteil der Instruktionsverweigerer und p_i die Randomisierungswahrscheinlichkeit. Um ein identifizierbares Modell zu erhalten, sind zwei unabhängige Gruppen von Befragten mit zwei unterschiedlichen Randomisierungswahrscheinlichkeiten p_1 und p_2 nötig.

Maximum-Likelihood-Schätzer für π und β können nach Clark und Desharnais (1998) anhand folgender Formeln berechnet werden:

$$\hat{\pi} = \frac{p_2 \frac{n_1'}{n_1} - p_1 \frac{n_2'}{n_2}}{p_2 - p_1} \quad , \tag{4}$$

$$\hat{\beta} = \frac{\frac{n_2'}{n_2} - \frac{n_1'}{n_1}}{p_2 - p_1} \quad . \tag{5}$$

Hierbei entsprechen n_1 und n_2 der Stichprobengröße und n_1' und n_2' der absoluten Anzahl der „Ich stimme zu"-Antworten in den beiden Gruppen mit den Randomisierungswahrscheinlichkeiten p_1 und p_2. Der Schätzer $\hat{\gamma}$ kann dann berechnet werden als $\hat{\gamma} = 1 - \hat{\pi} - \hat{\beta}$.

Das CDM ist bereits einigen schwachen Validierungen unterzogen worden. In den meisten dieser Studien war der Anteil der Instruktionsverweigerer an den Befragten substantiell (Ostapczuk, Musch, & Moshagen, 2011). Im Vergleich zu einer direkten Selbstauskunft führte das CDM zu höheren und daher nach einem schwachen Validierungskriterium mutmaßlich valideren Prävalenzschätzungen für viele sozial unerwünschte Merkmale (Moshagen & Musch, 2012; Moshagen, Musch, Ostapczuk, & Zhao, 2010; Musch, Bröder, & Klauer, 2001; Ostapczuk et al., 2011). Für einige andere sensible Merkmale bot das CDM jedoch keinen Vorteil gegenüber einer direkten Selbstauskunft (Moshagen & Musch, 2012; Ostapczuk, Moshagen, Zhao, & Musch, 2009; Ostapczuk & Musch, 2011). Außerdem konnte gezeigt werden, dass die Instruktionen des CDM signifikant schlechter verständlich waren als die Instruktionen einer direkten Selbstauskunft (Hoffmann et al., 2017). Da bislang noch keine starke Validierung des CDM vorgenommen wurde, ist unklar, inwieweit das CDM tatsächlich die Validität von Selbstauskünften steigern kann und wie es um seine Spezifität und Sensitivität bestellt ist. Die vorliegende Dissertation präsentiert die Ergebnisse einer ersten starken Validierung des CDM mit bekanntem individuellen Merkmalsstatus der Befragten und kann so erstmalig auch die Spezifität und Sensitivität des Modells untersuchen (Experiment 4). Zudem wird in Experiment 4 mit dem CDTRM ein neues Modell vorgeschlagen, welches das Messmodell des CDM mit den leichter verständlichen Instruktionen des TRM kombiniert und so mutmaßlich die Validität der Messung zu steigern vermag.

2.3 Non-Randomized-Response-Technik mit Verweigererdetektion: Cheating-Detection-Triangular-Modell (CDTRM)

In der vorliegenden Dissertation wird ein neues Modell vorgeschlagen, das *Cheating-Detection-Triangular-Modell* (CDTRM). Das CDTRM vereint die leicht verständlichen In-

struktionen des TRM mit dem Mechanismus zur Verweigererdetektion des CDM. Die Idee des CDTRM ist, dass Befragte dieselben Instruktionen und Antwortoptionen erhalten wie im TRM. Im Gegensatz zum TRM und analog zum CDM wird im Messmodell des CDTRM explizit berücksichtigt, dass manche Befragte die Instruktionen missachten und die selbstschützende Antwortoption auswählen könnten. Der Anteil der ehrlichen Merkmalsträger (π), der ehrlichen Nicht-Merkmalsträger (β), und der Instruktionsverweigerer ($\gamma = 1 - \pi - \beta$) kann im CDTRM analog zum CDM geschätzt werden. Ebenso wie das CDM macht das CDTRM keine Annahme über den wahren Merkmalsstatus der Instruktionsverweigerer. Daher kann auch im CDTRM die Prävalenz des sensiblen Merkmals nur innerhalb des Intervalls von π (unter der Annahme, dass kein Instruktionsverweigerer Merkmalsträger ist) bis $\pi + \gamma$ (unter der Annahme, dass alle Instruktionsverweigerer Merkmalsträger sind) geschätzt werden. Um ein identifizierbares Modell zu erhalten, müssen die Instruktionen des TRM analog zum CDM auf zwei unabhängige, randomisiert zugewiesene Gruppen von Befragten angewandt werden, für die unterschiedliche Randomisierungswahrscheinlichkeiten p_1 und p_2 zum Einsatz kommen, beispielsweise $p_1 = .158$ und $p_2 = .842$. Abbildung 4 stellt das CDTRM als Baummodell für eine der beiden Gruppen von Befragten dar.

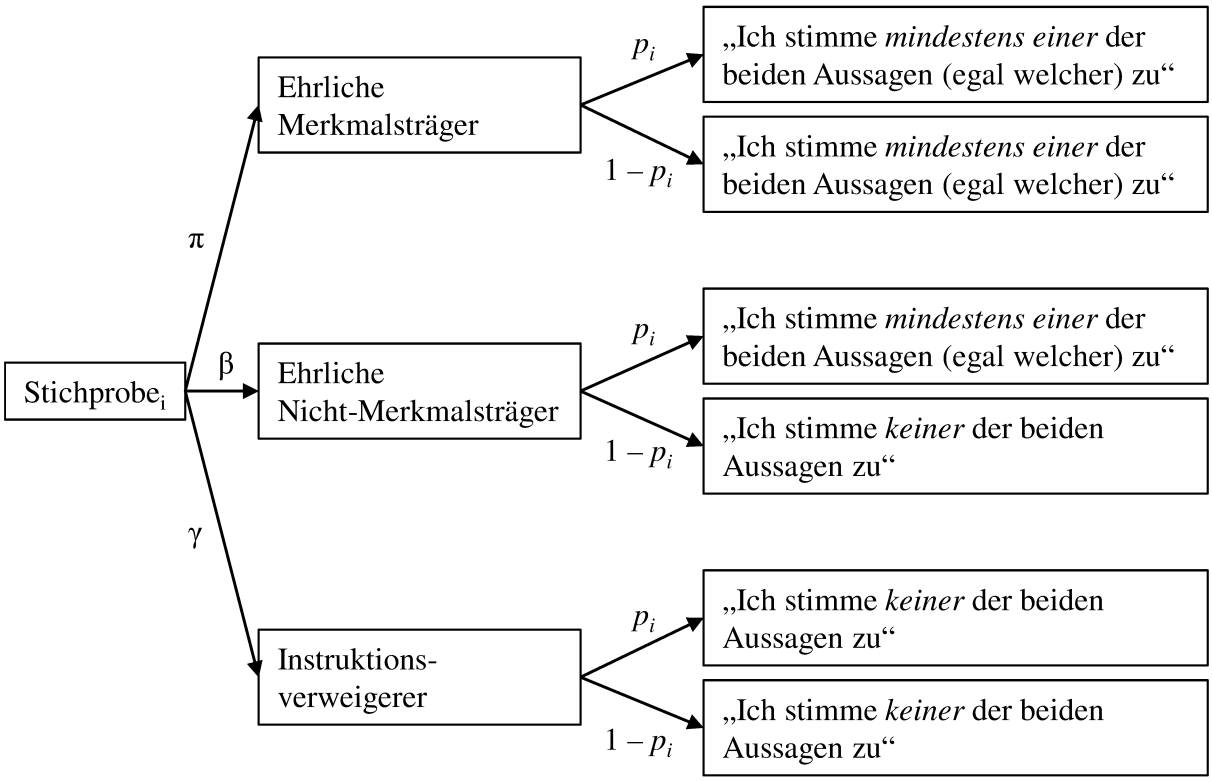


Abbildung 4: Baummodell des Cheating-Detection-Triangular-Modells. Der Parameter π bezeichnet den Anteil der ehrlichen Merkmalsträger des sensiblen Merkmals, β bezeichnet den Anteil der ehrlichen Nicht-Merkmalsträger, γ den Anteil der Instruktionsverweigerer und p_i die Randomisierungswahrscheinlichkeit. Um ein identifizierbares Modell zu erhalten, sind zwei unabhängige Gruppen von Befragten mit zwei unterschiedlichen Randomisierungswahrscheinlichkeiten p_1 und p_2 nötig.

Maximum-Likelihood-Schätzer für π und β können anhand der Gleichungen (4) und (5) bestimmt werden. Hierbei entsprechen n_1 und n_2 der Stichprobengröße und n_1' und n_2' der absoluten Anzahl der Antworten „Ich stimme *mindestens einer* der beiden Aussagen (egal welcher) zu“ in den beiden Gruppen mit den Randomisierungswahrscheinlichkeiten p_1 und p_2. Der Schätzer $\hat{\gamma}$ kann berechnet werden als $\hat{\gamma} = 1 - \hat{\pi} - \hat{\beta}$.

In Experiment 4 der vorliegenden Dissertation wurde das CDTRM erstmalig angewandt. In einer starken Validierung mit bekanntem individuellen Merkmalsstatus der Befragten wurde untersucht, ob das CDTRM validere Prävalenzschätzungen liefert als eine direkte Selbstauskunft. Zudem wurde die Validität des CDTRM mit der Validität des TRM und CDM verglichen.

3 Forschungsfragen

Im Rahmen der vorliegenden Dissertation sollte untersucht werden, ob verschiedene Varianten der RRT zu valideren Prävalenzschätzungen für sozial unerwünschte Merkmale führen als eine direkte Selbstauskunft. Weiterhin sollte untersucht werden, welche Faktoren die Validität der RRT erhöhen und wie verschiedene RRT hinsichtlich der Spezifität und Sensitivität abschneiden. Zum einen wurden schwache Validierungen durchgeführt, in denen Prävalenzschätzungen der RRT mit Prävalenzschätzungen aus einer direkten Selbstauskunft verglichen wurden (Experimente 1 und 2). Zum anderen wurden starke Validierungen durchgeführt, in denen die Prävalenz des sensiblen Merkmals auf Stichprobenebene bekannt war und als starkes Außenkriterium herangezogen werden konnte (Experimente 1, 3 und 4) beziehungsweise in denen der Merkmalsstatus der Befragten sogar auf individueller Ebene bekannt war (Experimente 3 und 4).

In Experiment 1 sollte geprüft werden, ob Antwortsymmetrie dazu beitragen kann, die Validität der RRT zu erhöhen. Hierzu wurde erstmalig eine schwache Validierung des symmetrischen CWM und des asymmetrischen TRM an derselben Stichprobe vorgenommen. Beide Modelle sollten zu höheren und daher nach einem schwachen Validierungskriterium potentiell valideren Prävalenzschätzungen für ein sozial unerwünschtes Merkmal führen als eine direkte Selbstauskunft. Sofern Antwortsymmetrie die Validität von Prävalenzschätzungen zu steigern vermag, sollte das symmetrische CWM dem asymmetrischen TRM überlegen sein. Anhand eines nicht-sensiblen Kontrollmerkmals mit bekannter Prävalenz konnte außerdem eine starke Validierung durchgeführt werden, um zu überprüfen, ob eines der beiden Modelle eine generelle Tendenz zur Über- oder Unterschätzung von Prävalenzen aufweist.

In Experiment 2 wurde untersucht, inwiefern eine Weiterentwicklung des CWM, das erst kürzlich vorgeschlagene ECWM, einer ersten schwachen Validierung standhält. Das ECWM sollte zu höheren Prävalenzschätzungen für ein sozial unerwünschtes Merkmal führen als eine direkte Selbstauskunft. Zudem konnte mit Hilfe eines Modelltests evaluiert werden, ob eine Voraussetzung für die Validität des (E)CWM erfüllt ist, die im ursprünglichen CWM nicht überprüft werden kann, und zwar, dass Befragte *keine* systematische Präferenz für eine der beiden Antwortoptionen aufweisen.

In Experiment 3 wurde mit Hilfe einer Validierung, in welcher der individuelle Merkmalsstatus der Befragten bekannt war, untersucht, ob sich die Spezifität des CWM durch die Implementierung detaillierter Instruktionen und Verständnisfragen steigern lässt. Hierzu

wurde das CWM mit ausführlichen Instruktionen und Verständnisfragen dem CWM mit kurzen Instruktionen ohne Verständnisfragen sowie einer direkten Selbstauskunft gegenübergestellt. Es sollte geklärt werden, welche Implementierung des CWM hinsichtlich der Validität, Spezifität und Sensitivität des Modells optimal ist.

In Experiment 4 wurde mit dem Cheating-Detection-Triangular-Modell (CDTRM) eine neue RRT-Variante vorgeschlagen, welche die leicht verständlichen Instruktionen des TRM mit dem Mechanismus zur Verweigererdetektion des CDM kombiniert. Es wurde eine erste starke Validierung des CDTRM, des TRM und des CDM mit bekanntem individuellen Merkmalsstatus der Befragten vorgenommen. Es sollte untersucht werden, welches der genannten Modelle die bekannte Prävalenz eines sensiblen Merkmals am genauesten zu schätzen vermag und welches der Modelle die höchste Sensitivität und Spezifität aufweist.

4 Einzelarbeiten

Im Folgenden werden die verwendete Methodik und die ermittelten Ergebnisse der vier Einzelarbeiten dieser Dissertation zusammengefasst und diskutiert. Alle in den Einzelarbeiten untersuchten Modelle wurden zur Auswertung als multinomiale Modelle (Batchelder, 1998; Batchelder & Riefer, 1999) formuliert (siehe auch Moshagen, Hilbig, & Musch, 2011; Moshagen, Musch, & Erdfelder, 2012; Ostapczuk et al., 2011). Hierbei bezeichnet der Parameter π jeweils die unbekannte Prävalenz des sensiblen Merkmals und der Parameter p die bekannte Randomisierungswahrscheinlichkeit. Prävalenzschätzungen wurden auf Basis der empirisch beobachteten Antworthäufigkeiten mit Hilfe des in der Software multiTree (Moshagen, 2010) implementierten Expectation-Maximization-Algorithmus (Dempster, Laird, & Rubin, 1977; Hu & Batchelder, 1994) gewonnen. Die Modellpassung wurde anhand der asymptotisch χ^2-verteilten Log-Likelihood-Statistik G^2 bestimmt. Um Parameter miteinander oder mit Konstanten zu vergleichen, wurde der Unterschied in der Modellpassung (ΔG^2) zwischen einem Basismodell ohne Parameterrestriktionen und einem restringierten Alternativmodell (beispielsweise mit $\pi_{CWM} = \pi_{TRM}$ oder $\pi_{CWM} = .22$) bestimmt. Um die geringere Effizienz der RRT im Vergleich zu einer direkten Selbstauskunft zu kompensieren (Ulrich et al., 2012), wurden den RRT-Bedingungen in den Experimenten 2 bis 4 doppelt so viele Teilnehmer zugewiesen wie den Bedingungen mit der direkten Selbstauskunft. In Experiment 1 wurde die Fragetechnik als Innersubjektfaktor manipuliert, so dass hier die Teilnehmerzahl in beiden Bedingungen notwendigerweise identisch war.

4.1 Experiment 1: Vergleich der Validität des Crosswise-Modells (CWM) und des Triangular-Modells (TRM)

In Experiment 1 wurden erstmalig die Prävalenzschätzungen zweier *Non-Randomized-Response-Techniken* (NRRT), des symmetrischen *Crosswise-Modells* (CWM) und des asymmetrischen *Triangular-Modells* (TRM), miteinander und mit der Prävalenzschätzung aus einer direkten Selbstauskunft verglichen. Das CWM und das TRM sind verwandte Modelle, die sich lediglich hinsichtlich ihrer Antwortsymmetrie unterscheiden: Das CWM bietet symmetrische, das TRM asymmetrische Antwortoptionen. Im Kontext anderer RRT-Varianten konnte bereits gezeigt werden, dass sich Antwortsymmetrie vorteilhaft auf die Befolgung der Instruktionen auswirkt (Ostapczuk, Moshagen, et al., 2009). In dem Maße, in dem Antwortsymmetrie auch die Validität von Prävalenzschätzungen zu steigern vermag, sollte das sym-

metrische CWM zu höheren und damit nach einem schwachen Validierungskriterium valideren Prävalenzschätzungen führen als das asymmetrische TRM. Bislang wurde jedoch die Validität des CWM und des TRM noch nie innerhalb derselben Stichprobe miteinander verglichen, sondern lediglich an zwei unterschiedlichen Stichproben mit demselben sensiblen Merkmal. In diesen Studien führte das TRM zu Prävalenzschätzern, die sich nicht signifikant von denen aus einer direkten Selbstauskunft unterschieden, während das CWM zu signifikant höheren Prävalenzschätzern führte als eine direkte Selbstauskunft (Jann et al., 2012; Jerke & Krumpal, 2013). Bei Vergleichen über unterschiedliche Stichproben hinweg bieten Unterschiede zwischen den Stichproben jedoch immer eine Alternativerklärung für die gefundenen Unterschiede in den Prävalenzschätzungen. Nur anhand einer echten experimentellen Manipulation kann geklärt werden, ob das symmetrische CWM im Vergleich zum asymmetrischen TRM zu valideren Prävalenzschätzern für sensible Merkmale führt. In Experiment 1 wurde daher ein erster experimenteller Vergleich beider Modelle innerhalb derselben Stichprobe durchgeführt. Hierbei wurden zwei sensible Merkmale mit unbekannter Prävalenz als weiche Validierungskriterien sowie ein nicht-sensibles Kontrollmerkmal mit bekannter Prävalenz als starkes Außenkriterium genutzt.

Als sensible Merkmale mit unbekannter Prävalenz dienten *Xenophobie* und die *Ablehnung einer weiteren Aufnahme von Geflüchteten*. Xenophobie bezeichnet die Angst vor oder die negative Einstellung gegenüber Menschen mit Migrationshintergrund und ist in Deutschland relativ weit verbreitet (Heitmeyer, 2012; Krumpal, 2012; Wagner & van Dick, 2001). Seit der sogenannten „Flüchtlingskrise“ von 2015 sind einige Deutsche insbesondere gegenüber Geflüchteten negativ eingestellt und lehnen deren weitere Aufnahme in Deutschland ab (Bertelsmann Stiftung, 2017). Viele Leute sind jedoch der Ansicht, dass fremdenfeindliche Einstellungen zentralen Werten vieler westlicher Gesellschaften widersprechen; eine öffentliche Äußerung solcher Einstellungen ist daher vermutlich sozial unerwünscht. Entsprechend konnte bereits gezeigt werden, dass indirekte Fragetechniken wie die RRT signifikant höhere und damit potentiell validere Prävalenzschätzungen für Xenophobie und die Ablehnung einer weiteren Aufnahme von Geflüchteten ergaben als eine direkte Selbstauskunft (Krumpal, 2012; Moshagen & Musch, 2012; Ostapczuk, Musch, & Moshagen, 2009). Xenophobie wurde in Experiment 1 mit der folgenden Aussage erfasst: „Es würde mich stören, wenn meine 20 Jahre alte Tochter eine Beziehung mit einem Türken eingehen würde“. Diese Aussage ist angelehnt an ein Item aus der Skala zur sozialen Distanz von Bogardus (1933) und wurde in gleicher oder ähnlicher Form bereits von Hoffmann und Musch (2016), Jimenez (1999), Ostapczuk, Musch, et al. (2009) sowie Silbermann und Hüsers (1995) eingesetzt. Die Ableh-

nung einer weiteren Aufnahme von Geflüchteten wurde mit der folgenden Aussage erfasst: „Ich finde, dass Deutschland bereits mehr als genug Flüchtlinge aufgenommen hat“. Sowohl das CWM als auch das TRM sollten zu Prävalenzschätzungen führen, die höher und daher vermutlich valider sind als Prävalenzschätzungen aus einer direkten Selbstauskunft. In dem Maße, in dem Antwortsymmetrie sich vorteilhaft auf die Validität der Prävalenzschätzungen auswirkt, sollte das CWM zudem zu valideren Schätzungen führen als das TRM.

Um zu testen, ob eines der Modelle eine generelle, methoden-inhärente Tendenz zur Über- oder Unterschätzung von Prävalenzen aufweist, wurde außerdem die bekannte Prävalenz eines nicht-sensiblen Kontrollmerkmals als starkes Außenkriterium herangezogen. Das nicht-sensible Kontrollmerkmal wurde anhand der folgenden Aussage erfasst: „Mein Nachname beginnt mit einem der folgenden Buchstaben: K, L, M“. In Deutschland hat dieses Merkmal nach Angaben des Statistischen Bundesamtes eine Prävalenz von 22% (Reinders, 1996). Sofern weder das CWM noch das TRM oder eine direkte Selbstauskunft zu einer generellen Über- oder Unterschätzung von Prävalenzen neigen, sollte die bekannte Prävalenz dieses nicht-sensiblen Kontrollmerkmals von allen drei Fragetechniken korrekt geschätzt werden.

Die Daten von 1382 Studierenden (60.06% weiblich) mit einem mittleren Alter von 21.40 Jahren (SD = 5.66) wurden analysiert. Befragte füllten auf dem Campus der Universität Düsseldorf vor Beginn der Vorlesungen einen einseitigen Papier-Bleistift-Fragebogen aus. Dieser enthielt Fragen nach Alter und Geschlecht der Befragten sowie insgesamt drei experimentelle Fragen. Die ersten beiden experimentellen Fragen bezogen sich auf die beiden sensiblen Merkmale Xenophobie und Ablehnung einer weiteren Aufnahme von Geflüchteten; die dritte experimentelle Frage bezog sich auf das nicht-sensible Kontrollmerkmal. Das Studiendesign war ein gekreuztes Innersubjekt-Design: Alle Teilnehmer beantworteten alle drei experimentellen Fragen mit jeweils einer der drei Fragetechniken (CWM, TRM, direkte Selbstauskunft). Die Reihenfolge der experimentellen Fragen war in jeder Bedingung gleich; die Reihenfolge der Fragetechnik war jedoch randomisiert, um Reihenfolgeeffekte zu vermeiden. Tabelle 1 zeigt die sechs verschiedenen Versionen des Fragebogens. Dieses Design erlaubte es, die Fragetechnik als Zwischensubjektfaktor für jedes sensible Merkmal zu manipulieren und zu analysieren. Nach der Datenerhebung wurden die Antworten der sechs experimentellen Bedingungen zusammengeführt, um Antworthäufigkeiten für alle drei Fragetechniken (CWM, TRM, direkte Selbstauskunft) und für jede experimentelle Frage (Xenophobie, Ablehnung einer weiteren Aufnahme von Geflüchteten, Anfangsbuchstabe des Nachnamens) zu erhalten. In der CWM- und TRM-Bedingung lautete die mit der ersten sensiblen Aussage

gepaarte nicht-sensible Aussage: „Mein Vater ist im November oder Dezember geboren“ (p = .158) und die mit der zweiten sensiblen Aussage gepaarte nicht-sensible Aussage: „Meine Mutter ist im November oder Dezember geboren“ (p = .158, Pötzsch, 2012).

Tabelle 1

Fragebogen-Versionen in Experiment 1.

	Fragebogen-Version					
Experimentelle Frage	1	2	3	4	5	6
Frage 1 (Xenophobie)	CWM	CWM	TRM	TRM	DS	DS
Frage 2 (Ablehnung einer weiteren Aufnahme von Geflüchteten)	TRM	DS	CWM	DS	CWM	TRM
Frage 3 (Anfangsbuchstabe des Nachnamens)	DS	TRM	DS	CWM	TRM	CWM

Anmerkung: CWM = Crosswise-Modell, TRM = Triangular-Modell, DS = direkte Selbstauskunft.

Die Prävalenzschätzung für Xenophobie war signifikant höher, wenn Xenophobie mit dem CWM ($\hat{\pi}$ = 31.65%, SE = 3.32%) statt mit einer direkten Selbstauskunft erfasst wurde ($\hat{\pi}$ = 15.45%, SE = 1.67%), $\Delta G^2(1) = 19.61$, $p < .001$. Die Prävalenzschätzung des TRM für Xenophobie ($\hat{\pi}$ = 20.05%, SE = 2.59%) überstieg die Schätzung der direkten Selbstauskunft zwar deskriptiv, aber nicht signifikant, $\Delta G^2(1) = 2.24$, p = .135. Das CWM ergab eine signifikant höhere Schätzung für Xenophobie als das TRM, $\Delta G^2(1) = 7.65$, p = .006. Für die Ablehnung einer weiteren Aufnahme von Geflüchteten ergab das CWM ($\hat{\pi}$ = 43.56%, SE = 3.38%) zwar eine deskriptiv, aber nicht signifikant höhere Schätzung als eine direkte Selbstauskunft ($\hat{\pi}$ = 36.73%, SE = 2.27%), $\Delta G^2(1) = 2.82$, p = .093. Auch die mit dem TRM ermittelte Prävalenzschätzung ($\hat{\pi}$ = 37.43%, SE = 2.75%) überstieg die aus einer direkten Selbstauskunft resultierende Schätzung nicht signifikant, $\Delta G^2(1) = 0.04$, p = .844. Die Schätzer aus CWM und TRM waren ebenfalls nicht signifikant voneinander verschieden, $\Delta G^2(1) = 1.99$, p = .159.

Für das nicht-sensible Kontrollmerkmal ergaben alle drei Fragetechniken Prävalenzschätzungen (CWM: $\hat{\pi}$ = 23.32%, SE = 3.16%; TRM: $\hat{\pi}$ = 22.22%, SE = 2.65%; DS: $\hat{\pi}$ = 24.35%, SE = 1.99%), die sich nicht signifikant von der bekannten Prävalenz (22%) des Kontrollmerkmals unterschieden; CWM vs. bekannte Prävalenz: $\Delta G^2(1) = 0.18$, p = .675;

TRM vs. bekannte Prävalenz: $\Delta G^2(1) = 0.01$, $p = .935$; direkte Selbstauskunft vs. bekannte Prävalenz: $\Delta G^2(1) = 1.46$, $p = .227$.

Zusammenfassend zeigte sich, dass die Prävalenz von *Xenophobie* durch eine direkte Selbstauskunft vermutlich unterschätzt wurde. Das CWM ergab eine signifikant höhere und daher nach einem schwachen Validierungskriterium potentiell validere Schätzung als eine direkte Selbstauskunft. Im Gegensatz zum CWM ergab das TRM keine höhere Prävalenzschätzung als eine direkte Selbstauskunft. Zudem führte das symmetrische CWM zu einer signifikant höheren Schätzung als das asymmetrische TRM. Dieses Ergebnis lässt sich nach einem schwachen Validierungskriterium als Beleg dafür interpretieren, dass das CWM zu valideren Prävalenzschätzungen führte als das TRM. Die Abwesenheit einer selbstschützenden Antwortoption scheint sich also positiv auf die Validität der ermittelten Schätzungen auszuwirken (siehe auch Ostapczuk, Moshagen, et al., 2009). Bezüglich des Merkmals *Ablehnung einer weiteren Aufnahme von Geflüchteten* fand sich jedoch kein Vorteil des CWM gegenüber einer direkten Selbstauskunft oder gegenüber dem TRM. Eine mögliche Erklärung für diesen Befund ist, dass das Merkmal als zu wenig sensibel empfunden wurde, denn der Einsatz der RRT ist vorteilhafter, wenn das untersuchte Merkmal sensibler ist (Lensvelt-Mulders et al., 2005). Dafür spricht, dass der entsprechenden Aussage bereits in einer direkten Selbstauskunft 37% der Befragten zustimmten, während es bei Xenophobie nur 15% waren. Insgesamt scheint das symmetrische CWM zu valideren Prävalenzschätzungen zu führen als das asymmetrische TRM; dies scheint insbesondere dann der Fall zu sein, wenn das untersuchte Merkmal sehr sensibel ist.

4.2 Experiment 2: Erste schwache Validierung des Extended-Crosswise-Modells (ECWM)

Nach der erfolgreichen Validierung des CWM in Experiment 1 wurde in Experiment 2 die kürzlich vorgeschlagene Erweiterung des CWM, das *Extended-Crosswise-Modell* (ECWM; Heck et al., 2018), erstmalig schwach validiert. Das ECWM erlaubt im Gegensatz zum CWM ein Aufdecken von systematischen Antwortpräferenzen bei den Befragten, bietet dabei jedoch dieselbe statistische Effizienz wie das CWM. Die Validität des ECWM wurde bislang allerdings noch nicht im Vergleich zu einer direkten Selbstauskunft evaluiert.

In Experiment 2 wurde Islamophobie als sensibles Merkmal genutzt. Islamophobie bezeichnet die Angst vor oder negative Einstellung gegenüber dem Islam und Menschen muslimischen Glaubens. In Deutschland sind relativ viele Menschen dem Islam gegenüber

negativ eingestellt; dabei wird insbesondere das muslimische Frauenbild häufig als problematisch angesehen (Petersen, 2012; Pollack, 2014; Zick, Küpper, & Hövermann, 2011). Islamophobie ist ein mit Xenophobie verwandtes Konstrukt (Heitmeyer, 2012); vermutlich ist eine öffentliche Äußerung islamophober Einstellungen daher ebenfalls sozial unerwünscht. Entsprechend konnte auch für Islamophobie bereits gezeigt werden, dass Prävalenzschätzer, die auf einer direkten Selbstauskunft beruhen, mutmaßlich durch soziale Erwünschtheit verzerrt sind (Hoffmann & Musch, 2016; Krumpal, 2012; Ostapczuk, Musch, et al., 2009). Die in Experiment 2 verwendete sensible Aussage zu Islamophobie lautete: „Viele muslimische Studenten verhalten sich frauenverachtend". Das ECWM sollte zu höheren und insofern nach einem schwachen Validierungskriterium potentiell valideren Prävalenzschätzungen für Islamophobie führen als eine direkte Selbstauskunft. Außerdem kann das ECWM anzeigen, ob Befragte systematisch die Instruktionen missachtetet und eine Präferenz für eine der beiden Antwortoptionen gezeigt haben.

In Experiment 2 wurden 1361 Studierende (55.69% weiblich) der Universität Düsseldorf befragt. Von diesen waren 55.55% jünger als 20 Jahre, 40.71% waren zwischen 20 und 29 Jahre alt und 3.75% waren mindestens 30 Jahre alt. Es wurden nur die Daten von Nicht-Muslimen ausgewertet, da Vorurteile gegenüber Muslimen unter Nicht-Muslimen untersucht werden sollten. Analog zum Vorgehen in Experiment 1 füllten die Befragten in Experiment 2 vor Beginn der Vorlesungen einen einseitigen Papier-Bleistift-Fragebogen aus. Dieser enthielt neben Fragen zu Geschlecht, Alter und Religion der Befragten die sensible Frage zur Islamophobie. Die Fragetechnik wurde hierbei als Zwischensubjektfaktor mit den beiden Stufen a) direkte Selbstauskunft und b) ECWM manipuliert. Innerhalb der ECWM-Bedingung lautete für die Hälfte der Befragten die nicht-sensible Aussage „Mein Vater ist im November oder Dezember geboren" (Randomisierungswahrscheinlichkeit p_1 = .158; Pötzsch, 2012) und für die andere Hälfte der Befragten „Mein Vater ist zwischen Januar und Oktober geboren" (p_2 = .842).

Das ECWM passte gut auf die ermittelten Daten, $G^2(1) = 0.10$, $p = .756$. Dieser Befund deutet darauf hin, dass sich die Prävalenzschätzer π_1 und π_2 aus den beiden ECWM-Gruppen mit den unterschiedlichen Randomisierungswahrscheinlichkeiten nicht signifikant voneinander unterschieden haben und dass Befragte daher keine systematische Präferenz für eine der beiden Antwortoptionen gezeigt haben. Die Prävalenzschätzer beider Gruppen konnten daher in einen gemeinsamen Schätzer des ECWM integriert werden. Dieser Schätzer war signifikant höher ($\hat{\pi}$ = 21.19%, SE = 2.23%) als der Schätzer aus einer direkten Selbstauskunft ($\hat{\pi}$ = 10.89%, SE = 1.47%), $\Delta G^2(1) = 14.69$, $p < .001$.

Zusammenfassend wurde die Prävalenz von Islamophobie von einer direkten Selbstauskunft vermutlich unterschätzt; das ECWM lieferte jedoch eine höhere und daher nach einem schwachen Validierungskriterium mutmaßlich validere Prävalenzschätzung. Im Gegensatz zu dem verwandten CWM erlaubte das ECWM einen Modelltest der bislang impliziten Annahme, dass Befragte die Instruktionen befolgen und keine der Antwortoptionen systematisch bevorzugen. Insofern ist das ECWM gegenüber dem CWM zu präferieren, da es einen Test dieser Modellannahme ohne Verlust von Effizienz erlaubt.

4.3 Experiment 3: Untersuchung zur Verbesserung der Validität des Crosswise-Modells (CWM)

Viele Validierungsstudien haben gezeigt, dass das CWM eine vielversprechende Fragetechnik darstellt, mit deren Hilfe potentiell validere Prävalenzschätzungen für sozial unerwünschte Merkmale ermittelt werden können als mit einer direkten Selbstauskunft. Für diese positiven Evaluationen des CWM könnte neben den einfachen Instruktionen des Modells auch dessen Antwortsymmetrie verantwortlich sein (siehe Experiment 1). Experiment 2 demonstrierte zudem, dass das ECWM eine empfehlenswerte Weiterentwicklung des CWM darstellt, da das ECWM zu einer potentiell valideren Prävalenzschätzung führte als eine direkte Selbstauskunft und darüber hinaus einen Test einer zentralen Modellannahme ohne Verlust von Effizienz erlaubt. In zwei kürzlich veröffentlichten Studien wurde jedoch die Validität des CWM in Frage gestellt. Es konnte gezeigt werden, dass das CWM eine signifikant von 100% verschiedene Spezifität aufwies, das heißt, dass einige der Nicht-Merkmalsträger vom CWM nicht als solche erkannt, sondern fälschlicherweise als Merkmalsträger klassifiziert wurden (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018). Bislang ist unklar, warum und unter welchen Bedingungen die Spezifität im CWM nicht optimal ist und wie die Spezifität im CWM erhöht werden kann. Es scheint unwahrscheinlich, dass Nicht-Merkmalsträger absichtlich so antworten wie Merkmalsträger; daher liegt die Vermutung nahe, dass ein Missverständnis der Instruktionen verantwortlich für die mangelnde Spezifität sein könnte. Zwar war das CWM in einer Studie verständlicher als andere indirekte Fragetechniken, allerdings traten bei einigen der Befragten noch Verständnisprobleme auf (Hoffmann et al., 2017). In Experiment 3 sollte daher untersucht werden, ob ein verbessertes Instruktionsverständnis die Spezifität im CWM zu erhöhen vermag. Gleichzeitig sollte auch die Sensitivität des CWM untersucht werden, also der Anteil der Merkmalsträger, die auch als solche identifiziert werden. Zudem wurden

die Validität, Spezifität und Sensitivität des CWM und einer direkten Selbstauskunft verglichen.

Die Bestimmung von Spezifität und Sensitivität ist nur anhand einer relativ aufwändigen starken Validierung mit Daten zum individuellen Merkmalsstatus der Befragten möglich. Um den Merkmalsstatus individueller Befragter bestimmen zu können, wurde in Experiment 3 ein Online-Anagramm-Paradigma verwendet, durch welches ein sensibles Merkmal experimentell induziert werden konnte (Hoffmann et al., 2015). Befragte sollten hierbei zunächst drei durcheinander gebrachte Buchstabenreihenfolgen, sogenannte *Anagramme*, lösen. Anschließend hatten sie die Möglichkeit, bei der Angabe der Anzahl gelöster Anagramme zu übertreiben, das heißt zu betrügen, um an einer Verlosung von drei Geldpreisen teilnehmen zu können. Das sensible Merkmal *Betrug im Anagramm-Paradigma* wurde anschließend mit der folgenden Aussage abgefragt: „Ich habe angegeben, mehr Anagramme gelöst zu haben, als ich tatsächlich gelöst habe." Die Spezifität wurde berechnet als der Anteil der mit der Fragetechnik (direkte Selbstauskunft, CWM) geschätzten Nicht-Merkmalsträger ($1 - \pi$) innerhalb der Teilstichprobe ehrlicher Teilnehmer; die Sensitivität wurde berechnet als der Anteil der mit der Fragetechnik geschätzten Merkmalsträger (π) innerhalb der Teilstichprobe der Betrüger.

Um zu klären, welche Implementierung des CWM hinsichtlich der Validität, Spezifität und Sensitivität optimal ist, wurde das CWM in zwei Versionen umgesetzt: Einmal mit detaillierten Instruktionen und Verständnisfragen (*CWM detailliert*) und einmal mit kurzen Instruktionen ohne Verständnisfragen (*CWM kurz*). In beiden CWM-Bedingungen lautete die mit der sensiblen Aussage gepaarte nicht-sensible Aussage: „Ich bin im November oder Dezember geboren" ($p = .158$, Pötzsch, 2012). Befragte in der Bedingung *CWM kurz* erhielten eine kurze Instruktion dazu, wie sie die sensible Aussage beantworten sollten. Befragte in der Bedingung *CWM detailliert* erhielten darüber hinaus Informationen zur Auswertung der Frage durch die Versuchsleiter sowie insgesamt sechs Verständnisfragen. Die Verständnisfragen sollten erfassen, ob die Befragten verstanden haben, wie sie im CWM antworten sollen und wie das CWM die Vertraulichkeit individueller Antworten garantiert. Anschließend erhielten die Befragten Rückmeldungen zu ihren Antworten und falsch beantwortete Verständnisfragen wurden so lange wiederholt präsentiert, bis sie entweder korrekt oder maximal drei Mal falsch beantwortet wurden. Das Bildungsniveau der Befragten wurde als Moderatorvariable untersucht, da frühere Untersuchungen ergeben haben, dass Befragte mit niedrigem Bildungsniveau mehr Probleme haben, die Instruktionen indirekter Fragetechniken zu verstehen (Hoffmann et al., 2017).

Insgesamt wurden in Experiment 3 die Daten von 2713 deutschen Muttersprachlern (50.31% weiblich) einer Online-Stichprobe analysiert. Ungefähr die Hälfte der Befragten (50.02%) hatte ein hohes Bildungsniveau (mindestens Abitur), die andere Hälfte hatte ein niedrigeres Bildungsniveau (maximal mittlere Reife). Alle Befragten waren zwischen 30 und 40 Jahre alt ($M = 34.73$, $SD = 3.15$), um eine Konfundierung von Alter und Bildung auszuschließen und so die Teststärke zu erhöhen. Beide CWM-Bedingungen ergaben signifikant höhere Prävalenzschätzer für das sensible Merkmal *Betrug im Anagrammparadigma* (CWM detailliert: $\hat{\pi} = 25.48\%$, $SE = 2.21\%$; CWM kurz: $\hat{\pi} = 30.78\%$, $SE = 2.07\%$) als eine direkte Selbstauskunft ($\hat{\pi} = 11.79\%$, $SE = 1.34\%$); CWM detailliert vs. direkte Selbstauskunft: $\Delta G^2(1) = 27.94$, $p < .001$; CWM kurz vs. direkte Selbstauskunft: $\Delta G^2(1) = 56.74$, $p < .001$. Die Prävalenzschätzer aus beiden CWM-Bedingungen unterschieden sich nicht signifikant voneinander, $\Delta G^2(1) = 3.06$, $p = .080$. Die bekannte Prävalenz des sensiblen Merkmals (direkte Selbstauskunft: 58.93%, CWM kurz: 56.70%, CWM detailliert: 59.26%) wurde in allen Bedingungen signifikant unterschätzt; CWM detailliert vs. bekannte Prävalenz: $\Delta G^2(1) = 210.75$, $p < .001$; CWM kurz vs. bekannte Prävalenz: $\Delta G^2(1) = 147.47$, $p < .001$; direkte Selbstauskunft vs. bekannte Prävalenz: $\Delta G^2(1) = 558.69$, $p < .001$.

In allen drei Bedingungen war die Spezifität signifikant geringer als 100%. Die geringste Spezifität trat zwar in den beiden CWM-Bedingungen auf (CWM detailliert: 86.92%, $SE = 3.17\%$; CWM kurz: 85.68%, $SE = 2.84\%$), allerdings fand sich auch in einer direkten Selbstauskunft mit 97.47% ($SE = 1.02\%$) eine signifikant von 100% verschiedene Spezifität. Die Spezifität unterschied sich nicht signifikant zwischen den beiden CWM-Bedingungen; die Spezifität in der direkten Selbstauskunft war signifikant höher als in den beiden CWM-Bedingungen. Die Sensitivität war mit 18.23% ($SE = 2.09\%$) bei einer direkten Selbstauskunft am geringsten, war jedoch auch in den CWM-Bedingungen signifikant geringer als 100% (CWM detailliert: 34.01%, $SE = 2.97\%$; CWM kurz: 43.35%, $SE = 2.83\%$). Die Sensitivität unterschied sich nicht signifikant zwischen den beiden CWM-Bedingungen; die Sensitivität in der direkten Selbstauskunft war signifikant geringer als in den beiden CWM-Bedingungen. Tabelle 2 gibt die Inferenzstatistik der Vergleiche der Sensitivität und Spezifität zwischen den Bedingungen sowie gegen 100% an.

Tabelle 2

Inferenzstatistische Vergleiche der Spezifität und Sensitivität in den verschiedenen Bedingungen.

				Modellpassung	
	Parameter 1	Parameter 2	|Differenz|	$\Delta G^2(1)$	*p*
Spezifität					
$Spez_{CWM\ detailliert} = 100\%$	86.92	100.00	13.08	20.99	< .001*
$Spez_{CWM\ kurz} = 100\%$	85.68	100.00	14.32	31.70	< .001*
$Spez_{DS} = 100\%$	97.47	100.00	2.53	165.09	< .001*
$Spez_{CWM\ detailliert} = Spez_{CWM\ kurz}$	86.92	85.68	1.24	0.08	= .771
$Spez_{CWM\ detailliert} = Spez_{DS}$	86.92	97.47	10.55	10.82	= .001*
$Spez_{CWM\ kurz} = Spez_{DS}$	85.68	97.47	11.79	15.86	< .001*
Sensitivität					
$Sens_{CWM\ detailliert} = 100\%$	34.01	100.00	365.99	601.97	< .001*
$Sens_{CWM\ kurz} = 100\%$	43.44	100.00	56.65	522.21	< .001*
$Sens_{DS} = 100\%$	18.23	100.00	81.77	9918.94	< .001*
$Sens_{CWM\ detailliert} = Sens_{CWM\ kurz}$	34.01	43.44	9.34	5.15	= .023*
$Sens_{CWM\ detailliert} = Sens_{DS}$	34.01	18.24	15.78	18.47	< .001*
$Sens_{CWM\ kurz} = Sens_{DS}$	43.44	18.24	25.12	47.71	< .001*

Anmerkung: CWM detailliert = Crosswise-Modell mit detaillierten Instruktionen und Verständnisfragen, CWM kurz = Crosswise-Modell mit kurzen Instruktionen ohne Verständnisfragen, DS = Direkte Selbstauskunft.

Abbildung 5 bildet die Spezifität und Sensitivität nach Bedingung und Bildungsniveau ab. Ein Vergleich von Befragten mit hohem und niedrigem Bildungsniveau zeigte, dass die Spezifität in beiden CWM-Bedingungen bei Befragten mit hohem Bildungsniveau signifikant höher war als bei Befragten mit niedrigem Bildungsniveau; CWM kurz: $\Delta G^2(1) = 4.33$, $p = .038$; CWM detailliert: $\Delta G^2(1) = 8.13$, $p = .004$. Für die direkte Selbstauskunft war diese Tendenz nicht signifikant, $\Delta G^2(1) = 3.23$, $p = .072$. Die Sensitivität wurde hingegen nicht

durch das Bildungsniveau moderiert; direkte Selbstauskunft: $\Delta G^2(1) = 0.32$, $p < .574$; CWM kurz: $\Delta G^2(1) = 1.51$, $p = .220$; CWM detailliert: $\Delta G^2(1) = 0.43$, $p = .511$.

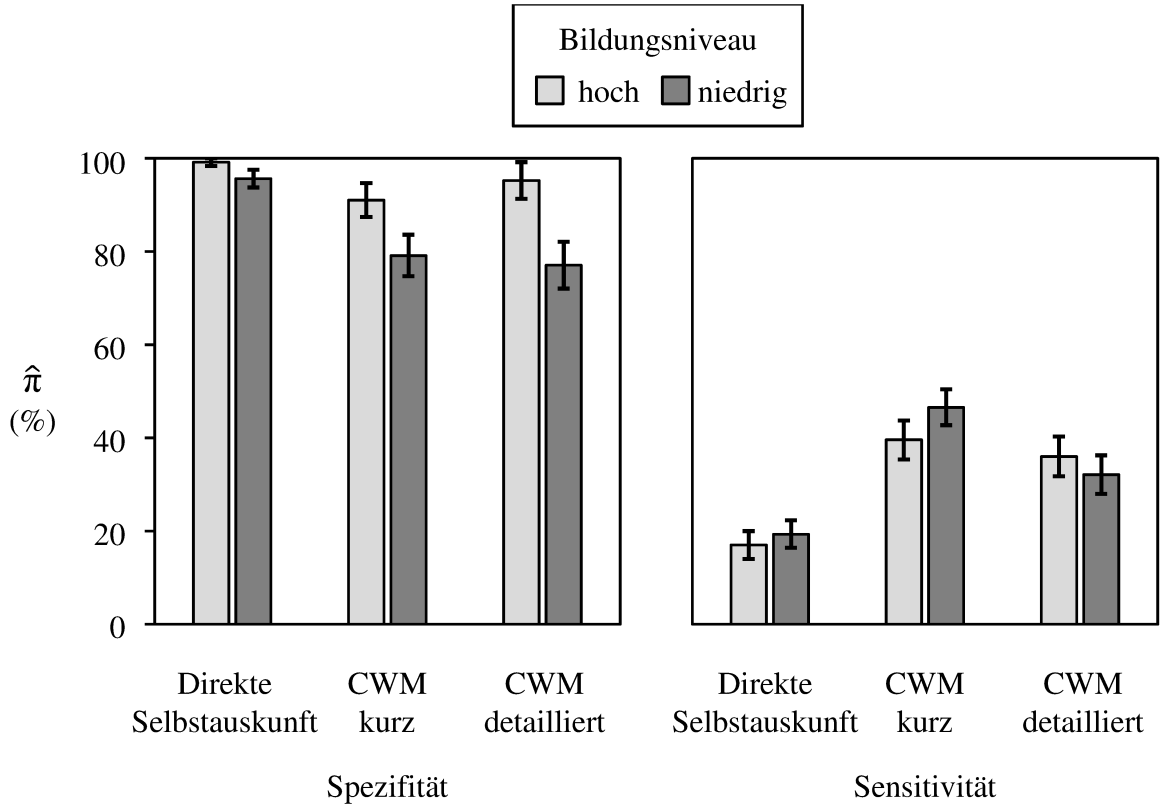


Abbildung 5: Spezifität und Sensitivität als Funktion der Bedingung und des Bildungsniveaus. CWM kurz = Crosswise-Modell mit kurzen Instruktionen ohne Verständnisfragen, CWM detailliert = Crosswise-Modell mit detaillierten Instruktionen und Verständnisfragen. Die Fehlerbalken repräsentieren den Standardfehler der Prävalenzschätzungen.

Um herauszufinden, ob die Verständnisfragen in der Bedingung *CWM detailliert* zu valideren Prävalenzschätzungen beitrugen, wurden explorativ die Analysen der Spezifität und Sensitivität nur innerhalb der Substichprobe der Befragten berechnet, die alle Verständnisfragen in einer der drei Runden korrekt beantwortet und daher vermutlich die Instruktionen verstanden haben (n = 669; davon 390 Befragte mit hohem und 279 Befragte mit niedrigem Bildungsniveau). Die Spezifität unter Befragten mit hohem Bildungsniveau stieg von 95.22% (SE = 3.92%) in der Gesamtstichprobe auf 100.00% (SE = 4.35%) in der Substichprobe der Befragten, die alle Verständnisfragen in einer der drei Runden korrekt beantwortet haben. Unter den Befragten mit niedrigem Bildungsniveau sank die Spezifität jedoch von 77.06% (SE = 5.05%) auf 74.83% (SE = 6.68%). Gleichzeitig verringerte sich auch die Sensitivität von 36.00% (SE = 4.27%, Befragte mit hohem Bildungsniveau) beziehungsweise 32.10%

(*SE* = 4.13%, Befragte mit niedrigem Bildungsniveau) in der Gesamtstichprobe auf 29.21% (*SE* = 4.75%, Befragte mit hohem Bildungsniveau) beziehungsweise 26.76% (*SE* = 5.27%, Befragte mit niedrigem Bildungsniveau) in der Substichprobe der Befragten, die alle Verständnisfragen in einer der drei Runden korrekt beantwortet haben.

Zusammenfassend wurde in Experiment 3 mit Hilfe einer starken Validierung mit Daten zum individuellen Merkmalsstatus der Befragten erstmalig untersucht, wie die suboptimale Spezifität des CWM erhöht werden kann. Es zeigte sich, dass die Spezifität des CWM deutlich höher war, wenn das Bildungsniveau der Befragten hoch statt niedrig war. Unter Anwendung ausführlicher Instruktionen und Verständnisfragen konnte die Spezifität bei Befragten mit hohem Bildungsniveau sogar auf 100% erhöht werden. Die Spezifität einer direkten Selbstauskunft lag jedoch auch signifikant unter 100% – mangelnde Spezifität scheint daher nicht nur ein Nachteil des CWM zu sein, sondern von Selbstauskünften zu sensiblen Merkmalen im Allgemeinen. Des Weiteren wurde in Experiment 3 gezeigt, dass auch die Sensitivität bei Befragungen mit dem CWM und einer direkten Selbstauskunft suboptimal und sogar deutlich niedriger war als die Spezifität (siehe auch Höglinger & Diekmann, 2017; Höglinger & Jann, 2018). Die Sensitivität fiel jedoch in den CWM-Bedingungen signifikant höher aus als in einer direkten Selbstauskunft. Dieser Befund stimmt mit dem ursprünglichen Zweck indirekter Fragetechniken wie dem CWM überein, den Anteil der Merkmalsträger, welche auch als solche erkannt werden, das heißt die Sensitivität, zu erhöhen.

Insgesamt unterschätzten die resultierenden Prävalenzschätzungen die bekannte Prävalenz des sensiblen Merkmals immer noch deutlich. Somit stimmen die Befunde aus Experiment 3 mit den Befunden zweier Meta-Analysen zur RRT überein (Lensvelt-Mulders et al., 2005): Prävalenzschätzungen für sensible Merkmale, die mit Hilfe einer RRT wie dem CWM ermittelt werden, sind bei weitem nicht perfekt; RRT-Schätzungen liegen jedoch deutlich näher an der bekannten Prävalenz sensibler Merkmale als Schätzungen aus einer direkten Selbstauskunft. In Anbetracht der Ergebnisse von Experiment 3 kann das CWM allerdings derzeit nur dann für die Anwendung empfohlen werden, wenn die Stichprobe aus Befragten mit einem hohen Bildungsniveau besteht und wenn detaillierte Instruktionen und Verständnisfragen verwendet werden. Im Hinblick auf Stichproben, die aus Befragten mit einem niedrigen Bildungsniveau bestehen, stellen die vorliegenden Ergebnisse jedoch die Anwendbarkeit des CWM in seiner derzeitigen Implementierung in Frage. Um auch von weniger gebildeten Befragten valide Antworten zu erhalten, ist es daher dringend notwendig, Wege zu finden, das Instruktionsverständnis auch innerhalb dieser Gruppe der Befragten zu verbessern. Darüber

hinaus verdeutlicht Experiment 3 die Bedeutung starker Validierungen mit bekanntem individuellen Merkmalsstatus der Befragten, da nur Daten auf individueller Ebene eine Berechnung der Spezifität und Sensitivität erlauben und damit eine umfassende Bewertung der Validität einer Methode ermöglichen (Höglinger & Jann, 2018; Umesh & Peterson, 1991).

4.4 Experiment 4: Validierung des neuen Cheating-Detection-Triangular-Modells (CDTRM)

In Experiment 1 zeigte sich in einer schwachen Validierung, dass das TRM Prävalenzschätzungen ergab, die sich nicht von denen aus einer direkten Selbstauskunft unterschieden. Ein möglicher Grund für diesen negativen Befund könnte die fehlende Antwortsymmetrie des TRM sein. Wenn Befragte im TRM die Instruktionen missachten und die selbstschützende Antwortoption wählen, kann dies allerdings im TRM nicht entdeckt werden. Daher wurde in Experiment 4 ein neues Modell getestet, das *Cheating-Detection-Triangular-Modell* (CDTRM), welches die leicht verständlichen Instruktionen des TRM mit dem Mechanismus zur Verweigererdetektion des CDM vereint. In einer starken Validierung mit bekanntem individuellen Merkmalsstatus der Befragten sollte erstmalig die Validität, Spezifität und Sensitivität des CDTRM sowie des TRM und des CDM untersucht werden. Die Prävalenzschätzungen dieser Modelle sollten näher an der bekannten Prävalenz eines sozial unerwünschten Merkmals liegen als die Prävalenzschätzung aus einer direkten Selbstauskunft. Außerdem sollte das CDTRM validere Schätzungen ergeben als das CDM und das TRM, da das CDTRM die Stärken der beiden anderen Modelle kombiniert. Es wurde außerdem untersucht, wie die Modelle hinsichtlich ihrer Spezifität und Sensitivität abschneiden.

In Experiment 4 wurden die Daten von 2787 Muttersprachlern (57.73% weiblich) einer Online-Stichprobe analysiert. Analog zu Experiment 3 wurde in Experiment 4 das Online-Anagramm-Paradigma genutzt, um ein sensibles Merkmal mit bekanntem individuellen Merkmalsstatus der Befragten zu induzieren (Hoffmann et al., 2015). Anschließend sollten die Befragten Stellung nehmen zu der sensiblen Aussage „Ich habe angegeben, mehr Anagramme gelöst zu haben, als ich tatsächlich gelöst habe“. Die Fragetechnik wurde als Zwischensubjektfaktor mit den Stufen a) direkte Selbstauskunft, b) TRM und c) CDM manipuliert. Da die Prävalenzschätzung für das CDTRM ebenfalls aus der Bedingung mit den Instruktionen des TRM gewonnen werden sollte, wurde das TRM genau wie das CDM zwei Gruppen von Befragten mit unterschiedlichen Randomisierungswahrscheinlichkeiten vorgegeben. Die nicht-sensiblen Aussagen in den indirekten Fragebedingungen lauteten in der

jeweiligen Gruppe mit der Randomisierungswahrscheinlichkeit $p_1 = .158$ „Ich bin im November oder Dezember geboren“ und in der jeweiligen Gruppe mit der Randomisierungswahrscheinlichkeit $p_2 = .842$ „Ich bin zwischen Januar und Oktober geboren“ (Pötzsch, 2012). Den Befragten in den indirekten Fragebedingungen wurden vier der in Experiment 3 entwickelten Verständnisfragen gestellt. Befragte erhielten Rückmeldungen zu ihren Antworten und falsch beantwortete Verständnisfragen wurden so lange wiederholt präsentiert, bis sie entweder korrekt oder mindestens drei Mal falsch beantwortet wurden. Im Anschluss an die Präsentation der sensiblen Frage sollten die Befragten die Fragetechnik anhand folgender neun Aussagen evaluieren:

- „Ich fand die Frage verständlich“
- „Die Frage hat die Vertraulichkeit meiner Antwort gewährleistet“
- „Die Art und Weise, wie die Frage gestellt wurde, war interessant“
- „Die Art und Weise, wie die Frage gestellt wurde, war sinnvoll“
- „Die Frage war umständlich gestellt“ (umgekehrt kodiert)
- „Ich habe alle Erläuterungen sorgfältig gelesen und befolgt“
- „Mir war völlig klar, was ich antworten musste“
- „Ich fühlte mich durch die Art der Fragestellung überfordert“ (umgekehrt kodiert)
- „Ich habe einfach irgendetwas angekreuzt“ (umgekehrt kodiert).

Die Zustimmung zu diesen Aussagen wurde auf einer siebenstufigen Likert-Skala von 1 (*trifft gar nicht zu*) bis 7 (*trifft vollkommen zu*) angegeben.

Um Prävalenzschätzungen für das CDTRM zu erhalten, mussten die Instruktionen des TRM zwei Gruppen von Befragten mit unterschiedlichen Randomisierungswahrscheinlichkeiten vorgegeben werden. Dies hatte zur Folge, dass auch das TRM in zwei Gruppen mit unterschiedlichen Randomisierungswahrscheinlichkeiten vorlag. Diese Erweiterung des TRM erlaubte einen Vergleich der Prävalenzschätzer beider Gruppen und somit einen ersten empirischen Test der dem TRM inhärenten Annahme, dass die Prävalenzschätzungen unabhängig von der Randomisierungswahrscheinlichkeit sind. Ein solcher Test konnte bislang noch nie durchgeführt werden, da im ursprünglichen TRM (Yu et al., 2008) nur eine der beiden Gruppen benötigt und erhoben wird. Experiment 4 zeigte, dass sich die Prävalenzschätzungen beider Gruppen ($\hat{\pi}_1 = 22.58\%$, $SE = 2.38\%$; $\hat{\pi}_2 = 0.00\%$, $SE = 8.36\%$) signifikant voneinander unterschieden, $\Delta G^2(1) = 45.79$, $p < .001$. Somit waren die Prävalenzschätzungen im TRM offensichtlich von der verwendeten Randomisierungswahrscheinlichkeit abhängig und eine wichtige Annahme des Modells war verletzt. Daher konnte die Validität des TRM in Experi-

ment 4 nicht gegen die Validität der anderen Modelle getestet und auch die Spezifität und Sensitivität des Modells nicht ermittelt werden.

Die Prävalenz des sensiblen Merkmals *Betrug im Anagramm-Paradigma* wurde in der direkten Selbstauskunft auf 15.59% (SE = 1.54%) geschätzt. Da das CDM und CDTRM keine Annahme über den wahren Merkmalsstatus der Instruktionsverweigerer machen, kann in beiden Modellen die Prävalenz des sensiblen Merkmals nur innerhalb des Intervalls von π (unter der Annahme, dass kein Instruktionsverweigerer Merkmalsträger ist) bis $\pi + \gamma$ (unter der Annahme, dass alle Instruktionsverweigerer Merkmalsträger sind) geschätzt werden. Im CDM wurde dieses Intervall auf 29.70% (SE = 2.58%) bis 38.94% (SE = 3.89%) und im CDTRM auf 25.07% (SE = 2.51%) bis 38.36% (SE = 3.93%) geschätzt. Der Anteil der Instruktionsverweigerer wurde im CDM auf 9.24% (SE = 2.10%) und im CDTRM auf 13.28% (SE = 2.25%) geschätzt; beide Schätzer waren signifikant größer als 0%; CDM: $\Delta G^2(1) = 24.28$, $p < .001$; CDTRM: $\Delta G^2(1) = 45.79$, $p < .001$. In allen Bedingungen wurde die bekannte Prävalenz des sensiblen Merkmals (direkte Selbstauskunft: 57.71%, CDM: 59.73%, CDTRM: 59.25%) signifikant unterschätzt; direkte Selbstauskunft vs. bekannte Prävalenz: $\Delta G^2(1) = 423.32$, $p < .001$; CDM_{untere_Grenze} vs. bekannte Prävalenz: $\Delta G^2(1) = 131.94$, $p < .001$; CDM_{obere_Grenze} vs. bekannte Prävalenz: $\Delta G^2(1) = 26.53$, $p < .001$; $CDTRM_{untere\_Grenze}$ vs. bekannte Prävalenz: $\Delta G^2(1) = 172.27$, $p < .001$; $CDTRM_{obere\_Grenze}$ vs. bekannte Prävalenz: $\Delta G^2(1) = 26.11$, $p < .001$. Tabelle 3 zeigt die paarweisen Vergleiche der Prävalenzschätzungen zwischen den einzelnen Bedingungen. Die Prävalenzschätzer des CDM und CDTRM unterschieden sich nicht signifikant voneinander, waren jedoch signifikant höher als die Prävalenzschätzer aus der direkten Selbstauskunft.

Tabelle 3

Inferenzstatistische Vergleiche der Prävalenzschätzungen aus den verschiedenen Bedingungen.

				Modellpassung	
	Parameter 1	Parameter 2	|Differenz|	$\Delta G^2(1)$	*p*
$\hat{\pi}_{DS} = \hat{\pi}_{CDM\_uG}$	15.59	29.70	14.11	22.17	< .001*
$\hat{\pi}_{DS} = \hat{\pi}_{CDM\_oG}$	15.59	38.94	23.35	33.38	< .001*
$\hat{\pi}_{DS} = \hat{\pi}_{CDTRM\_uG}$	15.59	25.07	9.48	10.52	= .001*
$\hat{\pi}_{DS} = \hat{\pi}_{CDTRM\_oG}$	15.59	38.36	22.77	31.29	< .001*
$\hat{\pi}_{CDM\_uG} = \hat{\pi}_{CDTRM\_uG}$	29.70	25.07	4.63	1.65	= .199
$\hat{\pi}_{CDM\_oG} = \hat{\pi}_{CDTRM\_oG}$	38.94	38.36	0.58	0.01	= .916
$\hat{\gamma}_{CDM} = \hat{\gamma}_{CDTRM}$	9.24	13.28	4.04	1.73	= .188
$\hat{\beta}_{CDM} = \hat{\beta}_{CDTRM}$	61.06	61.65	0.59	0.01	= .916

Anmerkung: DS = Direkte Selbstauskunft, CDM = Cheating-Detection-Modell, CDTRM = Cheating-Detection-Triangular-Modell, uG = untere Grenze (unter der Annahme, dass kein Instruktionsverweigerer Merkmalsträger ist), oG = obere Grenze (unter der Annahme, dass alle Instruktionsverweigerer Merkmalsträger sind).

Außerdem wurden die Spezifität und Sensitivität des CDM, des CDTRM und der direkten Selbstauskunft ermittelt. Die Spezifität berechnet sich genau wie in Experiment 3 als der Anteil von der Fragetechnik geschätzter Nicht-Merkmalsträger ($1 - \pi$) innerhalb der Gruppe ehrlicher Teilnehmer. Da im CDM und CDTRM jedoch der Merkmalsstatus der Instruktionsverweigerer ungeklärt bleibt, liegt der geschätzte Anteil der Nicht-Merkmalsträger dabei zwischen β (= $1 - \pi - \gamma$; unter der Annahme, dass alle Instruktionsverweigerer Merkmalsträger sind) und $\beta + \gamma$ (= $1 - \pi$; unter der Annahme, dass kein Instruktionsverweigerer Merkmalsträger ist); es kann daher auch für die Spezifität nur eine untere und eine obere Grenze angegeben werden. Die Spezifität lag in der direkten Selbstauskunft bei 97.88% (*SE* = 0.94%), im CDM zwischen 79.13% (*SE* = 5.74%) und 83.69% (*SE* = 3.68%), und im CDTRM zwischen 79.46% (*SE* = 5.75%) und 90.19% (*SE* = 3.40%). Tabelle 4 zeigt die inferenzstatistischen Vergleiche. Die Spezifität aller Fragetechniken lag signifikant unter 100%; die Spezifität der direkten Selbstauskunft war jedoch signifikant höher als die Spezifität des

CDM und des CDTRM. Die Spezifität des CDM und CDTRM unterschied sich nicht signifikant voneinander.

Die Sensitivität berechnet sich jeweils als der Anteil von der Fragetechnik geschätzter Merkmalsträger (π) innerhalb der Gruppe der Betrüger. Im CDM und CDTRM liegt der Anteil der Merkmalsträger hierbei zwischen π und $\pi + \gamma$; es kann auch hier also nur eine obere und untere Grenze der Sensitivität angegeben werden. Die Sensitivität lag in der direkten Selbstauskunft bei 25.47% (*SE* = 2.43%), in der CDM-Bedingung zwischen 39.59% (*SE* = 3.48%) und 52.00% (*SE* = 5.16%), und in der CDTRM-Bedingung zwischen 36.18% (*SE* = 3.42%) und 51.34% (*SE* = 5.23%). Tabelle 4 zeigt die inferenzstatischen Vergleiche. Die Sensitivität aller Fragetechniken lag signifikant unter 100%; die Sensitivität von CDM und CDTRM war jedoch signifikant höher als die Sensitivität der direkten Selbstauskunft. Die Sensitivität von CDM und CDTRM unterschied sich nicht signifikant voneinander.

Tabelle 4

Inferenzstatistische Vergleiche der Spezifität und Sensitivität in den verschiedenen Bedingungen.

				Modellpassung	
	Parameter 1	Parameter 2	|Differenz|	$\Delta G^2(1)$	*p*
Spezifität					
$Spez_{DS} = 100\%$	97.88	100.00	2.12	135.77	< .001*
$Spez_{CDM\_uG} = 100\%$	79.13	100.00	20.87	25.55	< .001*
$Spez_{CDM\_oG} = 100\%$	83.69	100.00	16.31	25.55	< .001*
$Spez_{CDTRM\_uG} = 100\%$	79.46	100.00	20.54	15.24	< .001*
$Spez_{CDTRM\_oG} = 100\%$	90.19	100.00	9.81	10.14	= .001*
$Spez_{DS} = Spez_{CDM\_uG}$	97.88	79.13	18.75	15.92	< .001*
$Spez_{DS} = Spez_{CDM\_oG}$	97.88	83.69	14.19	15.91	= .001*
$Spez_{DS} = Spez_{CDTRM\_uG}$	97.88	79.46	18.42	11.46	< .001*
$Spez_{DS} = Spez_{CDTRM\_oG}$	97.88	90.19	7.69	5.26	= .022*
$Spez_{CDM\_uG} = Spez_{CDTRM\_uG}$	79.13	79.46	0.33	0.00	= .968
$Spez_{CDM\_oG} = Spez_{CDTRM\_oG}$	83.69	90.19	6.50	1.68	= .195

Sensitivität					
$Sens_{DS} = 100\%$	25.47	100.00	74.53	8476.52	< .001*
$Sens_{CDM\_uG} = 100\%$	39.59	100.00	60.41	7946.30	< .001*
$Sens_{CDM\_oG} = 100\%$	52.00	100.00	48.00	79.38	< .001*
$Sens_{CDTRM\_uG} = 100\%$	36.18	100.00	63.82	8769.51	< .001*
$Sens_{CDTRM\_oG} = 100\%$	51.34	100.00	48.66	78.78	< .001*
$Sens_{DS} = Sens_{CDM\_uG}$	25.47	39.59	14.12	10.94	< .001*
$Sens_{DS} = Sens_{CDM\_oG}$	25.47	52.00	26.53	22.55	< .001*
$Sens_{DS} = Sens_{CDTRM\_uG}$	25.47	36.18	10.71	6.48	= .011*
$Sens_{DS} = Sens_{CDTRM\_oG}$	25.47	51.34	25.87	21.14	< .001*
$Sens_{CDM\_uG} = Sens_{CDTRM\_uG}$	39.59	36.18	3.41	0.49	= .485
$Sens_{CDM\_oG} = Sens_{CDTRM\_oG}$	52.00	51.34	0.66	0.01	= .928

Anmerkung: DS = Direkte Selbstauskunft, CDM = Cheating-Detection-Modell, CDTRM = Cheating-Detection-Triangular-Modell, uG = untere Grenze (unter der Annahme, dass kein Instruktionsverweigerer Merkmalsträger ist), oG = obere Grenze (unter der Annahme, dass alle Instruktionsverweigerer Merkmalsträger sind).

Da das CDTRM im Vergleich zum CDM einfachere Instruktionen aufweist, sollte das CDTRM auch besser verständlich sein, das heißt, es sollten mehr Befragte auf Anhieb die Instruktionen verstehen und daher die Verständnisfragen korrekt beantworten. Tatsächlich beantworteten im CDM nur 13.42% der Befragten auf Anhieb die Verständnisfragen korrekt, im CDTRM waren es 26.99%, $\chi^2(1) = 63.53$, $p < .001$, *Cramer-V* = .169. Nach spätestens der dritten Runde Verständnisfragen hatten jedoch in beiden Bedingungen vergleichbar viele Befragte die Verständnisfragen korrekt beantwortet (CDM: 76.1%, CDTRM: 76.1%), $\chi^2(1) < 0.001$, $p = .994$, *Cramer-V* < .001.

Die neun Aussagen zur Evaluation der Fragetechnik erfassten ein homogenes Konstrukt (Cronbachs Alpha = .84). Daher konnte ein Mittelwert über alle Aussagen berechnet werden, sodass höhere Werte eine positivere Evaluation der Fragetechnik bedeuten. Es zeigte sich, dass das CDTRM ($M = 5.21$, $SD = 1.08$) von den Befragten signifikant positiver evaluiert wurde als das CDM ($M = 4.75$; $SD = 1.08$), $t(2227) = 10.08$, $p < .001$, $d = 0.43$.

Zusammenfassend zeigte Experiment 4, dass auch das CDM und das CDTRM ähnlich wie das CWM in Experiment 3 eine nicht optimale Spezifität und Sensitivität von signifikant unter 100% aufwiesen. Ebenso wie in Experiment 3 war die Spezifität einer direkten Selbstauskunft zwar besser als die Spezifität der RRT-Varianten, aber auch noch signifikant gerin-

ger als 100%; die Sensitivität hingegen war genau wie in Experiment 3 in der direkten Selbstauskunft am geringsten. Zudem bestätigte Experiment 4 den Befund von Experiment 3, dass die Spezifität aller Verfahren deutlich höher war als ihre Sensitivität und dass insgesamt die bekannte Prävalenz des sensiblen Merkmals von allen Fragetechniken signifikant unterschätzt wurde. Allerdings ergaben sowohl das CDM als auch das CDTRM Schätzungen, die näher an der bekannten Prävalenz des sensiblen Merkmals lagen als die Schätzungen aus einer direkten Selbstauskunft.

Zudem zeigte Experiment 4, dass die Erweiterung des TRM um eine Verweigererdetektion, wie mit dem CDTRM vorgeschlagen, sinnvoll und notwendig ist. Das TRM in seiner derzeitigen Form ohne Verweigererdetektion scheint nicht sinnvoll anwendbar zu sein, da die resultierenden Prävalenzschätzungen stark von der verwendeten Randomisierungswahrscheinlichkeit abhängen. Wenn einige der Befragten im TRM die Instruktionen missachten und die selbstschützende Antwort wählen, wirkt sich dies vermutlich unterschiedlich stark auf die beiden Gruppen mit den unterschiedlichen Randomisierungswahrscheinlichkeiten aus. Das TRM kann nicht abbilden, wenn der Anteil der „Ich stimme *mindestens einer* der beiden Aussagen zu“-Antworten unterhalb der Randomisierungswahrscheinlichkeit p liegt. Daher sind Prävalenzschätzungen des TRM bei hohen Randomisierungswahrscheinlichkeiten eher verzerrt. Dies könnte erklären, warum das TRM in der Bedingung mit einer hohen Randomisierungswahrscheinlichkeit ($p = .842$) zu einer unplausiblen Grenzschätzung von 0% führte. Die Erweiterung des TRM um eine Verweigererdetektion, wie im CDTRM vorgeschlagen, hat hingegen den Vorteil, dass der Anteil der Instruktionsverweigerer zumindest sichtbar gemacht werden kann. Das CDTRM schätzte den Anteil der Instruktionsverweigerer auf ca. 13% und liefert somit indirekte Evidenz dafür, dass auch im TRM Instruktionsverweigerung stattgefunden hat.

Ein Vergleich des CDTRM mit dem CDM zeigte, dass beide Modelle zu vergleichbaren Prävalenzschätzungen führten und eine vergleichbare Spezifität und Sensitivität aufwiesen. Das CDTRM war gegenüber dem CDM jedoch deutlich besser verständlich – ein Befund, der mit der Grundidee von NRRT übereinstimmt, die Verständlichkeit zu erleichtern (Yu et al., 2008). Darüber hinaus war das CDTRM dem CDM auch hinsichtlich der subjektiven Evaluation der Befragten überlegen. Die Wahrnehmung der Fragetechnik durch die Befragten könnte insbesondere für messwiederholte Designs wichtig sein, denn Befragte mit einer positiven Einstellung zur RRT könnten wahrscheinlicher an einem zweiten Messzeitpunkt teilnehmen als Befragte mit einer negativen Einstellung. Unter Berücksichtigung dieser Gesichtspunkte ist die Verwendung des CDTRM der Verwendung des CDM vorzuziehen.

5 Diskussion

Die RRT soll soziale Erwünschtheit in Selbstauskünften zu sensiblen Merkmalen kontrollieren, indem sie mit Hilfe einer Zufallsverschlüsselung die Vertraulichkeit individueller Antworten garantiert. Bisherige Befunde zeigen, dass die Validität der RRT zwar höher ist als die Validität einer direkten Selbstauskunft, aber RRT-Schätzungen immer noch von der bekannten Prävalenz sensibler Merkmale abweichen (Lensvelt-Mulders et al., 2005). In der vorliegenden Dissertation wurde die Validität verschiedener RRT-Varianten untersucht. Es wurde geprüft, wie gut diese RRT-Varianten hinsichtlich der Validität, Spezifität und Sensitivität abschneiden. Zu diesem Zweck wurden zum einen schwache Validierungen durchgeführt, in denen die Prävalenzschätzungen der RRT mit den Prävalenzschätzungen aus einer direkten Selbstauskunft verglichen wurden. Zum anderen wurden starke Validierungen durchgeführt, in denen die RRT-Schätzer mit der bekannten Prävalenz sensibler Merkmale verglichen wurden oder in denen gar der Merkmalsstatus der Befragten auf Individualebene bekannt war und zur Berechnung der Spezifität und Sensitivität herangezogen werden konnte. Außerdem wurde untersucht, welche Faktoren die Validität der RRT-Schätzungen erhöhen.

Insgesamt ermöglichte die RRT validere Prävalenzschätzungen als eine direkte Selbstauskunft. Allerdings unterschätzte die RRT die bekannte Prävalenz sozial unerwünschter Merkmale noch und wies außerdem eine nicht optimale Spezifität und Sensitivität auf. Die Validität der RRT konnte verbessert werden, wenn statt eines asymmetrischen Modells ein symmetrisches Modell genutzt wurde, oder das asymmetrische Modell um eine Verweigererdetektion erweitert wurde. Außerdem zeigte sich eine verbesserte Spezifität für hochgebildete Befragte unter Anwendung von Verständnisfragen und ausführlichen Instruktionen.

Abgesehen vom TRM führten alle im Rahmen der vorliegenden Dissertation untersuchten RRT (CWM, ECWM, CDM, CDTRM) zu signifikant höheren Schätzungen für sozial unerwünschte Merkmale als eine direkte Selbstauskunft (Experimente 1 bis 4). Nach einem schwachen Validierungskriterium ermöglichten die untersuchten RRT daher potentiell validere Prävalenzschätzungen als eine direkte Selbstauskunft. Dieser Befund stimmt mit den Ergebnissen vieler Studien zur RRT überein (beispielsweise Hoffmann et al., 2019; Hoffmann & Musch, 2016; Hoffmann & Musch, 2019; Jann et al., 2012; Jerke & Krumpal, 2013; Korndörfer et al., 2014; Kundt et al., 2017; Moshagen & Musch, 2012; Moshagen et al., 2010; Musch et al., 2001; Nakhaee et al., 2013; Ostapczuk et al., 2011; Thielmann et al., 2016; Waubert de Puiseau et al., 2017). Schwache Validierungen können jedoch nicht zeigen, wie gut die resultierenden Prävalenzschätzungen die wahre Prävalenz eines sensiblen Merkmals

abbilden und ob es sich um Über- oder Unterschätzungen handelt. Eine solche Evaluation ermöglichen nur starke Validierungen (Moshagen et al., 2014; Umesh & Peterson, 1991). Nach einem starken Validierungskriterium ergaben die im Rahmen der vorliegenden Arbeit untersuchten RRT (CWM, CDM, CDTRM) im Vergleich zu einer direkten Selbstauskunft zwar Schätzungen, die näher an der bekannten Prävalenz sozial unerwünschter Merkmale lagen, diese jedoch immer noch unterschätzten (Experimente 3 und 4). Dieser Befund deckt sich mit den Ergebnissen einer Meta-Analyse über andere RRT-Varianten (Lensvelt-Mulders et al., 2005). Starke Validierungen, in denen der individuelle Merkmalsstatus Befragter bekannt ist, ermöglichen darüber hinaus, die Spezifität und Sensitivität von Schätzern zu ermitteln. Derartige starke Validierungen mit Daten zum individuellen Merkmalsstatus der Befragten wurden in den Experimenten 3 und 4 der vorliegenden Arbeit durchgeführt. Beide Experimente zeigten, dass die Sensitivität und Spezifität der untersuchten RRT (CWM, CDM, CDTRM) sowie einer direkten Selbstauskunft signifikant unter 100% lagen und daher nicht optimal waren. Die Sensitivität der RRT war jedoch der Sensitivität einer direkten Selbstauskunft überlegen. Dieser Befund stimmt mit der ursprünglichen Idee der RRT überein, den Anteil der ehrlich antwortenden Merkmalsträger zu erhöhen. Die Spezifität hingegen war in einer direkten Selbstauskunft höher als in den untersuchten RRT. Insgesamt war die Sensitivität aller Fragetechniken deutlich geringer als ihre Spezifität (Experiment 3 und 4).

Vor dem Hintergrund der suboptimalen Spezifität und Sensitivität aller Selbstauskünfte zu sensiblen Merkmalen in dieser Arbeit stellt sich zunächst die Frage nach den unmittelbaren Implikationen. Eine reduzierte Sensitivität könnte zu einer Unterschätzung der Prävalenz sozial unerwünschter Merkmale führen, eine reduzierte Spezifität im schlimmsten Fall zu einer Überschätzung (Höglinger & Diekmann, 2017). In der vorliegenden Arbeit zeigte sich jedoch, dass die RRT die bekannte Prävalenz sozial unerwünschter Merkmale immer noch unterschätzte. Daher stellen auch bisherige Prävalenzschätzungen auf Grundlage der RRT trotz unzureichender Spezifität vermutlich eher Unter- als Überschätzungen der wahren Prävalenz dar. Ausnahmen bilden allerdings Merkmale mit Nullprävalenz; bei diesen führt die unzureichende Spezifität zwangsläufig zu einer Überschätzung (Höglinger & Diekmann, 2017).

Angesichts der Ergebnisse der vorliegenden Arbeit scheint der Einsatz der RRT statt einer direkten Selbstauskunft dann gerechtfertigt, wenn das Ziel der Untersuchung ist, entweder die Validität des Gesamtschätzers oder die Sensitivität zu maximieren. In Kontexten, in denen eine Unterschätzung der Prävalenz eines sensiblen Merkmals schwerwiegende negative Folgen hätte, könnten Forscher beispielsweise die Sensitivität priorisieren wollen. Dies könn-

te insbesondere dann der Fall sein, wenn auf der Basis von Prävalenzschätzungen die Notwendigkeit von Bildungs- oder Präventionsmaßnahmen abgeleitet werden soll, beispielsweise in Studien zur Steuerhinterziehung (Korndörfer et al., 2014; Kundt et al., 2017) oder Xenophobie (Hoffmann & Musch, 2016; Ostapczuk, Musch, et al., 2009). Sollte jedoch das primäre Ziel einer Untersuchung sein, die Spezifität zu maximieren, so ist eine direkte Selbstauskunft der RRT überlegen. Die Spezifität könnten Forscher beispielsweise dann maximieren wollen, wenn eine Überschätzung der Prävalenz eines sensiblen Merkmals mit schwerwiegenden negativen Folgen verbunden wäre, wie beispielsweise einem Reputationsverlust. Der Einsatz einer direkten Selbstauskunft ist auch hinsichtlich der Effizienz der Schätzung günstiger, denn eine direkte Selbstauskunft ist wesentlich effizienter als die RRT (Ulrich et al., 2012).

Die Befunde zur suboptimalen Sensitivität und Spezifität von Selbstauskünften werfen zahlreiche weitere Forschungsfragen auf. Zu diesen gehört beispielsweise, welche anderen Fragetechniken von einer suboptimalen Sensitivität und Spezifität betroffen sind, wie Spezifität und Sensitivität erhöht werden können und welche kognitiven Prozesse bei den Befragten letztendlich zu einer suboptimalen Spezifität und Sensitivität der Schätzer führen. Auf diese Fragen soll in den folgenden Absätzen genauer eingegangen werden.

Bislang wurde eine suboptimale Spezifität und Sensitivität im CWM, CDM, CDTRM und interessanterweise auch in der direkten Selbstauskunft nachgewiesen (Experimente 3 und 4 der vorliegenden Arbeit; Höglinger & Diekmann, 2017; Höglinger & Jann, 2018). In zwei weiteren Varianten der RRT zeigte sich nur ein Hinweis auf eine suboptimale Sensitivität, nicht aber auf eine suboptimale Spezifität (Höglinger & Jann, 2018). Suboptimale Sensitivität und Spezifität scheinen also nicht das Problem einer einzelnen Fragetechnik zu sein, sondern vielmehr ein allgemeines Problem verschiedener Methoden zur Erfassung von Selbstauskünften zu sensiblen Merkmalen. In einem nächsten Schritt sollte untersucht werden, ob andere indirekte Fragetechniken, wie beispielweise die Unmatched-Count-Technik (Miller, 1984) oder der Stochastische Lügendetektor (Moshagen et al., 2012), ebenfalls eine suboptimale Sensitivität und Spezifität aufweisen und welche Fragetechniken besonders stark betroffen sind. Hierzu sollten möglichst viele verschiedene Fragetechniken innerhalb einer Stichprobe hinsichtlich ihrer Sensitivität und Spezifität verglichen werden. Um überhaupt die Sensitivität und Spezifität bestimmen zu können, ist der Einsatz von starken Validierungen mit bekanntem individuellen Merkmalsstatus der Befragten notwendig. Es ist jedoch schwierig, auf Individualebene an sensible Daten zu Befragten zu gelangen, um diese dann als starkes Außenkriterium nutzen zu können (siehe aber Landsheer et al., 1999; van der Heijden,

van Gils, Bouts, & Hox, 2000). Daher sind Paradigmen, in denen ein sensibles Merkmal experimentell induziert werden kann, besonders nützlich (siehe beispielsweise Hoffmann et al., 2015; Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; John et al., 2018; Moshagen et al., 2014).

Abgesehen von der Frage, inwieweit die Sensitivität und Spezifität von Selbstauskünften von der verwendeten Fragetechnik abhängen, ist auch interessant, inwieweit Sensitivität und Spezifität vom Kontext abhängen und verbessert werden können. In Experiment 3 der vorliegenden Dissertation wurde erstmalig untersucht, wie die Spezifität im CWM erhöht werden kann. Bei Befragten mit hohem Bildungsniveau erhöhte sich die Spezifität unter Anwendung detaillierter Instruktionen und Verständnisfragen im Vergleich zu kurzen Instruktionen ohne Verständnisfragen. Eine Erhöhung der Spezifität ging jedoch mit einer Reduktion der Sensitivität einher, und bei Befragten mit niedrigem Bildungsniveau konnte die Spezifität überhaupt nicht erhöht werden. Somit zeigt Experiment 3, dass die konkrete Implementierung der RRT die Validität der Schätzungen erheblich beeinflussen kann (siehe auch John et al., 2018). Derzeit ist die Implementierung häufig nicht ausführlich dokumentiert, sodass eventuell problematische Implementierungen gar nicht bemerkt werden können. Zukünftige Forschung sollte daher auch die Implementierung der RRT näher untersuchen. Experiment 3 verdeutlichte zudem, dass es dringend erforderlich ist, geeignete RRT-Instruktionen für Befragte mit niedrigem Bildungsniveau zu entwickeln. RRT-Untersuchungen sollten nicht nur auf Stichproben mit hohem Bildungsniveau beschränkt bleiben, sondern in breiteren Stichproben anwendbar sein. Die unterschiedliche Validität der Prävalenzschätzungen für Stichproben mit unterschiedlichem Bildungsniveau ist insbesondere dann ein Problem, wenn die Prävalenz eines sensiblen Merkmals ebenfalls vom Bildungsniveau der Befragten moderiert wird, wie es beispielsweise bei Xenophobie der Fall ist (Ostapczuk, Musch, et al., 2009).

Bislang ist ungeklärt, welche kognitiven Prozesse bei den Befragten zu einer verringerten Spezifität und Sensitivität führen. Es ist anzunehmen, dass mangelnde Spezifität hauptsächlich auf ein Missverständnis der Instruktionen zurückzuführen ist, denn es scheint unwahrscheinlich, dass Nicht-Merkmalsträger absichtlich so antworten wie Merkmalsträger. Diese Annahme wird dadurch gestützt, dass in der vorliegenden Arbeit ein Anheben des Verständnisses zu besserer Spezifität führte (Experiment 3). Mangelnde Sensitivität hingegen ist wahrscheinlich auf eine Mischung aus einem Missverständnis der Instruktionen und absichtlichen Falschantworten zurückzuführen. In beiden Fällen ist jedoch unklar, wie genau die Befragten vorgehen, um eine Antwortoption auszuwählen. Weiterführende Studien sollten daher beispielsweise mit Hilfe qualitativer Interviews (wie beispielsweise in Boeije &

Lensvelt-Mulders, 2002; Lensvelt-Mulders & Boeije, 2007) untersuchen, welche kognitiven Prozesse bei den Befragten bei der Beantwortung von RRT-Fragen auftreten.

Die vorliegende Dissertation hatte weiterhin zum Ziel, zu untersuchen, welche Modelleigenschaften günstig für die Validität von RRT-Schätzungen sind. Hierbei stellten sich einerseits die Antwortsymmetrie von Modellen und andererseits ein Mechanismus zur Verweigererdetektion als wichtig heraus. Antwortsymmetrie bezeichnet die Abwesenheit einer selbstschützenden Antwortoption, die unehrliche Befragte wählen können, um eindeutig auszuschließen, Träger des sensiblen Merkmals zu sein. In Experiment 1 wurde erstmalig die Validität zweier RRT-Varianten (CWM und TRM) gegenübergestellt, die sich lediglich hinsichtlich ihrer Antwortsymmetrie unterscheiden. Verglichen mit einer direkten Selbstauskunft führte das symmetrische CWM zu signifikant höheren und daher nach einem schwachen Validierungskriterium potentiell valideren Prävalenzschätzungen für ein sozial unerwünschtes Merkmal, das asymmetrische TRM hingegen nicht. In Experiment 4 zeigte sich darüber hinaus in einer erstmaligen Erweiterung des TRM auf zwei Gruppen mit unterschiedlichen Randomisierungswahrscheinlichkeiten, dass die Schätzer des TRM abhängig von der Randomisierungswahrscheinlichkeit sind. Damit ist eine wichtige Voraussetzung des Modells verletzt, die bislang in Untersuchungen mit nur einer Randomisierungswahrscheinlichkeit nicht getestet werden konnte. Sowohl die bislang fehlenden Hinweise auf eine erhöhte Validität des TRM im Vergleich zu einer direkten Selbstauskunft (siehe auch Erdmann, 2019; Jerke & Krumpal, 2013) als auch die Modellverletzung des TRM können vermutlich mit der fehlenden Antwortsymmetrie des Modells erklärt werden. Das Vorhandensein einer eindeutig selbstschützenden Antwortoption begünstigt Instruktionsverweigerung (Ostapczuk, Moshagen, et al., 2009). Es ist daher anzunehmen, dass im asymmetrischen TRM einige Befragte die Instruktionen missachten und die selbstschützende Antwortoption wählen, um eindeutig auszuschließen, für einen Träger des sensiblen Merkmals gehalten zu werden. Instruktionsverweigerung in Form einer Wahl der selbstschützenden Antwortoption hat jedoch eine Verringerung der Validität von Prävalenzschätzungen zur Folge und wirkt sich wahrscheinlich im TRM bei hohen Randomisierungswahrscheinlichkeiten stärker aus. In symmetrischen Modellen wie dem CWM gibt es hingegen keine vollkommen selbstschützende Antwortoption. Zwar beinhaltet im CWM eine der beiden Optionen abhängig von der Randomisierungswahrscheinlichkeit ein geringeres Risiko, als Merkmalsträger identifiziert zu werden, allerdings ist unklar, inwieweit Befragte überhaupt in der Lage sind, diese Option zu identifizieren. Eine Studie, in der die Hälfte der Befragten ehrlich antworten sollte und die andere Hälfte eine „Fake-good“-Instruktion erhielt, zeigte, dass es Befragten in der „Fake-good“-Bedingung

nicht gelang, die Prävalenzschätzer des symmetrischen CWM zu verfälschen (Hoffmann et al., 2019).

Die vorliegende Arbeit zeigte weiterhin die Bedeutung eines Mechanismus zur Verweigererdetektion für die Validität der RRT. Die Asymmetrie des TRM ist vor allem deshalb problematisch, weil eine Instruktionsverweigerung zwar wahrscheinlich ist und die Prävalenzschätzungen verzerrt, vom Modell aber nicht entdeckt werden kann. In Experiment 4 wurde mit dem CDTRM eine Kombination der Instruktionen des TRM mit dem Mechanismus zur Verweigererdetektion des CDM vorgeschlagen. Das CDTRM zeigte einen signifikanten Anteil an Instruktionsverweigerern an und lieferte daher indirekte Evidenz dafür, dass im ursprünglichen TRM ebenfalls Instruktionsverweigerung aufgetreten ist. Die Kombination asymmetrischer Modelle mit einer Verweigererdetektion hat sich auch bereits in anderen Studien als vorteilhaft erwiesen (Moshagen & Musch, 2012; Wu & Tang, 2016). Das CDTRM hat zudem gegenüber dem CDM, dessen Messmodell es verwendet, den Vorteil verständlicherer Instruktionen; entsprechend war das CDTRM dem CDM hinsichtlich seiner objektiven Verständlichkeit und auch hinsichtlich der subjektiven Bewertung der Befragten überlegen (Experiment 4). Die vorliegende Arbeit untersuchte mit dem ECWM darüber hinaus ein weiteres Modell, das einen Mechanismus zur Verweigererdetektion beinhaltet. Das ECWM ist eine Erweiterung des CWM und kann zumindest Instruktionsverweigerung in Form einer systematischen Präferenz für eine der beiden Antwortoptionen detektieren; hierbei hat es dieselbe statistische Effizienz wie das CWM (Heck et al., 2018). In Experiment 2 konnte anhand einer ersten schwachen Validierung gezeigt werden, dass das ECWM zu höheren und daher mutmaßlich valideren Prävalenzschätzungen führte als eine direkte Selbstauskunft.

Insgesamt scheint es sinnvoll, statt asymmetrischen Modellen wie dem TRM entweder symmetrische Modelle wie das CWM einzusetzen oder asymmetrische Modelle um eine Verweigererdetektion zu erweitern, wie es im CDM oder dem in Experiment 4 vorgeschlagenen CDTRM der Fall ist. Das ECWM vereint in gewisser Weise die Vorteile beider Ansätze, da es ein symmetrisches Modell ist, das zumindest eine bestimmte Art der Instruktionsverweigerung entdecken kann. Zudem nutzt das ECWM auch leicht verständliche Instruktionen. Gegenüber dem CDTRM hat das ECWM den Vorteil, dass es effizienter und darüber hinaus testbar ist. Es ist allerdings noch unklar, wie das ECWM in einem direkten Vergleich mit dem CDTRM hinsichtlich der Validität, Spezifität und Sensitivität abschneidet. Daher sollten beide Modelle in einer weiterführenden starken Validierung mit bekanntem individuellen Merkmalsstatus der Befragten an derselben Stichprobe miteinander verglichen werden.

Zusammenfassend zeigte die vorliegende Dissertation, dass die Validität der RRT besser war als die Validität einer direkten Selbstauskunft, da RRT-Schätzer näher an der bekannten Prävalenz sozial unerwünschter Merkmale lagen als Prävalenzschätzer aus einer direkten Selbstauskunft. Allerdings erwies sich die Validität der RRT als nicht perfekt, denn zum einen unterschätzte die RRT die Validität von sozial unerwünschten Merkmalen noch (siehe auch Lensvelt-Mulders et al., 2005) und zum anderen waren Sensitivität und Spezifität der Methode suboptimal. In der vorliegenden Arbeit wurden jedoch Faktoren identifiziert, die dazu beitragen, die Validität der RRT zu verbessern. Hierzu zählen Modelleigenschaften – wie Antwortsymmetrie oder ein Mechanismus zur Verweigererdetektion –, Eigenschaften des Kontexts – wie die Verwendung ausführlicher Instruktionen und Verständnisfragen – und Eigenschaften der Stichprobe – wie zum Beispiel ein hohes Bildungsniveau der Befragten.

Literaturverzeichnis

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10*, 331-344. doi:10.1037/1040-3590.10.4.331

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86. doi:10.3758/Bf03210812

Bertelsmann Stiftung. (2017). *Willkommenskultur im „Stresstest“ - Einstellungen in der Bevölkerung 2017 und Entwicklungen und Trends seit 2011/2012 - Ergebnisse einer repräsentativen Bevölkerungsumfrage [Welcoming culture under 'stress test' - attitudes in the population in 2017 and developments and trends since 2011/2012 - findings of a representative survey of the population]*. Retrieved from http://www.bertelsmann-stiftung.de/fileadmin/files/Projekte/28_Einwanderung_und_Vielfalt/IB_Umfrage_Willkommenskultur_2017.pdf

Boeije, H. R., & Lensvelt-Mulders, G. J. L. M. (2002). Honest by chance: A qualitative interview study to clarify respondents (non-) compliance with computer-assisted randomized response. *Bulletin Methodologie Sociologique, 75*, 24-39.

Bogardus, E. S. (1933). A Social Distance Scale. *Sociology & Social Research, 17*, 265-271.

Boruch, R. F. (1971). Assuring Confidentiality of Responses in Social Research: A Note on Strategies. *American Sociologist, 6*, 308-311.

Chaudhuri, A., & Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys*. Berlin, Heidelberg: Springer.

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160-168.

Dawes, R. M., & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen [Guttman scaling of orthodox and randomized reactions]. In F. Petermann (Ed.), *Einstellungsmessung, Einstellungsforschung [Attitude measurement, attitude research]* (pp. 117–133). Göttingen: Hogrefe.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39*, 1-38.

Edgell, S. E., Duchan, K. L., & Himmelfarb, S. (1992). An Empirical-Test of the Unrelated Question Randomized-Response Technique. *Bulletin of the Psychonomic Society, 30*, 153-156.

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of Forced Responses in a Randomized-Response Model. *Sociological Methods & Research, 11*, 89-100. doi:10.1177/0049124182011001005

Erdmann, A. (2019). Non-Randomized Response Models: An Experimental Application of the Triangular Model as an Indirect Questioning Method for Sensitive Topics. *methods, data, analyses, 13*, 139-167. doi:10.12758/mda.2018.07

Heck, D. W., Hoffmann, A., & Moshagen, M. (2018). Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior Research Methods, 50*, 1895-1905. doi:10.3758/s13428-017-0957-8

Heitmeyer, W. (2012). Gruppenbezogene Menschenfeindlichkeit (GMF) in einem entsicherten Jahrzehnt [Group-focused enmity (GFE) in a ??? decade]. In W. Heitmeyer (Ed.), *Deutsche Zustände: Folge 10* (pp. 15-41). Berlin: Suhrkamp.

Hoffmann, A., Diedenhofen, B., Verschuere, B. J., & Musch, J. (2015). A strong validation of the Crosswise Model using experimentally induced cheating behavior. *Experimental Psychology, 62*, 403-414. doi:10.1027/1618-3169/a000304

Hoffmann, A., Meisters, J., & Musch, J. (2019). Nothing but the truth? Effects of faking on the validity of the crosswise model. [Manuscript submitted].

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: a Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods, 48*, 1032-1046. doi:10.3758/s13428-015-0628-6

Hoffmann, A., & Musch, J. (2019). Prejudice against Women Leaders: Insights from an Indirect Questioning Approach. *Sex Roles, 80*, 681–692. doi:10.1007/s11199-018-0969-6

Hoffmann, A., Waubert de Puiseau, B., Schmidt, A. F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods, 49*, 1470-1483. doi:10.3758/s13428-016-0804-3

Höglinger, M., & Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis, 25*, 131-137. doi:10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *Plos One, 13*. doi:10.1371/journal.pone.0201770

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling into Question the Method's Validity. *Public Opinion Quarterly, 74*, 328-343. doi:10.1093/Poq/Nfq012

Hu, X., & Batchelder, W. H. (1994). The Statistical-Analysis of General Processing Tree Models with the Em Algorithm. *Psychometrika, 59*, 21-47. doi:10.1007/Bf02294263

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking Sensitive Questions Using the Crosswise Model. *Public Opinion Quarterly, 76*, 32-49. doi:10.1093/Poq/Nfr036

Jerke, J., & Krumpal, I. (2013). Plagiate in studentischen Arbeiten [Plagiarism in Student Papers]. *methoden, daten, analysen, 7*, 347-368. doi:10.12758/mda.2013.017

Jimenez, P. (1999). Weder Opfer noch Täter - die alltäglichen Einstellungen 'unbeteiligter' Personen gegenüber Ausländern [Neither victim nor offender—the common attitudes of 'non-involved' persons towards foreigners]. In R. Dollase, T. Kliche, & H. Moser (Eds.), *Politische Psychologie der Fremdenfeindlichkeit. Opfer - Täter - Mittäter* (pp. 293–306). Weinheim: Juventa.

John, L. K., Loewenstein, G., Acquisti, A., & Vosgerau, J. (2018). When and why randomized response techniques (fail to) elicit the truth. *Organizational Behavior and Human Decision Processes, 148*, 101-123. doi:10.1016/j.obhdp.2018.07.004

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology, 45*, 18-32. doi:10.1016/j.joep.2014.08.001

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research, 41*, 1387-1403. doi:10.1016/j.ssresearch.2012.05.015

Kundt, T. C., Misch, F., & Nerré, B. (2017). Re-assessing the merits of measuring tax evasion through business surveys: an application of the crosswise model. *International Tax and Public Finance, 24*, 112-133. doi:10.1007/s10797-015-9373-0

Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response - A study of a method for improving the estimate of social security fraud. *Quality & Quantity, 33*, 1-12. doi:10.1023/A:1004361819974

Lensvelt-Mulders, G. J. L. M., & Boeije, H. R. (2007). Evaluating compliance with a computer assisted randomized response technique: a qualitative study into the origins of lying and cheating. *Computers in Human Behavior, 23*, 591-608. doi:10.1016/j.chb.2004.11.001

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: thirty-five years of validation. *Sociological Methods & Research, 33*, 319-348. doi:10.1177/0049124104268664

Mangat, N. S. (1994). An Improved Randomized-Response Strategy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 56*, 93-95.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior.* (Unpublished Ph.D. dissertation), George Washington University, Department of Sociology,

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42-54. doi:10.3758/BRM.42.1.42

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An Experimental Validation Method for Questioning Techniques That Assess Sensitive Issues. *Experimental Psychology, 61*, 48-54. doi:10.1027/1618-3169/a000226

Moshagen, M., Hilbig, B. E., & Musch, J. (2011). Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *European Journal of Social Psychology, 41*, 638-644. doi:10.1002/Ejsp.793

Moshagen, M., & Musch, J. (2012). Surveying Multiple Sensitive Attributes using an Extension of the Randomized-Response Technique. *International Journal of Public Opinion Research, 24*, 508-523.

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44*, 222-231. doi:10.3758/s13428-011-0144-2 21858604

Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2010). Reducing Socially Desirable Responses in Epidemiologic Surveys. An Extension of the Randomized-response Technique. *Epidemiology, 21*, 379-382. doi:10.1097/Ede.0b013e3181d61dbc

Musch, J., Bröder, A., & Klauer, K. C. (2001). Improving Survey Research on the World-Wide Web using the Randomized Response Technique. In U. D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 179-192). Lengerich, Germany: Pabst.

Nakhaee, M. R., Pakravan, F., & Nakhaee, N. (2013). Prevalence of Use of Anabolic Steroids by Bodybuilders Using Three Methods in a City of Iran. *Addict Health, 5*, 1-6.

Ostapczuk, M., Moshagen, M., Zhao, Z., & Musch, J. (2009). Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. *Journal of Educational and Behavioral Statistics, 34*, 267-287. doi:10.3102/1076998609332747

Ostapczuk, M., & Musch, J. (2011). Estimating the prevalence of negative attitudes towards people with disability: A comparison of direct questioning, projective questioning and randomised response. *Disability and Rehabilitation, 33*, 1-13. doi:10.3109/09638288.2010.492067

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology, 39*, 920-931. doi:10.1002/ejsp.588

Ostapczuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. *Statistical Methods in Medical Research, 20*, 489-503. doi:10.1177/0962280210372843

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes, Vol. 1* (pp. 17-59). San Diego, CA: Academic Press.

Petersen, T. (2012, 21 November 2012). Die Furcht vor dem Morgenland im Abendland [The fear of the orient in the occident]. *Frankfurter Allgemeine Zeitung*. Retrieved from https://www.faz.net/aktuell/politik/inland/allensbach-studie-die-furcht-vor-dem-morgenland-im-abendland-11966471-p2.html

Pollack, D. (2014). Wahrnehmung und Akzeptanz religiöser Vielfalt in ausgewählten Ländern Europas: Erste Beobachtungen [Perception and acceptance of religious diversity in selected European countries: First observations]. In C. Gärtner, M. Koenig, G. Pickel, K. Sammet, & H. Winkel (Eds.), *Grenzen der Toleranz. Veröffentlichungen der Sektion Religionssoziologie der Deutschen Gesellschaft für Soziologie* (pp. 13-34). Wiesbaden: Springer Fachmedien.

Pötzsch, O. (2012). Geburten in Deutschland [Births in Germany]. Retrieved Jun 6, 2012, from German Federal Statistical Office https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/Bevoelkerungsbewegung/BroschuereGeburtenDeutschland0120007129004.pdf

Rasinski, K. A., Visser, P. S., Zagatsky, M., & Rickett, E. M. (2005). Using implicit goal priming to improve the quality of self-report data. *Journal of Experimental Social Psychology, 41*, 321-327. doi:10.1016/j.jesp.2004.07.001

Reinders, M. (1996). Häufigkeit von Namensanfängen [Frequency of first letters of surnames]. *Statistische Rundschau Nordrhein-Westfalen, 11*, 651-660.

Schwarz, N. (1999). Self-reports - How the questions shape the answers. *American Psychologist, 54*, 93-105. doi:10.1037/0003-066x.54.2.93

Silbermann, A., & Hüsers, F. (1995). *Der 'normale' Haß auf die Fremden. Eine sozialwissenschaftliche Studie zu Ausmaß und Hintergründen von Fremdenfeindlichkeit in Deutschland [The 'normal' xenophobia. A socio-scientific study on the extent and determinants of xenophobia in Germany]*. München: Quintessenz.

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. *Judgment and Decision Making, 11*, 527-536.

Tian, G.-L., & Tang, M.-L. (2014). *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883. doi:10.1037/0033-2909.133.5.859 17723033

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking Sensitive Questions: A Statistical Power Analysis of Randomized Response Models. *Psychological Methods, 17*, 623-641. doi:10.1037/A0029314

Umesh, U. N., & Peterson, R. A. (1991). A Critical Evaluation of the Randomized-Response Method - Applications, Validation, and Research Agenda. *Sociological Methods & Research, 20*, 104-138. doi:10.1177/0049124191020001004

van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning - Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research, 28*, 505-537.

Wagner, U., & van Dick, R. (2001). Fremdenfeindlichkeit "in der Mitte der Gesellschaft": Phänomenbeschreibung, Ursachen, Gegenmaßnahmen [Xenophobia 'in the middle of society': description of phenomena, causes, countermeasures]. *Zeitschrift für Politische Psychologie, 9*, 41-54.

Warner, S. L. (1965). Randomized-Response - a Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*, 63-69.

Waubert de Puiseau, B., Hoffmann, A., & Musch, J. (2017). How indirect questioning techniques may promote democracy: A pre-election polling experiment. *Basic And Applied Social Psychology, 39*, 209-217. doi:10.1080/01973533.2017.1331351

Wu, Q., & Tang, M.-L. (2016). Non-randomized response model for sensitive survey with noncompliance. *Statistical Methods in Medical Research, 25*, 2827-2839. doi:10.1177/0962280214533022

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika, 67*, 251-263. doi:10.1007/s00184-007-0131-x

Zick, A., Küpper, B., & Hövermann, A. (2011). *Intolerance, Prejudice and Discrimination: A European Report*. Berlin: Friedrich-Ebert-Stiftung.

Eidesstattliche Versicherung

Eidesstattliche Versicherung gemäß § 5 der Promotionsordnung vom 15.06.2018 der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf:

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist. Ferner versichere ich, dass die Arbeit in der vorgelegten oder in ähnlicher Form bisher bei keiner anderen Fakultät als Dissertation eingereicht wurde und dass ich bisher keine erfolglosen Promotionsversuche unternommen habe.

Düsseldorf, den _____ _____

Datum Julia Meisters

Anhang: Einzelarbeiten

Originalartikel zu Experiment 1:

Hoffmann, A., Meisters, J., & Musch, J. (2019). On the Validity of Non-Randomized Response Techniques: An Experimental Comparison of the Crosswise Model and the Triangular Model. *Manuscript submitted for publication*.

Ich bin geteilte Erstautorin dieses Manuskripts. Ich habe einen Teil der Datenauswertung dieser Studie übernommen sowie das Manuskript geschrieben.

Originalartikel zu Experiment 2:

Meisters, J., Hoffmann, A., & Musch, J. (2019). Controlling social desirability bias: An experimental validation of the extended crosswise model. *Manuscript submitted for publication*.

Ich bin geteilte Erstautorin dieses Manuskripts. Ich war für die Planung der Studie und die Erstellung des Fragebogens mitverantwortlich. Außerdem habe ich die Datenauswertung und das Schreiben des Manuskripts übernommen.

Originalartikel zu Experiment 3:

Meisters, J., Hoffmann, A., & Musch, J. (2019). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? *Manuscript submitted for publication*.

Ich bin geteilte Erstautorin dieses Manuskripts. Ich war für die Planung der Studie mitverantwortlich und habe die Programmierung des Online-Fragebogens, die Datenauswertung sowie das Schreiben des Manuskripts übernommen.

Originalartikel zu Experiment 4:

Meisters, J., Hoffmann, A., & Musch, J. (2019). A new approach to detecting cheating in sensitive surveys: The Cheating Detection Triangular Model. *Manuscript submitted for publication*.

Ich bin geteilte Erstautorin dieses Manuskripts. Ich war für die Planung der Studie mitverantwortlich und habe die Programmierung des Online-Fragebogens, die Datenauswertung sowie das Schreiben des Manuskripts übernommen.

On the Validity of Non-Randomized Response Techniques: An Experimental Comparison of the Crosswise Model and the Triangular Model

Adrian Hoffmann*, Julia Meisters*, and Jochen Musch

University of Duesseldorf

«fn» *A. Hoffmann and J. Meisters contributed equally to this work.

Author Note

Adrian Hoffmann, Julia Meisters, and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Grant number 393108549.

Correspondence concerning this article should be addressed to Adrian Hoffmann, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, Floor 00, Room 27, 40225 Duesseldorf, Germany.

E-mail: adrian.hoffmann@uni-duesseldorf.de

Abstract

Non-Randomized response techniques (NRRTs) such as the crosswise model and the triangular model (CWM and TRM; Yu et al., 2008) have been developed to control for socially desirable responding in surveys on sensitive personal attributes. We present the first study to directly compare the validity of the CWM and TRM and contrast their performance with a conventional direct questioning (DQ) approach. In a paper-pencil survey of 1382 students, we obtained prevalence estimates for two sensitive attributes (xenophobia and rejection of further refugee admissions) and one nonsensitive control attribute with a known prevalence (the first letter of respondents' surnames). Both NRRTs yielded descriptively higher prevalence estimates for the sensitive attributes than DQ; however, only the CWM estimates were significantly higher. We attribute the higher prevalence estimates for the CWM to its response symmetry, which is lacking in the TRM. Only the CWM provides symmetrical answer options, meaning that there is no "safe" alternative respondents can choose to distance themselves from being carriers of the sensitive attribute. Prevalence estimates for the nonsensitive control attribute with known prevalence confirmed that neither method suffered from method-specific bias towards over- or underestimation. Exploratory moderator analyses further suggested that the sensitive attributes were perceived as more sensitive among politically left-oriented than among politically right-oriented respondents. Based on our results, we recommend using the CWM over the TRM in future studies on sensitive personal attributes.

Keywords: non-randomized response technique, crosswise model, triangular model, validity, xenophobia

On the Validity of Non-Randomized Response Techniques: An Experimental Comparison of the Crosswise Model and the Triangular Model

In surveys of sensitive attributes, social desirability bias threatens the validity of direct self-reports (Tourangeau & Yan, 2007). If some respondents choose to reply in line with social or legal norms rather than truthfully, prevalence estimates will be distorted in the direction of the socially desirable answer (Paulhus, 1991; Tourangeau & Yan, 2007). This response bias poses a serious threat to the interpretation of the results of survey studies on socially (un-)desirable or illegal behaviors, such as drug use, prejudice, abortion, tax evasion, and plagiarism. Indirect questioning techniques try to control for socially desirable responding. In the present article, we focus on two non-randomized response techniques (NRRT) that have recently been proposed as an advancement upon traditional randomized response techniques (RRT; Warner, 1965). In particular, we investigate the crosswise model and the triangular model (CWM and TRM; Yu, Tian, & Tang, 2008) and present the first direct comparison of the two methods' validity.

Randomized Response Techniques (RRT) and Non-Randomized Response Techniques (NRRT)

Randomized response techniques (RRT; Warner, 1965) ensure that individual responses remain completely confidential in order to encourage respondents to provide more honest answers less distorted by social desirability bias. In the original RRT, respondents are presented with two statements: a sensitive statement (e.g., "I have taken cocaine") and its negation (e.g., "I have *never* taken cocaine"). Respondents are then asked to reply to only one of the statements with "true" or "false". The statement respondents are asked to react to is determined by the outcome of a randomization procedure (e.g. the roll of a die or the respondent's birth month). It is impossible for the experimenter to tell whether any individual respondent has admitted to

consuming cocaine because the experimenter is not informed of the outcome of the randomization. The distribution of the potential randomization outcomes is known, however; therefore, prevalence estimates for the sensitive attribute can be obtained on the sample level via appropriate statistical procedures (Warner, 1965). In a meta-analysis of 32 so-called "weak" validation studies, RRTs obtained higher prevalence estimates than direct questions (DQ) for various sensitive personal attributes (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). The superiority of RRTs over DQ was found to increase with the increasing sensitivity of the attribute under investigation.

According to the "more is better" criterion, higher prevalence estimates for sensitive attributes are assumed to be more valid because they are presumably less distorted by social desirability bias. However, finding such a pattern can only be considered "weak" validity evidence, since higher prevalence estimates may still underestimate - or overestimate - the true prevalence (Umesh & Peterson, 1991). Therefore, "strong" validation studies in which the prevalence of the sensitive attribute is known and can be used as an external validation criterion are considered the gold standard. However, such studies are costly and difficult to implement, making strong validation studies quite rare (Lensvelt-Mulders et al., 2005; Umesh & Peterson, 1991). A meta-analysis found only six strong validation studies comparing the prevalence estimates of RRTs to estimates obtained via DQ. Overall, RRTs were found to provide more valid estimates than DQ, but still substantially underestimated the true prevalence (Lensvelt-Mulders et al., 2005).

RRTs have also been criticized for their relative inefficiency, the complexity of their instructions, and the need to use an external randomization device (Landsheer, van der Heijden, & van Gils, 1999; Ulrich, Schröter, Striegel, & Simon, 2012; Yu et al., 2008). Consequently,

non-randomized response techniques (NRRTs) such as the CWM and the TRM (Tian & Tang, 2014; Yu et al., 2008) have recently been proposed as an advancement upon RRTs. In contrast to RRTs, NRRTs directly integrate the randomization procedure into the answer options. They thus have simpler instructions and are easier for the experimenter to administer and for respondents to understand. Moreover, at least in contrast to the original RRT (Warner, 1965), the TRM is more efficient in most cases (Yu et al., 2008).

The Crosswise Model. In the CWM, respondents are simultaneously presented with two statements. The first statement A refers to a sensitive attribute with unknown prevalence π (e.g., "I have taken cocaine"); the second statement B refers to a nonsensitive control attribute with known prevalence *r* that is used for randomization (e.g., "My mother was born in November or December"). Respondents are requested to provide a joint answer to both statements by indicating whether "both statements are true or none of the statements is true" or "exactly one statement (irrespective of which one) is true". As in the original RRT, respondents can honestly choose either of the answer while their individual status with respect to the sensitive statement A remains completely confidential. On the sample level, however, a maximum likelihood estimate for the prevalence π of the sensitive attribute can be obtained by using the formula (Yu et al., 2008):

$$\hat{\pi}_{CWM} = \frac{\hat{\lambda}_{CWM} + r - 1}{2 * r - 1} \quad (1)$$

where $\hat{\lambda}_{CWM}$ is the observed proportion of respondents choosing the first answer option ("both statements are true or none of the statements is true"). An essential advantage of the CWM is its response symmetry: Both answer options can and must be chosen by both carriers and

noncarriers of the sensitive attribute, depending on their status with respect to the nonsensitive statement B (e.g., their mother's birth month).

To quantify the objective confidentiality protection of the two answer options in the CWM, conditional probabilities can be derived using Bayes' theorem (Bayes, 1763) according to the procedure described by Lanke (1976) and Fligner, Policello, and Singh (1977). The conditional probabilities of being identified as a carrier of the sensitive attribute given that one has selected the first ("both statements are true or none of the statements is true") versus the second answer option ("exactly one statement (irrespective of which one) is true") are:

$$Pr_{\text{CWM}}(\text{carrier} \mid \text{“both/none true”}) = \frac{Pr_{\text{CWM}}(\text{carrier} \cap \text{“both/none true”})}{Pr_{\text{CWM}}(\text{“both/none true”})} \tag{2.1}$$

$$Pr_{\text{CWM}}(\text{carrier} \mid \text{“one true”}) = \frac{Pr_{\text{CWM}}(\text{carrier} \cap \text{“one true”})}{Pr_{\text{CWM}}(\text{“one true”})} \tag{2.2}$$

These equations can be reformulated using the parameters for prevalence estimation from Equation 1:

$$Pr_{\text{CWM}}(\text{carrier} \mid \text{“both/none true”}) = \frac{\hat{\pi}_{\text{CWM}} * r}{\hat{\lambda}_{\text{CWM}}} \tag{2.3}$$

$$Pr_{\text{CWM}}(\text{carrier} \mid \text{“one true”}) = \frac{\hat{\pi}_{\text{CWM}} * (1 - r)}{(1 - \hat{\lambda}_{\text{CWM}})} \tag{2.4}$$

As can be seen when comparing the numerators of Equations 2.3 and 2.4, the probability of being identified as a carrier of the sensitive attribute is dependent both on the randomization probability r, and its complement, $1 - r$. As can also be seen from the equations, the conditional probability of being identified as a carrier is lower when choosing the first ("both/none true") compared to the second answer option ("one true") when the randomization probability is $0 < r < .5$, because $r < 1 - r$. For $.5 < r < 1$, this reverses to $r > 1 - r$; hence, in these cases, choosing the

second answer option is associated with a lower risk. Importantly, however, the probability of being identified as a carrier of the sensitive attribute exceeds zero regardless of whether the respondent chooses the first or the second answer option for all cases of $0 < \hat{\pi}_{CWM} < 1$, $0 < r < 1$, and $0 < \hat{\lambda}_{CWM} < 1$. In practical applications of the CWM, these conditions are usually met because researchers typically ensure that the expected prevalence of the sensitive attribute, the randomization probability, and the proportion of respondents choosing the first answer option are different from 0% and 100%. In such cases, no CWM answer option provides a "safe" alternative respondents can choose to explicitly deny being a carrier of the sensitive attribute.

Even though a "safe" answer option is unavailable, respondents confronted with a CWM question might still be tempted to try to assess the relative risk of either answer option. To succeed, they would however have to a) correctly estimate the randomization probability (r), and b) derive and understand the relationship between the randomization probability and the conditional probabilities from Equations 2.3 and 2.4. Soeken and Macready (1982) have already demonstrated that with the exception of extreme randomization probabilities, which eliminate confidentiality, respondents are rather poor at estimating the relationship between the randomization probability and the objective privacy protection afforded by the RRT. In light of this finding, as well as the time-consuming computations that would be necessary, we argue that it is quite unlikely that respondents will successfully assess the relative risk of the answer options. Instead, considering that response symmetry reduces the incentive to provide untruthful answers (Ostapczuk, Moshagen, Zhao, & Musch, 2009), we propose that the high symmetry of the two CWM response options will lead to a higher proportion of honest responses compared to a direct question affording no confidentiality and offering a safe answer option that eliminates all risk of being associated with an undesirable behavior.

The validity of the CWM is supported by several weak validation studies in which higher, and therefore presumably more valid, prevalence estimates for sensitive attributes were obtained when using the CWM rather than DQ. The attributes investigated in these validation studies included crossing the street on a "No Walk" signal in plain view of children (Hoffmann, Meisters, & Musch, 2019), xenophobia (Hoffmann & Musch, 2016), the use of anabolic steroids among bodybuilders (Nakhaee, Pakravan, & Nakhaee, 2013), distrust in the Trust Game (Thielmann, Heck, & Hilbig, 2016), plagiarism (Jann, Jerke, & Krumpal, 2012), tax evasion (Korndörfer, Krumpal, & Schmukle, 2014; Kundt, Misch, & Nerré, 2017), prejudice against female leaders (Hoffmann & Musch, 2019), and the intention to vote for the German right-wing party *Alternative for Germany* (Waubert de Puiseau, Hoffmann, & Musch, 2017). Additionally, in a first strong validation study, the CWM provided highly accurate prevalence estimates for the known prevalence of an experimentally induced sensitive attribute while DQ provided a severe underestimate (Hoffmann, Diedenhofen, Verschuere, & Musch, 2015). The CWM has also been proven to be more comprehensible than other indirect questioning techniques, and to evoke a higher level of trust than conventional direct questions (Hoffmann, Waubert de Puiseau, Schmidt, & Musch, 2017). A recent study provided evidence that the CWM is quite robust even against deliberate faking, as "fake good" instructions impaired the validity of DQ, but not of CWM prevalence estimates (Hoffmann et al., 2019). This robustness against deliberate faking is likely attributable to respondents' inability to identify a "safe" answer in the symmetric CWM.

Some recent studies have reported a problematic tendency for the CWM to produce false positives, that is, some non-carriers of the sensitive attribute were falsely classified as carriers. This can potentially result in an overestimation of the prevalence of sensitive attributes (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018). In Höglinger and Diekmann (2017),

the prevalence of two sensitive attributes with a known prevalence near zero was overestimated by the CWM (at 5% and 8%, respectively). Höglinger and Jann (2018), also observed false positives in a CWM survey. However, Meisters, Hoffmann, and Musch (2019) found no evidence for an overestimation of the known prevalence of an experimentally induced sensitive attribute, and suggested that the problem of false positive can be addressed by providing respondents with more comprehensive and detailed instructions and by ensuring that they actually comprehend the procedure.

The Triangular Model. In the TRM, respondents are also presented with a sensitive statement A with unknown prevalence π (e.g., "I have taken cocaine") and a nonsensitive statement B with known prevalence r (e.g., "My mother was born in November or December") to which they must provide a joint response. The response options in the TRM are: "none of the statements is true" versus "at least one of the statements (irrespective of which one) is true". A maximum likelihood estimate for the prevalence π is given by (Yu et al., 2008):

$$\hat{\pi}_{TRM} = 1 - \frac{\hat{\lambda}_{TRM}}{1 - r} \tag{3}$$

where $\hat{\lambda}_{TRM}$ is the observed proportion of respondents choosing the first answer option ("none of the statements is true"). As in the CWM, carriers of the sensitive attribute can choose the second answer option ("at least one of the statements (irrespective of which one) is true") without disclosing their true status with respect to the sensitive statement, since this response must also be given by noncarriers of the sensitive attribute who carry the nonsensitive attribute used for randomization (e.g., respondents whose mother was born in November or December). However, in contrast to the CWM, the TRM is an asymmetric model, because the answer option "none of the statements is true" provides a "safe" answer alternative that explicitly precludes being a carrier of the sensitive attribute. Respondents who are eager to distance themselves from the

sensitive attribute may therefore likely be tempted to opt for this safe response option even when told otherwise by the randomization instructions.

The asymmetry of the TRM is reflected in the conditional probabilities of being identified as a carrier of the sensitive attribute given the first ("none of the statements is true") versus the second answer option ("at least one of the statements (irrespective of which one) is true"):

$$Pr_{\text{TRM}}(\text{carrier} \mid \text{“none true”}) = \frac{Pr_{\text{TRM}}(\text{carrier} \cap \text{“none true”})}{Pr_{\text{TRM}}(\text{“none true”})} \tag{4.1}$$

$$Pr_{\text{TRM}}(\text{carrier} \mid \text{“at least one true”}) = \frac{Pr_{\text{TRM}}(\text{carrier} \cap \text{“at least one true”})}{Pr_{\text{TRM}}(\text{“at least one true”})} \tag{4.2}$$

The reformulation of Equation 4.2 using the parameters for prevalence estimation from Equation 3 is straightforward because in the TRM, the numerator of Equation 4.1 refers to an impossible event. As per the TRM instructions, no carrier of the sensitive attribute may choose the first answer option ("none of the statements is true"), since for carriers, the sensitive statement A is true by definition. Therefore, according to the TRM, the probability of being a carrier when answering "none of the statements is true" is 0:

$$Pr_{\text{TRM}}(\text{carrier} \mid \text{“none true”}) = \frac{0}{\hat{\lambda}_{\text{TRM}}} = 0 \tag{4.3}$$

$$Pr_{\text{TRM}}(\text{carrier} \mid \text{“at least one true”}) = \frac{\hat{\pi}_{\text{TRM}}}{(1 - \hat{\lambda}_{\text{TRM}})} \tag{4.4}$$

As Equation 4.3 shows, respondents can be sure that the first answer option ("none of the statements is true") is associated with a zero probability of being identified as a carrier of the sensitive attribute, irrespective of the randomization probability. Choosing this "safe" answer option is likely to attract respondents who are keen to make a positive or avoid a negative

impression; this in turn is likely to result in underestimates due to dishonest responses (cf. Jerke & Krumpal, 2013).

Research on the validity of the TRM is relatively scarce. Two experimental validation studies have compared a TRM and a DQ control condition (Erdmann, 2019; Jerke & Krumpal, 2013). One study found prevalence estimates for plagiarism obtained via the TRM to descriptively exceed those obtained via DQ. However, this difference was not statistically significant (Jerke & Krumpal, 2013). In a second study, TRM estimates were comparable to and not significantly different from DQ estimates for three different sensitive attributes (Erdmann, 2019). Self-protective answering behavior facilitated by the asymmetric nature of the TRM is a possible explanation for these findings.

Comparison of the CWM and the TRM. In terms of theoretical properties, a potential advantage of the CWM over the TRM is that only the CWM offers response symmetry. If respondents confronted with CWM questions are therefore less tempted to provide evasive answers, or are less successful in identifying a self-protective choice, this should result in prevalence estimates with higher validity. On the other hand, the TRM is usually more efficient than the CWM (with some exceptions for high values of r; cf. Theorem 3 in Yu et al., 2008). Accordingly, the TRM would be preferable to the CWM if both models were equally valid. However, we argue that the validity of the prevalence estimates is even more important than the efficiency of parameter estimation.

The only existing evidence regarding this question is based on a comparison across studies. The prevalence of plagiarism has been assessed both with the TRM (Jerke & Krumpal, 2013) and the CWM (Jann et al., 2012). Comparing the two results reveals that the CWM estimate significantly exceeded the DQ estimate, whereas the TRM and DQ estimates were not

significantly different. Moreover, the CWM estimate (Jann et al., 2012) was descriptively higher than the TRM estimate (Jerke & Krumpal, 2013). This pattern of results tentatively suggests that the symmetrical CWM might be superior to the asymmetric TRM in discouraging dishonest responses and thus in obtaining more valid estimates (cf. Jann et al., 2012; Jerke & Krumpal, 2013). However, a more conclusive comparison would involve directly comparing the CWM and the TRM in a single sample using an experimental design, and thus allowing alternative explanations for the observed differences in prevalence estimates to be ruled out. Taking up this challenge, the current study extends the existing body of research by providing the first experimental comparison of the validity of the CWM and the TRM and contrasting the performance of the two models to a DQ control condition. Xenophobia and opposition to further refugee admissions were chosen as sensitive attributes for the purpose of this validation study.

Xenophobia and Opposition to Reception of Refugees in Germany

Xenophobia, a fear of - or negative attitude towards - foreigners, is quite prevalent in Germany (Heitmeyer, 2012; Krumpal, 2012; Wagner & van Dick, 2001). Since the so-called "refugee crisis" of 2015, attitudes towards foreigners in general and refugees in particular have become more negative among the German population (Bertelsmann Stiftung, 2017). A representative survey revealed that 54% of the German population opposes the further intake of refugees, whereas most Germans still perceive a "welcoming culture" in Germany (Bertelsmann Stiftung, 2017). This discrepancy is likely to lead to social pressure to deny xenophobic attitudes and endorse further refugee admissions (Zick, Hövermann, & Krause, 2012). Direct self-reports on xenophobic attitudes and reluctance to grant asylum to refugees have indeed been found to be distorted by social desirability bias, leading to underestimates of their prevalence (D'Ancona, 2013; Krumpal, 2012; Moshagen & Musch, 2012; Ostapczuk, Musch, & Moshagen, 2009).

Indirect questioning techniques have been shown to lead to higher, and thus presumably more valid, estimates of both the prevalence of xenophobic attitudes (Hoffmann & Musch, 2016; Krumpal, 2012; Ostapczuk, Musch, et al., 2009) and opposition to further refugee admissions (Moshagen & Musch, 2012).

We expected more xenophobic attitudes and greater opposition to further refugee admissions among respondents with a right-oriented vs. a left-oriented political orientation, as a right-oriented political orientation has been shown to be positively associated with xenophobic attitudes (cf. Alba & Johnson, 2000; Zick, Küpper, & Hövermann, 2011). However, we also expected that the perceived sensitivity of these attitudes might vary as a function of political orientation. Right-oriented respondents may be more willing to openly express their disapproval of foreigners and refugees, whereas left-oriented respondents might feel hesitant to openly admit to a xenophobic attitude, and therefore choose to respond in a socially desirable rather than truthful manner. Therefore, the differences in prevalence estimates obtained via direct vs. indirect questioning were used not only to assess the validity of the competing non-randomized response techniques, but also to investigate the influence of political orientation on question sensitivity.

The current study

This study is the first to experimentally compare the validity of two NRRTs (symmetric CWM and asymmetric TRM) and contrast their performance to conventional direct questioning (DQ). We expected that both NRRTs would outperform DQ in terms of delivering more valid prevalence estimates for two socially undesirable attributes. We also expected a beneficial effect of response symmetry (cf. Ostapczuk, Moshagen, et al., 2009), and therefore predicted that the symmetric CWM would outperform the asymmetric TRM with respect to a successful control of

social desirability bias. Furthermore, we investigated a potential moderating influence of self-ascribed political orientation (from “left” to “right”) on prevalence estimates for xenophobia and opposition to the further intake of refugees. Finally, a nonsensitive control attribute with known prevalence was included to test for method-specific biases in the form of a general tendency towards over- or underestimation. If the CWM and the TRM allow to obtain prevalence estimates for the control attribute that correspond to DQ estimates and to the known prevalence, this provides strong evidence for the validity of these indirect questioning techniques.

Method

Sample

The initial sample consisted of 1544 students from the University of Düsseldorf. Due to nonresponse to at least one of the experimental, demographic, or political orientation questions, 162 respondents (10.49%) had to be excluded from further analyses. Dropout rates were unaffected by experimental condition [condition 1: 8.27%, condition 2: 12.65%, condition 3: 9.77%, condition 4: 12.17%, condition 5: 10.00%, condition 6: 10.08%; $X^2(5) = 3.64$, $p = .603$, *Cramer-V* = .05], age [final sample: $M = 21.40$, dropouts: $M = 21.78$; $t(1515) = 0.71$, $p = .478$, $d = 0.06$] and political orientation [final sample: $M = -0.88$, dropouts: $M = -0.95$; $t(1445) = -0.32$, $p = .748$, $d = 0.04$]. The final sample consisted of 1382 respondents (60.1% female) with a mean age of 21.40 years ($SD = 5.66$). Respondents were contacted on the university campus prior to the start of lectures and asked to complete a short one-page survey. This survey study was carried out in accordance with the revised Declaration of Helsinki (World Medical Association, 2013) and the ethical guidelines of the German Society for Psychology (Berufsverband Deutscher Psychologinnen und Psychologen & Deutsche Gesellschaft für Psychologie, 2016). All respondents were informed of the purpose of the study and the strict anonymization of all data

prior to participation, and consented to participate on a voluntary basis without receiving any financial compensation.

Survey Design

The one-page paper-pencil questionnaire contained three experimental questions, one question concerning respondents' self-reported political orientation, and demographic questions asking for the respondents' age and gender. The first two experimental questions asked about the two sensitive attributes (xenophobia and opposition to further refugee admissions). The third question referred to a nonsensitive control attribute and asked about the first letter of the respondents' surname. The prevalence of this attribute is known to be $\pi_{control} = 22\%$ in Germany according to official statistics (Reinders, 1996), and was also confirmed by the university's student registry. All respondents were presented with all three experimental questions. The order of the sensitive attributes the experimental questions referred to was fixed (question 1: xenophobia, question 2: opposition to refugee admission, question 3: first letter of surname K, L, M). The order of the questioning techniques (CWM, TRM, DQ) that were assigned to each experimental question was randomized. This resulted in six different experimental conditions to which respondents were assigned randomly. An overview of questions, questioning techniques and number of respondents by experimental condition is given in Table 1. This design allowed us to manipulate and analyze the questioning technique as a between-subjects variable for each sensitive attribute. After data collection, responses were pooled across experimental conditions to obtain answer frequencies for all three questioning techniques and for each experimental question. Question 1 (xenophobia) was answered in CWM format by 233 + 221 = 454 respondents (32.85%), in TRM format by 231 + 231 = 462 respondents (33.43%), and in DQ format by 234 + 232 = 466 respondents (33.72%). Question 2 (opposition to refugee admission)

was answered in CWM format by 231 + 234 = 465 respondents (33.65%), in TRM format by 233 + 232 = 465 respondents (33.65%), and in DQ format by 221 + 231 = 452 respondents (32.71%). Question 3 (first letter of surname: K, L, M) was answered in CWM format by 231 + 232 = 463 respondents (33.50%), in TRM format by 221 + 234 = 455 respondents (32.92%), and in DQ format by 233 + 231 = 464 respondents (33.57%).

[INSERT TABLE 1]

Sensitive statements. The sensitive statement used to measure xenophobia read: "I would mind if my 20-year-old daughter had a relationship with a Turkish man." It was adapted from Bogardus' social distance scale (Bogardus, 1933) and had previously been used by Hoffmann and Musch (2016) in this form and by Jimenez (1999); Ostapczuk, Musch, et al. (2009) and Silbermann and Hüsers (1995) with respect to other minority groups. The sensitive statement regarding opposition to further intake of refugees read: "Germany has already received more than enough refugees."

CWM format. In the CWM format, two statements were presented simultaneously: One of the two sensitive statements and a nonsensitive control statement with known prevalence r (father's month of birth: r =.158 due to official birth statistics; Pötzsch, 2012).

> Statement A: "I would mind if my 20-year-old daughter had a relationship with a Turkish man."
>
> Statement B: "My father was born in November or December."

Respondents could choose from the two answer options "both statements are true or both statements are false" and "exactly one statement is true (irrespective of which one)". The statements regarding the two other topics were adapted accordingly.

TRM format. In the TRM format, two statements were presented simultaneously: The remaining sensitive statement and a nonsensitive control statement with known prevalence *r* (mother's month of birth: *r* =.158 due to official birth statistics; Pötzsch, 2012)

Statement A: "Germany has already received more than enough refugees."

Statement B: "My mother was born in November or December."

Respondents could choose from the two answer options "both statements are false" and "at least one statement is true (irrespective of which one)". The statements regarding the two other topics were adapted accordingly.

DQ format. In the DQ format, the nonsensitive control statement with known prevalence (π = 22%, Reinders, 1996) read as follows: "My surname begins with one of the following letters: K, L, M." Respondents then had to indicate whether the statement was "true" or "false". The statements regarding the two sensitive attributes were presented in the same way.

Political orientation. To assess respondents' political orientation, we presented the question: "Political beliefs are often labeled as rather 'left' or rather 'right'. Where on that scale would you place yourself?" Responses were recorded on an 11-point Likert scale from "left" (-5) to "right" (+5).

Statistical analyses

To obtain and compare prevalence estimates for the three attributes under investigation, we formulated multinomial processing tree models for all three questioning techniques (Batchelder, 1998; Batchelder & Riefer, 1999), following the procedure detailed in works such as Moshagen, Hilbig, and Musch (2011); Moshagen, Musch, and Erdfelder (2012); and Ostapczuk, Musch, and Moshagen (2011). The parameter π referred to the prevalence of the attribute to be estimated. In the CWM and the TRM question formats, the parameter *r* referred to the known

prevalence of the nonsensitive attributes used for randomization, that is, the respondents' mother or father being born in November or December. Official statistics on more than 2.3 million births in Germany provided by the Federal Statistical Office show that of all births over the course of a year, about 15.8% take place in November or December (Pötzsch, 2012, p. 17). This number was therefore considered to be the best estimate of the prevalence of the non-sensitive attributes. Consequently, the parameter r was set constant to .158 in both the CWM and TRM format, respectively. Processing tree diagrams for the CWM, TRM and DQ formats are shown in Figure 1.

[INSERT FIGURE 1]

Based on the empirically observed answer frequencies, prevalence estimates were obtained using the expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977; Hu & Batchelder, 1994) as implemented in the software multiTree (Moshagen, 2010). Model fit was assessed via the asymptotically X^2-distributed log-likelihood statistic G^2 as detailed in, for example, Ostapczuk, Musch, et al. (2009), Moshagen et al. (2011), and Hoffmann and Musch (2016). The multinomial processing tree models for all three questioning techniques per attribute were saturated with $df = 0$ and $G^2 = 0$ since the number of independent answer categories was just sufficient to estimate all parameters in the three questioning technique conditions. For each of the attributes under investigation (xenophobia, opposition to further refugee admission, first letter of surname: K, L, M) three prevalence estimates were obtained (CWM, TRM, and DQ) based on the response frequencies from three independent, non-overlapping groups (cf. Table 1). For example, one independent proportion of respondents answering "both/none true" to the xenophobia question in CWM format, another independent proportion of respondents answering "none true" to the xenophobia question in TRM format, and a third independent proportion of

respondents answering “true” to the xenophobia question in DQ format allowed us to obtain completely independent estimates for the prevalence of xenophobia for each of these questioning technique formats (π_{CWM}, π_{TRM}, and π_{DQ}). To compare these prevalence estimates, we assessed the difference in model fit (ΔG^2) between an unrestricted baseline model and a restricted alternative model in which the respective parameters were equalized or set to be constant (e.g. $\pi_{CWM} = \pi_{DQ}$ or $\pi_{CWM} = .22$).

To analyze the influence of political orientation, we split the sample into two independent, non-overlapping groups of left- versus right-oriented respondents via their answers to the Likert-scaled item on self-ascribed political orientation (from “left” = -5 to “right” = +5). For each of these groups, we established separate multinomial processing trees to obtain and compare prevalence estimates for the two sensitive attributes (xenophobia and an opposition to a further admission of refugees) following the procedure detailed above. Both political orientation (left, right) and questioning technique format (CWM, TRM, DQ) varied between-subjects. This allowed us to conduct pairwise comparisons of prevalence estimates between political orientation groups within a specific questioning technique format (e.g. $\pi_{CWM, left}$ versus $\pi_{CWM, right}$), and between questioning technique formats within a specific political orientation group (e.g. $\pi_{CWM, left}$ versus $\pi_{TRM, left}$) by assessing the difference in model fit (ΔG^2) between an unrestricted baseline model and a restricted alternative model in which the respective parameters were equalized or set to be constant (e.g. $\pi_{CWM, left} = \pi_{CWM, right}$ or $\pi_{CWM, left} = \pi_{TRM, left}$).

To assess interactions between questioning technique and political orientation on the prevalence estimates for the two sensitive attributes, we introduced parametric order constraints by reparameterizing the original multinomial models established for estimating the prevalence of xenophobia, and the prevalence of an opposition to further refugee admission, respectively

(Hoffmann & Musch, 2019; Knapp & Batchelder, 2004). Within each of these models, we replaced the parameter used for estimating the prevalence among left-oriented respondents in the DQ condition ($\pi_{DQ, left}$) with the corresponding parameter for right-oriented respondents ($\pi_{DQ, right}$), multiplied by a shrinkage factor ($\alpha_{DQ, left:right}$); the CWM and TRM conditions were reparameterized in an analogous way ($\pi_{CWM, left} = \pi_{CWM, right} * \alpha_{CWM, left:right}$; $\pi_{TRM, left} = \pi_{TRM, right} * \alpha_{TRM, left:right}$). The shrinkage factors $\alpha_{DQ, left:right}$, $\alpha_{CWM, left:right}$ and $\alpha_{TRM, left:right}$ thus represent the ratio of the prevalence for politically left-oriented to the prevalence for politically right-oriented respondents in the DQ, CWM and TRM conditions, respectively. For example, for the question on xenophobia, a shrinkage factor of $\alpha_{DQ, left:right} = 33\%$ means that in the DQ condition, respondents who label themselves as politically "left" are only .33 times as likely to admit to the xenophobic attitude as respondents who label themselves as politically "right". To test for interactions between questioning technique and political orientation, the shrinkage factors were compared using the ΔG^2 statistic, as described above (e.g. $\alpha_{DQ, left:right} = \alpha_{CWM, left:right}$). In this analysis, significant changes in model fit indicate significant interactions (Hoffmann & Musch, 2019). MultiTree model equations and empirically observed answer frequencies are available in appendices A and B in the electronic supplementary material.

Results

Table 2 contains the prevalence estimates and test statistics for parameter comparisons for both the two sensitive attributes and the nonsensitive control attribute obtained via the different questioning techniques.

[INSERT TABLE 2]

Xenophobia (Sensitive Attribute 1)

Prevalence estimates for xenophobic responses were significantly higher when assessed via the CWM (31.65%) than when assessed via the TRM (20.05%) or DQ (15.45%). The TRM resulted in descriptively but not significantly higher prevalence estimates for xenophobia than DQ. This finding suggests that the prevalence of xenophobia is presumably underestimated in DQ and TRM, while the CWM seems to successfully control for socially desirable responding.

Opposition to Refugee Intake (Sensitive Attribute 2)

The prevalence estimates for opposition to further intake of refugees obtained via the CWM (43.56%) were descriptively but not significantly higher than those obtained via the TRM (37.43%), or DQ (36.73%). Prevalence estimates obtained via TRM and DQ did not differ significantly, either.

Nonsensitive control attribute with known prevalence: First letter of surname

For the first letter of the respondents' surnames as a nonsensitive control attribute, all questioning techniques obtained similar prevalence estimates that did not differ from the known prevalence of 22% (CWM: 23.32%, TRM: 22.22%, DQ: 24.35%). These highly accurate prevalence estimates suggest that none of the questioning techniques under investigation was subject to systematic bias in the form of a general tendency towards over- or underestimation.

Exploratory moderator analyses

A median split of the sample by self-reported political orientation revealed a moderating influence on prevalence estimates for xenophobia and opposition to further refugee admissions (see Table 3).

[INSERT TABLE 3]

As expected, the prevalence estimates for both sensitive attributes were higher among politically right-oriented than among politically left-oriented respondents for all questioning techniques. For xenophobia, the prevalence estimates obtained via TRM and DQ were significantly higher for right-oriented than for left-oriented respondents; in the CWM, however, prevalence estimates varied only slightly as a function of political orientation. For opposition to refugee intake, all three questioning techniques provided significantly higher prevalence estimates among politically right-oriented compared to politically left-oriented respondents. In the CWM condition, however, this difference was descriptively smaller than in the two other conditions. Interaction analyses revealed a significant interaction between questioning technique and political orientation for both sensitive attributes. Shrinkage factors indicated that the difference in prevalence estimates between politically right-oriented respondents and politically left-oriented respondents was significantly higher in the DQ than in the CWM condition (see Figure 2).

[INSERT FIGURE 2]

In the subsample of left-oriented respondents, prevalence estimates in the CWM condition were higher and thus presumably more valid than prevalence estimates in the DQ condition. In the subsample of right-oriented respondents, however, the CWM estimates only slightly and insignificantly exceeded those obtained via DQ. Thus, social desirability bias seems to have exerted a substantially stronger influence on politically left-oriented compared to politically right-oriented respondents. This pattern of result suggests higher perceived sensitivity of the questions measuring xenophobia and opposition to refugee admissions among politically left-oriented compared to politically right-oriented respondents.

Discussion

In this study, we conducted the first experimental comparison of the validity of a symmetric (crosswise model; CWM) and an asymmetric NRRT (triangular model; TRM), and contrasted the performance of these two models to a conventional direct questioning approach (DQ). To this end, we assessed respondents' attitudes towards xenophobia and opposition to the further intake of refugees in Germany with two sensitive statements of unknown prevalence. Additionally, following a "strong" validation approach, we included a nonsensitive control statement (first letter of respondents' surnames) with known prevalence to test for potential method-specific biases leading to a general tendency towards over- or underestimation.

As expected, both NRRTs yielded higher estimates for xenophobia than DQ. However, only the CWM provided estimates that were significantly higher than the estimates obtained via DQ, thus sufficing the "more is better" criterion. Moreover, the CWM estimates were significantly higher than the TRM estimates, indicating the superiority of the symmetrical CWM over the asymmetrical TRM. For opposition to further refugee admissions, the CWM yielded descriptively higher prevalence estimates than TRM and DQ; however, none of the pairwise comparisons of parameter estimates were significant. All three questioning techniques accurately recovered the known prevalence of the nonsensitive control attribute. Thus, both the CWM and the TRM met the criteria of a successful "strong" validation. As expected, exploratory analyses of the influence of political orientation revealed that prevalence estimates for the sensitive attributes were higher among right-oriented respondents than among left-oriented respondents. Interestingly, interaction analyses showed that this discrepancy was less pronounced when prevalence estimates were obtained via the CWM compared to via the TRM or DQ.

Our results indicate that while both NRRTs outperformed DQ, only the CWM satisfied the "more is better" criterion for xenophobia and was thus better able to control for social desirability with respect to this question than DQ. The estimates obtained via the TRM were descriptively, but not significantly higher than estimates obtained via DQ. None of the questioning techniques under investigation exhibited a method-specific tendency towards over- or underestimation; instead, the prevalence estimates for the nonsensitive control attribute with known prevalence were highly accurate for all questioning techniques. In light of these findings, we recommend that the CWM be used to control for social desirability bias and maximize the validity of prevalence estimates in surveys of sensitive personal attributes.

We also found that for xenophobia, the symmetric CWM outperformed the asymmetric TRM in terms of validity. Response symmetry, or the absence of an objectively "safe" answer option, has been shown to increase the confidentiality of individual answers and thus also compliance with RRT instructions (Ostapczuk, Moshagen, et al., 2009). Response symmetry seems to prevent respondents from faking their answers (cf. Hoffmann et al., 2019), either because they understand that their privacy is perfectly protected and they cannot make a negative impression or because they are simply unable to identify a self-protective response. In contrast, the asymmetric TRM offers a "safe" answer option and therefore seems to be more prone to deliberate faking than the CWM. The TRM offers the theoretical advantage that under many conditions, it provides the more efficient estimates (Yu et al., 2008), as also confirmed by smaller standard errors for TRM compared to CWM estimates in the current study. However, the CWM was found to be clearly preferable to the TRM in terms of the superordinate criterion of measurement validity. Interestingly, our results are in line with the results of two previous studies examining plagiarism, one via the CWM, the other via the TRM (Jann et al., 2012; Jerke &

Krumpal, 2013). The study applying the CWM obtained descriptively higher prevalence estimates than the study applying the TRM. However, this observation is based on a comparison across studies and samples and therefore does not provide unequivocal evidence for the superiority of one of the two models. The present study conducted a first experimental comparison of the two models, and found direct evidence for the assumption first formulated by Jann et al. (2012) and Jerke and Krumpal (2013) that the CWM outperforms the TRM in terms of validity.

In an exploratory analysis, we found a moderating influence of self-reported political orientation on prevalence estimates. For left-oriented respondents, the CWM provided substantially higher prevalence estimates than DQ, indicating that within this subgroup, the prevalence estimates obtained via DQ were strongly distorted by social desirability bias. For right-oriented respondents, the CWM obtained only slightly higher prevalence estimates than DQ, indicating that within this subgroup, social desirability bias had a somewhat weaker impact, as the topics under investigation were likely perceived as less sensitive by respondents less reluctant to openly express negative attitudes towards foreigners. Hence, the CWM proved particularly effective among left-oriented respondents, affirming the assumption that the usefulness of indirect questioning techniques increases with topic sensitivity (cf. Lensvelt-Mulders et al., 2005). Consequently, we particularly recommend that indirect questioning techniques such as the CWM be applied when investigating issues that are highly sensitive for a particular group of respondents, as a strong social desirability influence might result in strongly biased results for such groups if only direct self-reports are used.

Limitations and Future Research Directions

It is necessary to acknowledge that the student population investigated in the present study is not representative of the population at large. Therefore, our pattern of findings and the generalizability of the prevalence estimates obtained in the present study are limited to the sample we investigated, and would need to be replicated in other populations. Estimates for the prevalence of xenophobic attitudes might turn out to be even higher in a more representative sample also including lower-educated respondents, as lower education has repeatedly been shown to be associated with a higher incidence of xenophobic attitudes (D'Ancona, 2013; Hjerm, 2001; Ostapczuk, Musch, et al., 2009; Zick et al., 2011). Student-only samples are also presumably more homogeneous, thereby increasing the statistical power to detect differences between questioning techniques. Lower-educated respondents generally exhibit greater problems understanding indirect questioning techniques (Hoffmann et al., 2017) and therefore tend to produce more false positives (Meisters et al., 2019). Developing better instructions that are easily comprehensible even for lower-educated respondents is therefore of considerable importance for future research using randomized and non-randomized response techniques in more heterogeneous samples.

The results of the current study revealed no method-specific tendency for over- or underestimation. This result is in line with several other studies that also found no deviation between CWM estimates and the known prevalence of a nonsensitive control attribute (Hoffmann & Musch, 2016) and a sensitive attribute (Hoffmann et al., 2015). To check for potential bias in the form of a general preference for one of the two answer options, future studies should try to replicate the present results using the Extended Crosswise Model (ECWM;

Heck, Hoffmann, & Moshagen, 2018), a recent modification of the CWM that allows for detecting instruction nonadherence without negatively affecting statistical efficiency.

As a final remark, it should be noted that the difference between the CWM and DQ condition was larger for self-reported ethnic discrimination than for self-reported opposition to further refugee admissions. A potential explanation for this finding is that opposition to further refugee admissions was perceived by respondents as less sensitive than the expression of xenophobic attitudes. This reasoning is supported by the higher percentage of respondents admitting that they opposed further intake of refugees (36.73%) compared to the much lower percentage of respondents admitting to xenophobia (15.45%) in the DQ condition. The lower perceived sensitivity of opposing further refugee intake might potentially be fueled by the increasing popularity of right-wing populist parties such as the *Alternative for Germany,* which cites social and economic concerns as a reason for limiting further refugee admissions.

Conclusions

The present research showed that non-randomized response techniques provide more valid prevalence estimates for socially undesirable attributes than conventional direct questions (DQ). The crosswise model (CWM) in particular was able to successfully control for the influence of social desirability bias, and outperformed the triangular model (TRM), presumably due to the favorable influence of the response symmetry found in the CWM but not the TRM. We also found that the sensitivity of two questions was contingent on respondents' political orientation, and that the CWM provided the most valid estimates for respondents for whom these questions were most sensitive. Based on these results, we recommend the use of the CWM over the TRM or DQ for topics that are highly sensitive in a survey's target population.

Open Practice Statement

All data and equation files necessary to reproduce the parameter estimates reported in this manuscript are provided in appendices A and B in the electronic supplementary material.

References

Alba, R., & Johnson, M. (2000). Zur Messung aktueller Einstellungsmuster gegenüber Ausländern in Deutschland [On the measurment of current attitudes towards foreigners in Germany]. In R. Alba, P. Schmidt, & M. Wasmer (Eds.), *Blickpunkt Gesellschaft 5. Deutsche und Ausländer: Freunde, Fremde oder Feinde?* (pp. 229-253). Wiesbaden: Westdeutscher Verlag.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10*, 331-344. doi:10.1037/1040-3590.10.4.331

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86. doi:10.3758/Bf03210812

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London, 53*, 370-418.

Bertelsmann Stiftung. (2017). *Willkommenskultur im „Stresstest“ - Einstellungen in der Bevölkerung 2017 und Entwicklungen und Trends seit 2011/2012 - Ergebnisse einer repräsentativen Bevölkerungsumfrage [Welcoming culture under 'stress test' - attitudes in the population in 2017 and developments and trends since 2011/2012 - findings of a representative survey of the population]*. Retrieved from http://www.bertelsmann-stiftung.de/fileadmin/files/Projekte/28_Einwanderung_und_Vielfalt/IB_Umfrage_Willkommenskultur_2017.pdf

Berufsverband Deutscher Psychologinnen und Psychologen & Deutsche Gesellschaft für Psychologie. (2016). Berufsethische Richtlinien des Berufsverbandes Deutscher Psychologinnen und Psychologen e.V. und der Deutschen Gesellschaft für Psychologie e.V. [Professional ethical guidelines of the Berufsverband Deutscher Psychologinnen und Psychologen e.V. and the Deutsche Gesellschaft für Psychologie e.V.]. Retrieved Sep 10th, 2018 https://www.dgps.de/fileadmin/documents/Empfehlungen/berufsethische_richtlinien_dgps.pdf

Bogardus, E. S. (1933). A Social Distance Scale. *Sociology & Social Research, 17*, 265-271.

D'Ancona, M. Á. C. (2013). Measuring xenophobia: Social desirability and survey mode effects. *Migration Studies*, 1-26.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39*, 1-38.

Erdmann, A. (2019). Non-Randomized Response Models: An Experimental Application of the Triangular Model as an Indirect Questioning Method for Sensitive Topics. *methods, data, analyses, 13*, 139-167. doi:10.12758/mda.2018.07

Fligner, M. A., Policello, G. E., & Singh, J. (1977). A Comparison of 2 Randomized Response Survey Methods with Consideration for Level of Respondent Protection. *Communications in Statistics Part a-Theory and Methods, 6*(15), 1511-1524. doi:10.1080/03610927708827593

Heck, D. W., Hoffmann, A., & Moshagen, M. (2018). Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior Research Methods, 50*, 1895-1905. doi:10.3758/s13428-017-0957-8

Heitmeyer, W. (2012). Gruppenbezogene Menschenfeindlichkeit (GMF) in einem entsicherten Jahrzehnt [Group-focused enmity in a disengaged decade]. In W. Heitmeyer (Ed.), *Deutsche Zustände: Folge 10* (pp. 15-41). Berlin: Suhrkamp.

Hjerm, M. (2001). Education, xenophobia and nationalism: A comparative analysis. *Journal of Ethnic and Migration Studies, 27*, 37-60. doi:10.1080/13691830124482

Hoffmann, A., Diedenhofen, B., Verschuere, B. J., & Musch, J. (2015). A strong validation of the Crosswise Model using experimentally induced cheating behavior. *Experimental Psychology, 62*, 403-414. doi:10.1027/1618-3169/a000304

Hoffmann, A., Meisters, J., & Musch, J. (2019). Nothing but the truth? Effects of faking on the validity of the Crosswise Model. [Manuscript submitted for publication].

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: a Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods, 48*, 1032-1046. doi:10.3758/s13428-015-0628-6

Hoffmann, A., & Musch, J. (2019). Prejudice against women leaders: Insights from an indirect questioning approach. *Sex Roles, 80*, 681-692. doi:10.1007/s11199-018-0969-6

Hoffmann, A., Waubert de Puiseau, B., Schmidt, A. F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods, 49*, 1470-1483. doi:10.3758/s13428-016-0804-3

Höglinger, M., & Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis, 25*, 131-137. doi:10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *Plos One, 13*. doi:10.1371/journal.pone.0201770

Hu, X., & Batchelder, W. H. (1994). The Statistical-Analysis of General Processing Tree Models with the Em Algorithm. *Psychometrika, 59*, 21-47. doi:10.1007/Bf02294263

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking Sensitive Questions Using the Crosswise Model. *Public Opinion Quarterly, 76*, 32-49. doi:10.1093/Poq/Nfr036

Jerke, J., & Krumpal, I. (2013). Plagiate in studentischen Arbeiten [Plagiarism in Student Papers]. *methoden, daten, analysen, 7*, 347-368. doi:10.12758/mda.2013.017

Jimenez, P. (1999). Weder Opfer noch Täter - die alltäglichen Einstellungen 'unbeteiligter' Personen gegenüber Ausländern [Neither victim nor offender—the common attitudes of 'non-involved' persons towards foreigners]. In R. Dollase, T. Kliche, & H. Moser (Eds.), *Politische Psychologie der Fremdenfeindlichkeit. Opfer - Täter - Mittäter* (pp. 293–306). Weinheim: Juventa.

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology, 48*, 215-229. doi:10.1016/j.jmp.2004.03.002

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology, 45*, 18-32. doi:10.1016/j.joep.2014.08.001

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research, 41*, 1387-1403. doi:10.1016/j.ssresearch.2012.05.015

Kundt, T. C., Misch, F., & Nerré, B. (2017). Re-assessing the merits of measuring tax evasion through business surveys: an application of the crosswise model. *International Tax and Public Finance, 24*, 112-133. doi:10.1007/s10797-015-9373-0

Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response - A study of a method for improving the

estimate of social security fraud. *Quality & Quantity, 33*, 1-12. doi:10.1023/A:1004361819974

Lanke, J. (1976). Degree of Protection in Randomized Interviews. *International Statistical Review, 44*, 197-203. doi:10.2307/1403277

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: thirty-five years of validation. *Sociological Methods & Research, 33*, 319-348. doi:10.1177/0049124104268664

Meisters, J., Hoffmann, A., & Musch, J. (2019). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? [Manuscript under preparation].

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42-54. doi:10.3758/BRM.42.1.42

Moshagen, M., Hilbig, B. E., & Musch, J. (2011). Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *European Journal of Social Psychology, 41*, 638-644. doi:10.1002/Ejsp.793

Moshagen, M., & Musch, J. (2012). Surveying Multiple Sensitive Attributes using an Extension of the Randomized-Response Technique. *International Journal of Public Opinion Research, 24*, 508-523. doi:10.1093/ijpor/edr034

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44*, 222-231. doi:10.3758/s13428-011-0144-2 21858604

Nakhaee, M. R., Pakravan, F., & Nakhaee, N. (2013). Prevalence of Use of Anabolic Steroids by Bodybuilders Using Three Methods in a City of Iran. *Addiction & Health, 5*, 77–82.

Ostapczuk, M., Moshagen, M., Zhao, Z., & Musch, J. (2009). Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. *Journal of Educational and Behavioral Statistics, 34*, 267-287. doi:10.3102/1076998609332747

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology, 39*, 920-931. doi:10.1002/ejsp.588

Ostapczuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. *Statistical Methods in Medical Research, 20*, 489-503. doi:10.1177/0962280210372843

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes, Vol. 1* (pp. 17-59). San Diego, CA: Academic Press.

Pötzsch, O. (2012). Geburten in Deutschland [Births in Germany]. Retrieved Jun 6, 2012, from German Federal Statistical Office, https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/Bevoelkerungsbewegung/BroschuereGeburtenDeutschland0120007129004.pdf

Reinders, M. (1996). Häufigkeit von Namensanfängen [Frequency of first letters of surnames]. *Statistische Rundschau Nordrhein-Westfalen, 11*, 651-660.

Silbermann, A., & Hüsers, F. (1995). *Der 'normale' Haß auf die Fremden. Eine sozialwissenschaftliche Studie zu Ausmaß und Hintergründen von Fremdenfeindlichkeit in Deutschland [The 'normal' xenophobia. A socio-scientific study on the extent and determinants of xenophobia in Germany]*. München: Quintessenz.

Soeken, K. L., & Macready, G. B. (1982). Respondents Perceived Protection When Using Randomized-Response. *Psychological Bulletin, 92*, 487-489. doi:10.1037/0033-2909.92.2.487

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. *Judgment and Decision Making, 11*, 527-536.

Tian, G.-L., & Tang, M.-L. (2014). *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883. doi:10.1037/0033-2909.133.5.859 17723033

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking Sensitive Questions: A Statistical Power Analysis of Randomized Response Models. *Psychological Methods, 17*, 623-641. doi:10.1037/A0029314

Umesh, U. N., & Peterson, R. A. (1991). A Critical Evaluation of the Randomized-Response Method - Applications, Validation, and Research Agenda. *Sociological Methods & Research, 20*, 104-138. doi:10.1177/0049124191020001004

Wagner, U., & van Dick, R. (2001). Fremdenfeindlichkeit "in der Mitte der Gesellschaft": Phänomenbeschreibung, Ursachen, Gegenmaßnahmen [Xenophobia "in the middle of society": description of phenomena, causes, countermeasures]. *Zeitschrift für Politische Psychologie, 9*, 41-54.

Warner, S. L. (1965). Randomized-Response - a Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*, 63-69.

Waubert de Puiseau, B., Hoffmann, A., & Musch, J. (2017). How indirect questioning techniques may promote democracy: A pre-election polling experiment. *Basic And Applied Social Psychology, 39*, 209-217. doi:10.1080/01973533.2017.1331351

World Medical Association. (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA, 310*, 2191-2194. doi:10.1001/jama.2013.281053

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika, 67*, 251-263. doi:10.1007/s00184-007-0131-x

Zick, A., Hövermann, A., & Krause, D. (2012). Die Abwertung von Ungleichwertigen - Erklärung und Prüfung eines erweiterten Syndroms der Gruppenbezogenen Menschenfeindlichkeit [The devaluation of inferiors - Explaining and examining an expanded syndrome of group-focused enmity]. In W. Heitmeyer (Ed.), *Deutsche Zustände: Folge 10* (pp. 64-86). Berlin: Suhrkamp.

Zick, A., Küpper, B., & Hövermann, A. (2011). *Intolerance, Prejudice and Discrimination: A European Report*. Berlin: Friedrich-Ebert-Stiftung.

Tables

Table 1

Questions, questioning techniques and number of respondents by experimental condition.

	Experimental Condition					
	1	2	3	4	5	6
Question 1 (xenophobia)	CWM	CWM	TRM	TRM	DQ	DQ
Question 2 (opposition to refugee admission)	TRM	DQ	CWM	DQ	CWM	TRM
Question 3 (first letter of surname: K, L, M)	DQ	TRM	DQ	CWM	TRM	CWM
Number of respondents	233 (16.86%)	221 (15.99%)	231 (16.71%)	231 (16.71%)	234 (16.93%)	232 (16.79%)

Note: CWM = Crosswise Model; TRM = Triangular Model; DQ = Direct Questioning.

Table 2

Parameter estimates (standard errors in parentheses) and parameter comparisons for the sensitive and nonsensitive attributes by questioning technique.

Parameter estimates	CWM	TRM	DQ
Xenophobia	31.65 (3.32)	20.05 (2.59)	15.45 (1.67)
Opposition to refugee admissions	43.56 (3.38)	37.43 (2.75)	36.73 (2.27)
First letter of surname: K, L, M	23.32 (3.16)	22.22 (2.65)	24.35 (1.99)
Parameter comparisons	|*difference*|	Model fit ΔG^2 (*df* = 1)	*p*
Xenophobia			
$\hat{\pi}_{CWM} = \hat{\pi}_{TRM}$	11.60	7.65	.006 *
$\hat{\pi}_{CWM} = \hat{\pi}_{DQ}$	16.20	19.61	< .001 *
$\hat{\pi}_{TRM} = \hat{\pi}_{DQ}$	4.60	2.24	.135
Opposition to refugee admissions			
$\hat{\pi}_{CWM} = \hat{\pi}_{TRM}$	6.13	1.99	.159
$\hat{\pi}_{CWM} = \hat{\pi}_{DQ}$	6.83	2.82	.093
$\hat{\pi}_{TRM} = \hat{\pi}_{DQ}$	0.70	0.04	.844
First letter of surname: K, L, M			
$\hat{\pi}_{CWM} = \hat{\pi}_{TRM}$	1.10	0.07	.789
$\hat{\pi}_{CWM} = \hat{\pi}_{DQ}$	1.03	0.08	.782
$\hat{\pi}_{TRM} = \hat{\pi}_{DQ}$	2.13	0.42	.519
$\hat{\pi}_{CWM} = \pi$ (22%)	1.32	0.18	.675
$\hat{\pi}_{TRM} = \pi$ (22%)	0.22	0.01	.935
$\hat{\pi}_{DQ} = \pi$ (22%)	2.35	1.46	.227

* significant at $p < .05$.

Table 3

Parameter estimates (standard errors in parentheses) and parameter comparisons for the sensitive attributes by questioning technique and political orientation.

Parameter estimates	CWM	TRM	DQ
Xenophobia			
$\hat{\pi}$ left political orientation	28.67 (4.84)	13.67 (3.51)	7.37 (1.77)
$\hat{\pi}$ right political orientation	34.19 (4.56)	26.21 (3.76)	22.49 (2.65)
$\hat{\alpha}$ shrinkage left : right	83.85 (18.03)	52.16 (15.34)	32.79 (8.78)
Opposition to refugee admissions			
$\hat{\pi}$ left political orientation	35.79 (4.88)	25.28 (3.93)	22.32 (2.78)
$\hat{\pi}$ right political orientation	50.29 (4.63)	47.69 (3.71)	50.88 (3.31)
$\hat{\alpha}$ shrinkage left : right	71.16 (11.71)	53.02 (9.22)	43.87 (6.17)

Parameter comparisons				|*difference*|	Model fit ΔG^2 ($df = 1$)	p
Xenophobia							
	$\hat{\pi}_{CWM}$	=	$\hat{\pi}_{TRM}$		15.00	6.41	.011 *
"left"	$\hat{\pi}_{CWM}$	=	$\hat{\pi}_{DQ}$		21.30	18.62	< .001 *
	$\hat{\pi}_{TRM}$	=	$\hat{\pi}_{DQ}$		6.30	2.63	.105
	$\hat{\pi}_{CWM}$	=	$\hat{\pi}_{TRM}$		7.98	1.83	.176
"right"	$\hat{\pi}_{CWM}$	=	$\hat{\pi}_{DQ}$		11.70	5.00	.025 *
	$\hat{\pi}_{TRM}$	=	$\hat{\pi}_{DQ}$		3.72	0.66	.417
CWM	$\hat{\pi}_{left}$	=	$\hat{\pi}_{right}$		5.52	0.69	.407
TRM	$\hat{\pi}_{left}$	=	$\hat{\pi}_{right}$		12.54	5.88	.015 *
DQ	$\hat{\pi}_{left}$	=	$\hat{\pi}_{right}$		15.12	21.50	< .001 *
$\hat{\alpha}_{CWM,\ left:right}$		=	$\hat{\alpha}_{TRM,\ left:right}$		31.69	1.79	.181
$\hat{\alpha}_{CWM,\ left:right}$		=	$\hat{\alpha}_{DQ,\ left:right}$		51.06	7.54	.006 *
$\hat{\alpha}_{TRM,\ left:right}$		=	$\hat{\alpha}_{DQ,\ left:right}$		19.37	1.29	.257
Opposition to refugee admissions							

	$\hat{\pi}_{CWM}$	= $\hat{\pi}_{TRM}$	10.51	2.83	.093
"left"	$\hat{\pi}_{CWM}$	= $\hat{\pi}_{DQ}$	13.47	5.84	.016 *
	$\hat{\pi}_{TRM}$	= $\hat{\pi}_{DQ}$	2.96	0.38	.538
	$\hat{\pi}_{CWM}$	= $\hat{\pi}_{TRM}$	2.60	0.19	.660
"right"	$\hat{\pi}_{CWM}$	= $\hat{\pi}_{DQ}$	0.59	0.01	.918
	$\hat{\pi}_{TRM}$	= $\hat{\pi}_{DQ}$	3.19	0.41	.521
CWM	$\hat{\pi}_{left}$	= $\hat{\pi}_{right}$	14.50	4.60	.032 *
TRM	$\hat{\pi}_{left}$	= $\hat{\pi}_{right}$	22.41	16.60	< .001 *
DQ	$\hat{\pi}_{left}$	= $\hat{\pi}_{right}$	28.56	40.49	< .001 *
$\hat{\alpha}_{CWM,\ left:right}$	=	$\hat{\alpha}_{TRM,\ left:right}$	18.14	1.52	.217
$\hat{\alpha}_{CWM,\ left:right}$	=	$\hat{\alpha}_{DQ,\ left:right}$	27.29	4.75	.029 *
$\hat{\alpha}_{TRM,\ left:right}$	=	$\hat{\alpha}_{DQ,\ left:right}$	9.15	0.70	.404

* significant at $p < .05$.

Figures

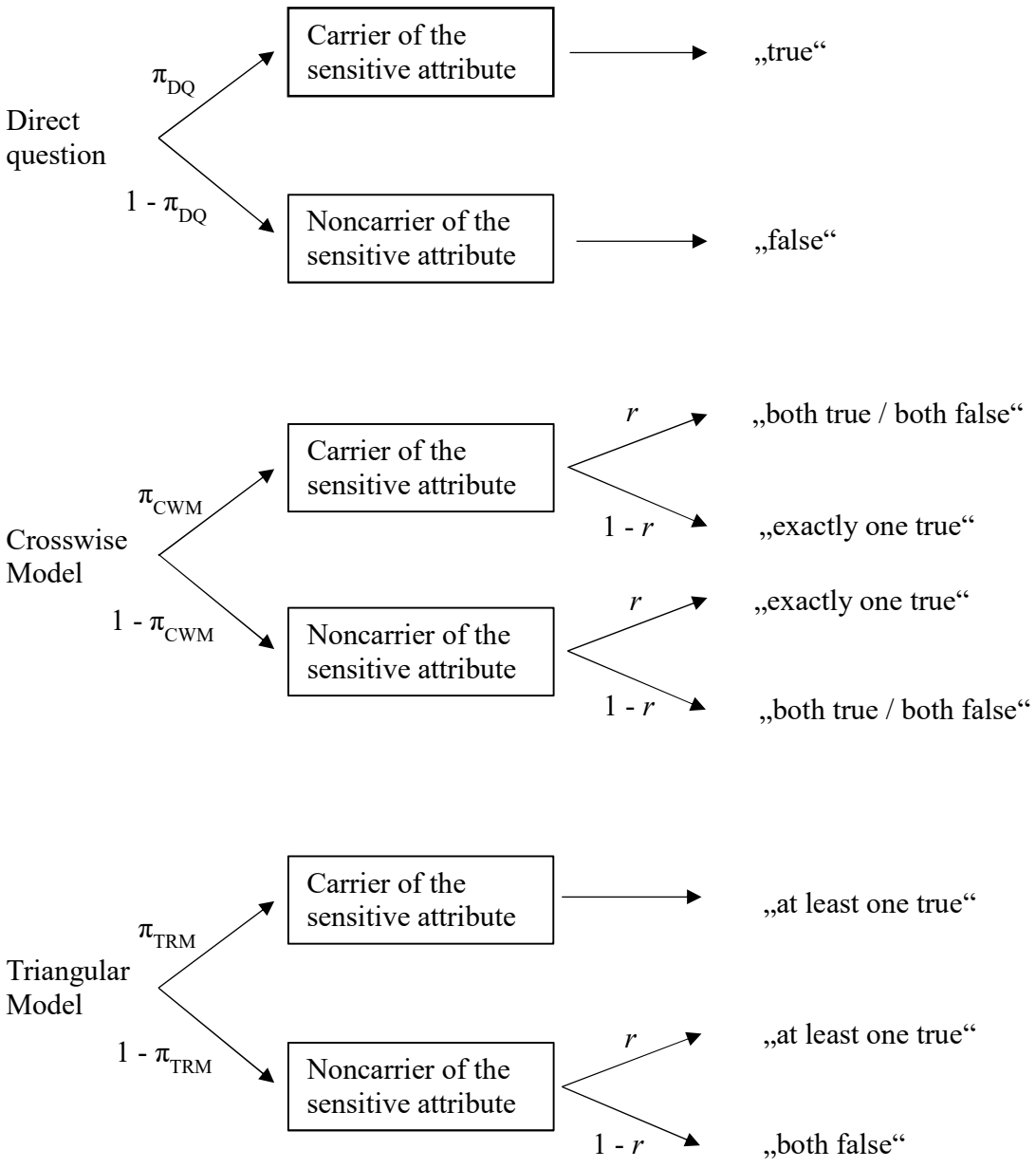


Figure 1. Tree diagram for direct questions and questions posed according to the crosswise model and the triangular model. The parameter π represents the unknown prevalence of the sensitive attribute and the parameter *r* represents the known randomization probability.

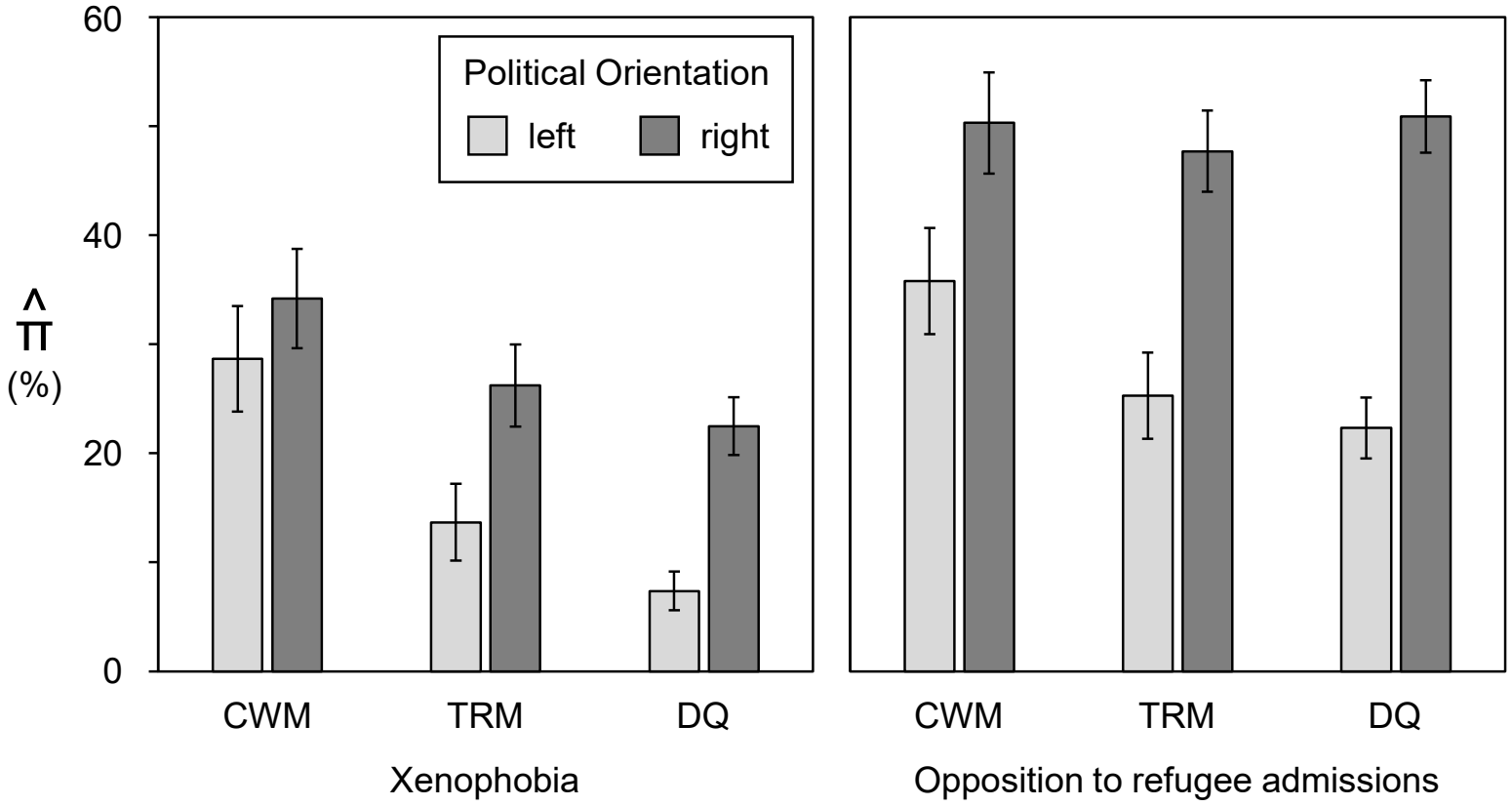


Figure 2. Prevalence estimates for xenophobia (left panel) and opposition to refugee admissions in Germany (right panel) by political orientation (median split). CWM = crosswise model, TRM = triangular model, DQ = direct questioning.

Appendix A

MultiTree equations for the estimation of π (pi) in a multinomial model. The (r)andomization parameter r denotes the known probability of being born in November or December, and is set constant to r = .158 in the multinomial model (Pötzsch, 2012). CWM = crosswise model, TRM = triangular model, DQ = direct questioning; q1 = Question 1 (xenophobia), q2 = Question 2 (opposition to refugee admissions in Germany); left = respondents with left political orientation, right = respondents with right political orientation.

Total Sample

```
q1_CWM     q1_CWM_bothnone     q1_1PiCWM*r
q1_CWM     q1_CWM_one          q1_1PiCWM*(1-r)
q1_CWM     q1_CWM_one          (1-q1_1PiCWM)*r
q1_CWM     q1_CWM_bothnone     (1-q1_1PiCWM)*(1-r)
q1_TRM     q1_TRM_minone       q1_2PiTRM*r
q1_TRM     q1_TRM_minone       q1_2PiTRM*(1-r)
q1_TRM     q1_TRM_minone       (1-q1_2PiTRM)*r
q1_TRM     q1_TRM_none         (1-q1_2PiTRM)*(1-r)
q1_DQ      q1_DQ_true          q1_3PiDQ
q1_DQ      q1_DQ_false         (1-q1_3PiDQ)
q2_CWM     q2_CWM_bothnone     q2_1PiCWM*r
q2_CWM     q2_CWM_one          q2_1PiCWM*(1-r)
q2_CWM     q2_CWM_one          (1-q2_1PiCWM)*r
q2_CWM     q2_CWM_bothnone     (1-q2_1PiCWM)*(1-r)
q2_TRM     q2_TRM_minone       q2_2PiTRM*r
q2_TRM     q2_TRM_minone       q2_2PiTRM*(1-r)
q2_TRM     q2_TRM_minone       (1-q2_2PiTRM)*r
q2_TRM     q2_TRM_none         (1-q2_2PiTRM)*(1-r)
q2_DQ      q2_DQ_true          q2_3PiDQ
q2_DQ      q2_DQ_false         (1-q2_3PiDQ)
q3_CWM     q3_CWM_bothnone     q3_1PiCWM*r
q3_CWM     q3_CWM_one          q3_1PiCWM*(1-r)
q3_CWM     q3_CWM_one          (1-q3_1PiCWM)*r
q3_CWM     q3_CWM_bothnone     (1-q3_1PiCWM)*(1-r)
q3_TRM     q3_TRM_minone       q3_2PiTRM*r
q3_TRM     q3_TRM_minone       q3_2PiTRM*(1-r)
q3_TRM     q3_TRM_minone       (1-q3_2PiTRM)*r
q3_TRM     q3_TRM_none         (1-q3_2PiTRM)*(1-r)
q3_DQ      q3_DQ_true          q3_3PiDQ
q3_DQ      q3_DQ_false         (1-q3_3PiDQ)
```

ON THE VALIDITY OF NON-RANDOMIZED RESPONSE TECHNIQUES

Split by political orientation

left_q1_CWM	left_q1_CWM_bothnone	left_q1_1PiCWM*r
left_q1_CWM	left_q1_CWM_one	left_q1_1PiCWM*(1-r)
left_q1_CWM	left_q1_CWM_one	(1-left_q1_1PiCWM)*r
left_q1_CWM	left_q1_CWM_bothnone	(1-left_q1_1PiCWM)*(1-r)
left_q1_TRM	left_q1_TRM_minone	left_q1_2PiTRM*r
left_q1_TRM	left_q1_TRM_minone	left_q1_2PiTRM*(1-r)
left_q1_TRM	left_q1_TRM_minone	(1-left_q1_2PiTRM)*r
left_q1_TRM	left_q1_TRM_none	(1-left_q1_2PiTRM)*(1-r)
left_q1_DQ	left_q1_DQ_true	left_q1_3PiDQ
left_q1_DQ	left_q1_DQ_false	(1-left_q1_3PiDQ)
left_q2_CWM	left_q2_CWM_bothnone	left_q2_1PiCWM*r
left_q2_CWM	left_q2_CWM_one	left_q2_1PiCWM*(1-r)
left_q2_CWM	left_q2_CWM_one	(1-left_q2_1PiCWM)*r
left_q2_CWM	left_q2_CWM_bothnone	(1-left_q2_1PiCWM)*(1-r)
left_q2_TRM	left_q2_TRM_minone	left_q2_2PiTRM*r
left_q2_TRM	left_q2_TRM_minone	left_q2_2PiTRM*(1-r)
left_q2_TRM	left_q2_TRM_minone	(1-left_q2_2PiTRM)*r
left_q2_TRM	left_q2_TRM_none	(1-left_q2_2PiTRM)*(1-r)
left_q2_DQ	left_q2_DQ_true	left_q2_3PiDQ
left_q2_DQ	left_q2_DQ_false	(1-left_q2_3PiDQ)
right_q1_CWM	right_q1_CWM_bothnone	right_q1_1PiCWM*r
right_q1_CWM	right_q1_CWM_one	right_q1_1PiCWM*(1-r)
right_q1_CWM	right_q1_CWM_one	(1-right_q1_1PiCWM)*r
right_q1_CWM	right_q1_CWM_bothnone	(1-right_q1_1PiCWM)*(1-r)
right_q1_TRM	right_q1_TRM_minone	right_q1_2PiTRM*r
right_q1_TRM	right_q1_TRM_minone	right_q1_2PiTRM*(1-r)
right_q1_TRM	right_q1_TRM_minone	(1-right_q1_2PiTRM)*r
right_q1_TRM	right_q1_TRM_none	(1-right_q1_2PiTRM)*(1-r)
right_q1_DQ	right_q1_DQ_true	right_q1_3PiDQ
right_q1_DQ	right_q1_DQ_false	(1-right_q1_3PiDQ)
right_q2_CWM	right_q2_CWM_bothnone	right_q2_1PiCWM*r
right_q2_CWM	right_q2_CWM_one	right_q2_1PiCWM*(1-r)
right_q2_CWM	right_q2_CWM_one	(1-right_q2_1PiCWM)*r
right_q2_CWM	right_q2_CWM_bothnone	(1-right_q2_1PiCWM)*(1-r)
right_q2_TRM	right_q2_TRM_minone	right_q2_2PiTRM*r
right_q2_TRM	right_q2_TRM_minone	right_q2_2PiTRM*(1-r)
right_q2_TRM	right_q2_TRM_minone	(1-right_q2_2PiTRM)*r
right_q2_TRM	right_q2_TRM_none	(1-right_q2_2PiTRM)*(1-r)
right_q2_DQ	right_q2_DQ_true	right_q2_3PiDQ
right_q2_DQ	right_q2_DQ_false	(1-right_q2_3PiDQ)

Appendix B

Empirically observed answer frequencies for the attributes used for parameter estimation in MultiTree (Moshagen, 2010). CWM = crosswise model, TRM = triangular model, DQ = direct questioning; q1 = Question 1 (xenophobia), q2 = Question 2 (opposition to refugee admissions in Germany); left = respondents with left political orientation, right = respondents with right political orientation.

Total Sample

q1_CWM_bothnone	284
q1_CWM_one	170
q1_TRM_minone	151
q1_TRM_none	311
q1_DQ_true	72
q1_DQ_false	394
q2_CWM_bothnone	253
q2_CWM_one	212
q2_TRM_minone	220
q2_TRM_none	245
q2_DQ_true	166
q2_DQ_false	286
q3_CWM_bothnone	316
q3_CWM_one	147
q3_TRM_minone	157
q3_TRM_none	298
q3_DQ_true	113
q3_DQ_false	351

Split by political orientation

left_q1_CWM_bothnone	135
left_q1_CWM_one	74
left_q1_TRM_minone	62
left_q1_TRM_none	165
left_q1_DQ_true	16
left_q1_DQ_false	201
left_q2_CWM_bothnone	129
left_q2_CWM_one	87
left_q2_TRM_minone	79
left_q2_TRM_none	134
left_q2_DQ_true	50
left_q2_DQ_false	174
right_q1_CWM_bothnone	149
right_q1_CWM_one	96

```
right_q1_TRM_minone       89
right_q1_TRM_none         146
right_q1_DQ_true          56
right_q1_DQ_false         193
right_q2_CWM_bothnone     124
right_q2_CWM_one          125
right_q2_TRM_minone       141
right_q2_TRM_none         111
right_q2_DQ_true          116
right_q2_DQ_false         112
```

Controlling social desirability bias:

An experimental validation of the extended crosswise model

Julia Meisters*, Adrian Hoffmann*, and Jochen Musch

University of Duesseldorf

«fn» *J. Meisters and A. Hoffmann contributed equally to this work.

Author Note

Julia Meisters, Adrian Hoffmann, and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Grant number 393108549.

Correspondence concerning this article should be addressed to Julia Meisters, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, Floor 00, Room 26, 40225 Duesseldorf, Germany. E-mail: julia.meisters@uni-duesseldorf.de

Abstract

Indirect questioning techniques such as the crosswise model (CWM; Yu, Tian, & Tang, 2008) aim to control for socially desirable responding in surveys on sensitive personal attributes. Recently, the extended crosswise model (ECWM; Heck, Hoffmann, & Moshagen, 2018) has been proposed as an improvement over the original crosswise model. It offers all of the advantages of the CWM while also enabling the detection of systematic response biases. We present the first experimental validation study comparing the ECWM with a direct questioning control condition. In a paper-pencil questionnaire, we investigated the prevalence of campus Islamophobia among 1,361 German university students. The ECWM successfully controlled for socially desirable responding and yielded significantly higher estimates of campus Islamophobia than a direct question. Moreover, an assessment of model fit indicated no systematic response bias, lending further credence to the estimates obtained with the ECWM. Our findings indicate that the ECWM is a promising new indirect questioning technique and highlight the importance of controlling for socially desirable responding when surveying for Islamophobia or other potentially sensitive attitudes.

Keywords: social desirability, indirect questioning, validity, extended crosswise model, Islamophobia.

Controlling social desirability bias:

An experimental validation of the extended crosswise model

Surveys of sensitive personal attributes often rely on self-reports. However, socially desirable responding, that is, the tendency to answer in accordance with social norms rather than truthfully, may result in underestimates of the prevalence of socially undesirable attributes and overestimates of the prevalence of socially desirable attributes (Paulhus, 1991; Tourangeau & Yan, 2007). To address this problem, indirect questioning techniques such as the randomized response technique (RRT; Warner, 1965) have been proposed. Based on an experimental randomization procedure, the RRT provides prevalence estimates of sensitive attributes on the sample level while preserving the confidentiality of individual responses. A comprehensive meta-analysis (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005) confirmed the usefulness of this approach and concluded that the RRT provides more valid prevalence estimates than direct questioning (DQ).

The crosswise model (CWM; Yu et al., 2008) is an improved version of the RRT. It presents respondents with a statement regarding a sensitive behavior or attitude (e.g., “Many Muslim students behave in misogynist ways”) in order to estimate its prevalence π, and a non-sensitive statement with known prevalence p that is used for randomization (e.g., “I was born in November or December”). To preserve confidentiality, respondents are not asked to respond to either of these statements individually, but to choose one of the following two combined response options: “I agree with *both* of the statements or *none* of the statements” versus “I agree with *exactly one* of the statements (irrespective of which one)”. Compared to other indirect questioning techniques, the CWM is easier to comprehend and provide instructions for (Hoffmann, Waubert de Puiseau, Schmidt, & Musch, 2017). The CWM has led to significantly higher and thus - according to the “more is better” criterion - presumably more valid prevalence estimates than direct questions in a number of studies investigating sensitive attributes such as xenophobia (Hoffmann, Meisters, & Musch, 2019b; Hoffmann &

Musch, 2016), plagiarism (Jann, Jerke, & Krumpal, 2012), tax evasion (Korndörfer, Krumpal, & Schmukle, 2014; Kundt, Misch, & Nerré, 2017), the use of anabolic steroids by bodybuilders (Nakhaee, Pakravan, & Nakhaee, 2013), the intention to vote for the far-right German party Alternative for Germany (Waubert de Puiseau, Hoffmann, & Musch, 2017), distrust in the Trust Game (Thielmann, Heck, & Hilbig, 2016), crossing the street on a "Don't Walk" sign in plain sight of children (Hoffmann, Meisters, & Musch, 2019a) and prejudice against female leaders (Hoffmann & Musch, 2019). Moreover, the CWM was able to accurately estimate the prevalence of an experimentally induced sensitive attribute, whereas DQ significantly underestimated the known prevalence of this attribute (Hoffmann, Diedenhofen, Verschuere, & Musch, 2015). Additionally, and unlike DQ, the CWM has been proven robust against attempts to deliberately distort answers in a socially desirable direction (Meisters, Hoffmann, & Musch, 2019).

These positive evaluations of the CWM might be partly attributable to the model's response symmetry. This symmetry encourages honest responding because no 'safe' answer option is available that excludes the possibility that a respondent is a carrier of the sensitive attribute. Recently, however, the validity of the CWM has been questioned because false positives have been observed in CWM surveys (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; Meisters et al., 2019). Like direct self-reports and other indirect questioning techniques, the CWM is based on the assumption that respondents will follow the model's instructions. However, this assumption may be violated under certain conditions (e.g. Clark & Desharnais, 1998; Edgell, Himmelfarb, & Duchan, 1982; Ostapczuk, Musch, & Moshagen, 2009). A potential reason for instruction non-adherence is that some respondents might not understand or trust RRT procedures (Hoffmann et al., 2017; Landsheer, van der Heijden, & van Gils, 1999). Moreover, a specific kind of instruction non-adherence in the CWM is a systematic preference for one of the two answer options, which cannot be detected in the original CWM. Two approaches are available to address the problems of false positives and

instruction non-adherence. First, at least in higher-educated samples, it might be possible to reduce false positives by offering detailed instructions and implementing comprehension checks to ensure that all instructions are properly understood (Meisters et al., 2019). Alternatively, the extended CWM (ECWM; Heck et al., 2018) can be applied to test for systematic response biases.

The ECWM offers all the advantages of the CWM and has identical statistical efficiency in parameter estimation, but additionally enables the identification of systematic preferences for one of the two answer options. The central idea is to apply the CWM to two non-overlapping groups with reversed randomization probabilities *p1* and *p2* (see Figure 1). Since the sensitive statement is identical for both groups to which the respondents are randomly assigned, the prevalence π of the sensitive attribute does not differ between groups ($\pi_{ECWM\_1} = \pi_{ECWM\_2}$). Because two independent response frequencies can be observed in the two groups, the resulting model has one degree of freedom and its fit can therefore be tested. If the prevalence estimates π_{ECWM_1} and π_{ECWM_2} do not significantly differ from one another, they can be pooled and are readily interpretable (Heck et al., 2018). If a model misfit is indicated by significant differences in the prevalence estimates π_{ECWM_1} and π_{ECWM_2}, the prevalence estimates should not be interpreted because the misfit indicates that a substantial share of respondents did not adhere to the instructions and exhibited a systematic preference for one of the two answer options (Heck et al., 2018). The ECWM even allows for detecting systematic preferences for one of the two answer options if this preference occurs only among carriers or among non-carriers of the sensitive attribute. However, the validity of the ECWM has never been evaluated in comparison to a direct questioning control condition. Therefore, the present study presents the first experimental validation of the ECWM using campus Islamophobia as a sensitive attribute.

[INSERT FIGURE 1]

Islamophobia is defined as a negative attitude towards, or fear of, Islam as a religion and people of Muslim faith. Muslims currently experience high levels of prejudice and discrimination in Western societies due to their religious affiliation. Surveys in several European countries show that attitudes towards Muslims are far more negative than attitudes towards members of other religions (Pickel & Yendell, 2016; Pollack, 2014; Yendell, 2013; Zick, Küpper, & Hövermann, 2011). In Germany, people of Muslim faith are often stereotyped as a problematic, delinquent and aggressive minority (Petersen, 2012; Pollack, 2014; Zick et al., 2011). Consequently, attitudes towards Muslims in Germany are even more negative than in other European countries (Savelkoul, Scheepers, van der Veld, & Hagendoorn, 2012), and fears of Islamist terrorism are widespread (Pollack, 2014; Yendell, 2013). A common concern is that Islam promotes intolerance and is therefore incompatible with Western open societies (e.g. Pollack, 2014; Yendell, 2013), particularly with respect to gender equality (Zick et al., 2011). In representative samples of eight European countries, 72%-82% of respondents agreed with the statement “The Muslim views on women are contrary to our values” (Zick et al., 2011) and in two representative German samples, more than 80% of respondents believed that Islam is characterized by discrimination against women (Petersen, 2012; Pollack, 2014).

Blatantly negative attitudes towards foreigners or Muslims have increasingly been replaced by more subtle forms of prejudice (e.g. Ganter, 2001; Meertens & Pettigrew, 1997; Pettigrew & Meertens, 1995). This is presumably because many prejudiced people are aware of the social undesirability of their attitudes, and therefore refrain from admitting them publicly (Stocké, 2007). Therefore, estimates for the prevalence of Islamophobia based on direct self-reports are likely underestimates of the true value, and indirect questioning techniques may help to obtain more valid estimates (Hoffmann et al., 2019b; Hoffmann & Musch, 2016; Krumpal, 2012; Ostapczuk et al., 2009). If the ECWM is suitable as a new

means of controlling for social desirability, it should provide higher and thus presumably more valid estimates for the prevalence of Islamophobia than a conventional direct question.

The results of numerous studies indicate that explicit prejudice against Muslims and other minorities is less common among better educated (Coenders & Scheepers, 2003; Easterbrook, Kuppens, & Manstead, 2016; Ostapczuk et al., 2009; Wagner & Zick, 1995) and younger respondents (Ganter, 2001; Strabag & Listhaug, 2008). In a study by Stocké (2007), younger and better educated respondents also reported higher perceived pressure to answer in accordance with social norms. However, even when controlling for social desirability, prejudice seems to be less prevalent among higher-educated compared to lower-educated respondents (Ostapczuk et al., 2009; Stocké, 2007; Wagner & Zick, 1995). Nevertheless, in a study by Kassis, Schallié, Strube, and van der Heyde (2014), more than 20% of respondents from a German university sample expressed strong anti-Muslim attitudes. Thus, even though higher-educated respondents seem to exhibit less prejudice towards Muslims or other minorities, such prejudice is still common; moreover, socially desirable responding is demonstrably an issue, especially among higher-educated respondents. This made Islamophobia a particularly well-suited subject for our experimental validation of the ECWM in a German university sample.

Method

Participants

The initial sample consisted of 1,629 students from the University of Duesseldorf, Germany. Due to item nonresponse, 98 respondents (6.02% of the initial sample) had to be excluded from further analyses. Dropout rates were significantly higher in the DQ (7.65%) than in the ECWM (5.19%) condition, although this effect was rather small, $\chi^2(1) = 3.91$, $p = .048$, *Cramer's V* = .05. The responses of 181 Muslims who participated in our study were excluded because we wanted to investigate Islamophobia among non-Muslim respondents.

Thus, the final sample consisted of $N = 1{,}361$ respondents (55.69% female). Age was only assessed in broad categories to increase the confidentiality of responses, and was distributed as follows: younger than 20 years (55.55%), 20-29 years (40.71%), 30-39 years (1.91%), 40-49 years (0.66%), 50-59 years (0.37%) and 60 and above (0.81%). Twice as many respondents were assigned to the ECWM condition ($n = 911$; 66.94%) as to the DQ condition ($n = 450$; 33.06%) to compensate for the lower efficiency of indirect questioning techniques (Ulrich, Schröter, Striegel, & Simon, 2012). Within the ECWM group, $n = 455$ respondents were assigned to the ECWM condition with randomization probability $p1 = .158$ and $n = 456$ to the ECWM condition with randomization probability $p2 = .842$. Respondents in the two questioning technique groups (DQ vs. ECWM) did not differ with regard to age group, $\chi^2(5) = 3.78$, $p = .581$, *Cramer's* $V = .05$ or gender, $\chi^2(1) = 0.15$, $p = .695$, *Cramer's* $V = .01$.

Survey Design

Between lectures, respondents filled in a one-page questionnaire consisting of the experimental question and additional questions about their gender, age group, and religious affiliation (Muslim vs. non-Muslim). The experimental question on the presumably sensitive topic of campus Islamophobia was presented in either the DQ format or the ECWM format. The survey also included three additional questions pertaining to the participant's political orientation, their frequency of contact with Muslims, and their perception of Muslims' attitudes towards gender roles. However, these questions did not moderate any of our main findings and are therefore not discussed further. The survey was carried out in accordance with the revised Declaration of Helsinki (World Medical Association, 2013) and the ethical guidelines of the German Society for Psychology (Berufsverband Deutscher Psychologinnen und Psychologen & Deutsche Gesellschaft für Psychologie, 2016). Prior to their participation, all respondents were informed of the purpose of the study and the strict anonymization of all

data, and consented to participate on a voluntary basis without receiving financial compensation.

Sensitive question formats.

DQ. Respondents in the DQ condition were simply presented with the sensitive statement ("Many Muslim students behave in misogynist ways") and had to indicate whether they agreed with this statement or not.

ECWM. In each of the ECWM conditions, the sensitive statement was paired with one of two non-sensitive statements. In the group with randomization probability $p1 = .158$, the non-sensitive statement read: "My father was born in November or December"; and in the group with randomization probability $p2 = .842$, it read: "My father was born between January and October". Respondents were asked to indicate whether they agreed with "*both* of the statements or *none* of them", or whether they agreed with "*exactly one* statement (irrespective of which one)".

Statistical Analyses

To obtain and compare parameter estimates, we established multinomial processing trees (Batchelder, 1998; Batchelder & Riefer, 1999) for both questioning techniques, as detailed in, for example, Moshagen, Hilbig, and Musch (2011); Moshagen, Musch, and Erdfelder (2012); Ostapczuk, Musch, and Moshagen (2011). A graphical representation of the processing trees for the DQ and ECWM conditions is shown in Figure 1. Based on the empirically observed answer frequencies, parameter estimates were obtained using the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977; Hu & Batchelder, 1994) as implemented in the software multiTree (Moshagen, 2010). To compare the parameter estimates, an unrestricted baseline model was compared to a restricted alternative model in which the respective parameters were set to be equal (e.g. $\pi_{ECWM} = \pi_{DQ}$) or set to a certain constant (e.g., 0). Model fit was assessed via the asymptotically Chi²-distributed log-likelihood ratio G^2. Significant differences in model fit indicated that the imposed restriction

was inadmissible and that the respective parameters differed significantly from each other ($\pi_{ECWM} \neq \pi_{DQ}$).

Results

The ECWM fit the empirically observed data well, $G^2(1) = 0.10$, $p = .756$, indicating that respondents did not show any systematic bias towards one of the available answer options. Prevalence estimates could therefore be pooled across the two groups with the two different randomization probabilities. The prevalence estimate of campus Islamophobia was significantly higher in the ECWM (21.19%; $SE = 2.23\%$) than in the DQ (10.89%; $SE = 1.47\%$) condition, $\Delta G^2(1) = 14.69$, $p < .001$, and both estimates were significantly higher than zero, DQ: $\Delta G^2(1) = 1495.46$, $p < .001$; ECWM: $\Delta G^2(1) = 119.40$, $p < .001$.

Discussion

We conducted the first experimental validation of the extended crosswise model (ECWM) by comparing prevalence estimates obtained with the ECWM to corresponding estimates obtained via DQ. Prevalence estimates of campus Islamophobia obtained via the ECWM (21.19%) significantly exceeded estimates obtained via DQ (10.89%). This result demonstrates that Islamophobia is perceived as a sensitive topic and that respondents on a university campus are hesitant to honestly admit Islamophobic attitudes in direct self-reports. Therefore, previous attempts to determine the prevalence of Islamophobia on the basis of direct self-reports likely provided underestimates and lower bounds only (e.g. Kassis et al., 2014; Petersen, 2012; Pickel & Yendell, 2016; Pollack, 2014; Yendell, 2013; Zick et al., 2011). Unlike direct questions, the ECWM seems to successfully control for the influence of socially desirable responding, thereby yielding higher and thus presumably more valid prevalence estimates. This finding is in line with previous positive evaluations of the related crosswise model (Hoffmann & Musch, 2016, 2019; Jann et al., 2012; Korndörfer et al., 2014; Kundt et al., 2017). Extending these findings, the current study presents the first experimental evidence for the validity of the ECWM. A good model fit indicated that the ECWM

prevalence estimates were not distorted by any systematic bias in favor of one of the two answer options. This prerequisite for obtaining valid prevalence estimates has long been implicitly assumed, but could not be explicitly tested within the standard crosswise model (CWM).

Limitations and future directions

A recent criticism of the CWM is that it has been observed to produce false positives under certain conditions (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; Meisters et al., 2019); non-carriers of a sensitive attribute were wrongly categorized as carriers. Moreover, the CWM has also been shown to sometimes produce false negatives by wrongly categorizing some carriers of a sensitive attribute as non-carriers. Since the present study applied only a weak ("more is better"), not a strong validation criterion, it is impossible to tell whether false positives or false negatives may have influenced the current results. We therefore recommend that future studies seek to create conditions under which a strong validation can be conducted. For example, researchers could experimentally induce a sensitive attribute with known prevalence (Hoffmann et al., 2015).

In Germany, two other recent studies found markedly higher shares of respondents directly admitting to being prejudiced against Muslims than the present study. In representative German samples, the share of respondents agreeing to the statement "The Muslim opinion on women contradicts our values" was 76.1% (Zick et al., 2011), and more than 80% of respondents associated Islam with discrimination against women (Petersen, 2012; Pollack, 2014). The relatively low rate of Islamophobic responses in our study may have been caused by two factors. First, we used a different question wording and thus a different operationalization of the sensitive statement. We explicitly asked about prejudice against "Muslim university students", and thus a higher-educated and more progressive subgroup of Muslims against whom prejudice might be less prevalent. Second, unlike previous studies, we employed a student sample comprised of younger, more highly-educated

people rather than a representative sample of the general population at large. This difference in samples might explain our relatively low prevalence estimates because several studies suggest that higher-educated samples are generally less prejudiced (e.g. Easterbrook et al., 2016; Ostapczuk et al., 2009; Wagner & Zick, 1995). However, the ECWM results indicate that prejudice against Muslims was still prevalent in more than 20% of our highly-educated university sample.

Conclusion

In a first experimental evaluation, the recently proposed extended crosswise model (ECWM) led to significantly higher prevalence estimates of campus Islamophobia than a conventional direct question. Moreover, an assessment of model fit indicated no systematic response bias, lending further credence to the estimates obtained with the ECWM. Our findings suggest that the ECWM is a promising new indirect questioning technique, and highlight the importance of controlling for socially desirable responding when surveying Islamophobia or other potentially sensitive attitudes. While direct questioning has been the standard method used for such surveys in previous studies, we strongly recommend the use of indirect questioning techniques in future studies on Islamophobia to better control for the influence of socially desirable responding and obtain more valid prevalence estimates. Out of the available methods, the ECWM seems to be a particularly promising candidate and a potential improvement over competing models because it allows for the detection of systematic response biases without a loss of statistical efficiency.

Open Data

Empirically observed answer frequencies and multiTree equations underlying the findings reported in this manuscript are available at https://osf.io/2fxjt/?view_only=10319e4ac8e042ce89dfab7fb5990c49.

Original instructions for the sensitive questioning formats are available upon request.

References

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10*, 331-344. doi:10.1037/1040-3590.10.4.331

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86. doi:10.3758/Bf03210812

Berufsverband Deutscher Psychologinnen und Psychologen & Deutsche Gesellschaft für Psychologie. (2016). Berufsethische Richtlinien des Berufsverbandes Deutscher Psychologinnen und Psychologen e.V. und der Deutschen Gesellschaft für Psychologie e.V. [Professional ethical guidelines of the Berufsverband Deutscher Psychologinnen und Psychologen e.V. and the Deutsche Gesellschaft für Psychologie e.V.]. Retrieved Sep 10th, 2018 https://www.dgps.de/fileadmin/documents/Empfehlungen/berufsethische_richtlinien_dgps.pdf

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160-168.

Coenders, M., & Scheepers, P. (2003). The Effect of Education on Nationalism and Ethnic Exclusionism: An International Comparison. *Political Psychology, 24*, 313-343. doi:10.1111/0162-895X.00330

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39*, 1-38.

Easterbrook, M. J., Kuppens, T., & Manstead, A. S. R. (2016). The Education Effect: Higher Educational Qualifications are Robustly Associated with Beneficial Personal and Socio-political Outcomes. *Social Indicators Research, 126*, 1261-1298. doi:10.1007/s11205-015-0946-1

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of Forced Responses in a Randomized-Response Model. *Sociological Methods & Research, 11*, 89-100. doi:10.1177/0049124182011001005

Ganter, S. (2001). Zu subtil? Eine empirische Überprüfung neuerer Indikatoren zur Analyse interethnischer Beziehungen [Too subtle? An empirical investigation of new indicators for the analysis of interethnic relations]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 53*, 111–135. doi:10.1007/s11577-001-0006-5

Heck, D. W., Hoffmann, A., & Moshagen, M. (2018). Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior Research Methods, 50*, 1895-1905. doi:10.3758/s13428-017-0957-8

Hoffmann, A., Diedenhofen, B., Verschuere, B. J., & Musch, J. (2015). A strong validation of the Crosswise Model using experimentally induced cheating behavior. *Experimental Psychology, 62*, 403-414. doi:10.1027/1618-3169/a000304

Hoffmann, A., Meisters, J., & Musch, J. (2019a). Nothing but the truth? Effects of faking on the validity of the crosswise model. [Manuscript submitted].

Hoffmann, A., Meisters, J., & Musch, J. (2019b). On the Validity of Non-Randomized Response Techniques: An Experimental Comparison of the Crosswise Model and the Triangular Model. [Manuscript submitted].

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: a Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods, 48*, 1032-1046. doi:10.3758/s13428-015-0628-6

Hoffmann, A., & Musch, J. (2019). Prejudice against Women Leaders: Insights from an Indirect Questioning Approach. *Sex Roles, 80*, 681–692. doi:10.1007/s11199-018-0969-6

Hoffmann, A., Waubert de Puiseau, B., Schmidt, A. F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods, 49*, 1470-1483. doi:10.3758/s13428-016-0804-3

Höglinger, M., & Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis, 25*, 131-137. doi:10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *Plos One, 13*. doi:10.1371/journal.pone.0201770

Hu, X., & Batchelder, W. H. (1994). The Statistical-Analysis of General Processing Tree Models with the Em Algorithm. *Psychometrika, 59*, 21-47. doi:10.1007/Bf02294263

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking Sensitive Questions Using the Crosswise Model. *Public Opinion Quarterly, 76*, 32-49. doi:10.1093/Poq/Nfr036

Kassis, W., Schallié, C., Strube, S., & van der Heyde, J. (2014). Prediction of Anti-Muslim Sentiment on Campus: A Cross-Cultural Analysis of Prejudice in Two University Populations. *HIKMA – Journal of Islamic Theology and Religious Education, 5*, 141-165.

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology, 45*, 18-32. doi:10.1016/j.joep.2014.08.001

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research, 41*, 1387-1403. doi:10.1016/j.ssresearch.2012.05.015

Kundt, T. C., Misch, F., & Nerré, B. (2017). Re-assessing the merits of measuring tax evasion through business surveys: an application of the crosswise model. *International Tax and Public Finance, 24*, 112-133. doi:10.1007/s10797-015-9373-0

Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response - A study of a method for improving the estimate of social security fraud. *Quality & Quantity, 33*, 1-12. doi:10.1023/A:1004361819974

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: thirty-five years of validation. *Sociological Methods & Research, 33*, 319-348. doi:10.1177/0049124104268664

Meertens, R. W., & Pettigrew, T. F. (1997). Is Subtle Prejudice Really Prejudice? *The Public Opinion Quarterly, 61*, 54-71.

Meisters, J., Hoffmann, A., & Musch, J. (2019). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? [Manuscript submitted].

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42-54. doi:10.3758/BRM.42.1.42

Moshagen, M., Hilbig, B. E., & Musch, J. (2011). Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *European Journal of Social Psychology, 41*, 638-644. doi:10.1002/Ejsp.793

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44*, 222-231. doi:10.3758/s13428-011-0144-2 21858604

Nakhaee, M. R., Pakravan, F., & Nakhaee, N. (2013). Prevalence of Use of Anabolic Steroids by Bodybuilders Using Three Methods in a City of Iran. *Addict Health, 5*, 1-6.

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology, 39*, 920-931. doi:10.1002/ejsp.588

Ostapczuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. *Statistical Methods in Medical Research, 20*, 489-503. doi:10.1177/0962280210372843

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes, Vol. 1* (pp. 17-59). San Diego, CA: Academic Press.

Petersen, T. (2012, 21 November 2012). Die Furcht vor dem Morgenland im Abendland [The fear of the orient in the occident]. *Frankfurter Allgemeine Zeitung*. Retrieved from https://www.faz.net/aktuell/politik/inland/allensbach-studie-die-furcht-vor-dem-morgenland-im-abendland-11966471-p2.html

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in western Europe. *European Journal of Social Psychology, 25*, 57-75.

Pickel, G., & Yendell, A. (2016). Islam als Bedrohung? Beschreibung und Erklärung von Einstellungen zum Islam im Ländervergleich [Islam as a threat? Description and explanation of attitudes towards Islam in a cross-country comparison]. *Zeitschrift für vergleichende Politikwissenschaft, 10*, 273–309. doi:10.1007/s12286-016-0309-6

Pollack, D. (2014). Wahrnehmung und Akzeptanz religiöser Vielfalt in ausgewählten Ländern Europas: Erste Beobachtungen [Perception and acceptance of religious diversity in selected European countries: First observations]. In C. Gärtner, M. Koenig, G. Pickel, K. Sammet, & H. Winkel (Eds.), *Grenzen der Toleranz. Veröffentlichungen der Sektion Religionssoziologie der Deutschen Gesellschaft für Soziologie* (pp. 13-34). Wiesbaden: Springer Fachmedien.

Savelkoul, M., Scheepers, P., van der Veld, W., & Hagendoorn, L. (2012). Comparing levels of anti-Muslim attitudes across Western countries. *Quality & Quantity, 46*, 1617-1624. doi:10.1007/s11135-011-9470-9

Stocké, V. (2007). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*, 125-138. doi:10.1027/1614-2241.3.3.125

Strabag, Z., & Listhaug, O. (2008). Anti-Muslim prejudice in Europe: A multilevel analysis of survey data from 30 countries. *Social Science Research, 37*, 268–286. doi:10.1016/j.ssresearch.2007.02.004

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. *Judgment and Decision Making, 11*, 527-536.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883. doi:10.1037/0033-2909.133.5.859 17723033

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking Sensitive Questions: A Statistical Power Analysis of Randomized Response Models. *Psychological Methods, 17*, 623-641. doi:10.1037/A0029314

Wagner, U., & Zick, A. (1995). The relation of formal education to ethnic prejudice: its reliability,validity and explanation. *European Journal of Social Psychology, 25*, 41-56. doi:10.1002/ejsp.2420250105

Warner, S. L. (1965). Randomized-Response - a Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*, 63-69.

Waubert de Puiseau, B., Hoffmann, A., & Musch, J. (2017). How indirect questioning techniques may promote democracy: A pre-election polling experiment. *Basic And Applied Social Psychology, 39*, 209-217. doi:10.1080/01973533.2017.1331351

World Medical Association. (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA, 310*, 2191-2194. doi:10.1001/jama.2013.281053

Yendell, A. (2013). Muslime unerwünscht? Zur Akzeptanz des Islam und dessen Angehörigen. Ein Vergleich zwischen Ost- und Westdeutschland [Muslims not welcome? On the acceptance of Islam and its followers. A comparison between East and West Germany]. In P. G. & H. O. (Eds.), *Religion und Politik im vereinigten Deutschland* (pp. 221-248). Wiesbaden: Springer VS.

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika, 67*, 251-263. doi:10.1007/s00184-007-0131-x

Zick, A., Küpper, B., & Hövermann, A. (2011). *Intolerance, Prejudice and Discrimination: A European Report*. Retrieved from Berlin: Friedrich-Ebert-Stiftung.

Figures

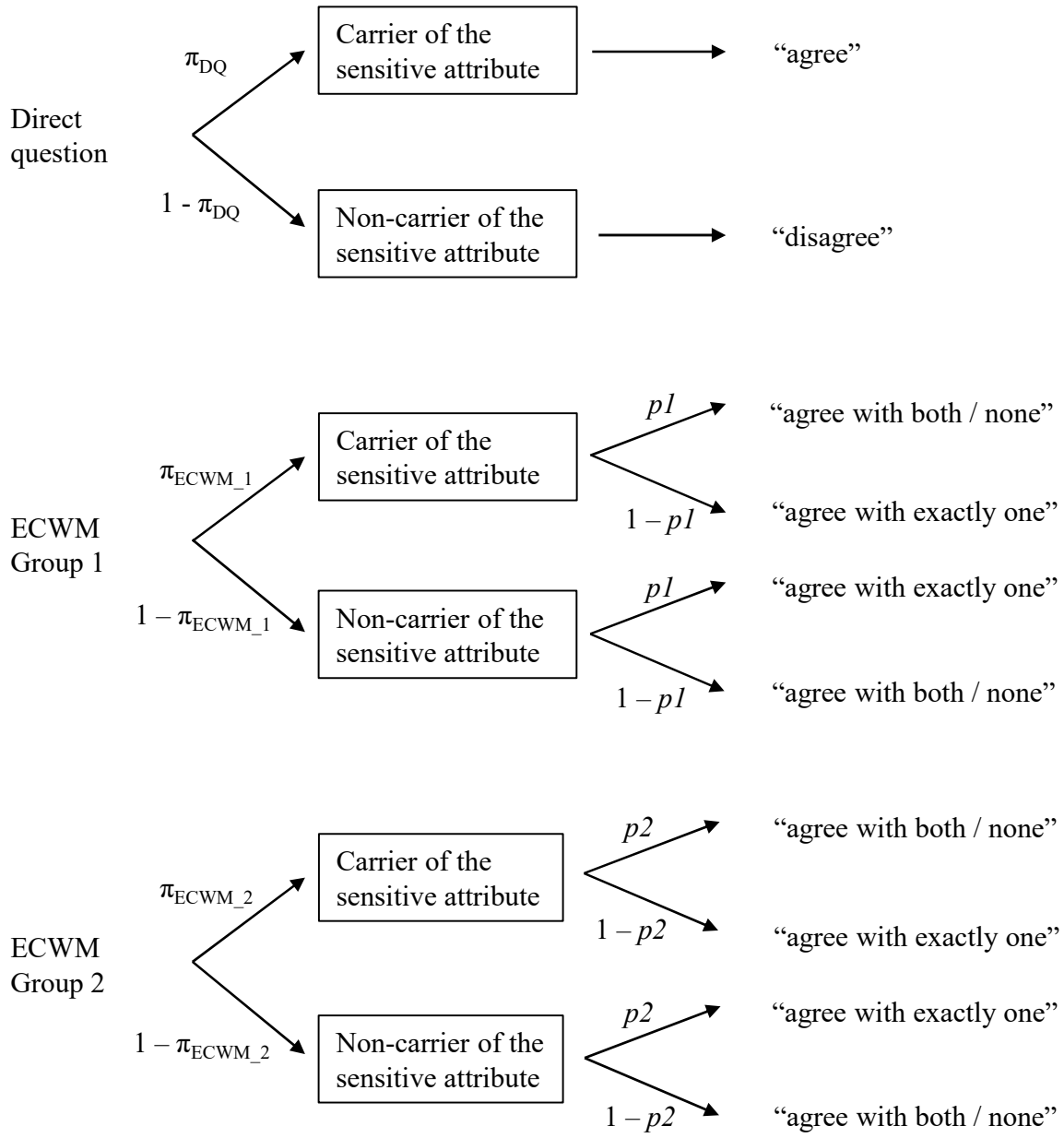


Figure 1. Tree diagrams for direct questioning and the extended crosswise model (ECWM). The parameter π represents the unknown prevalence of the sensitive attribute and the parameter *p1* and *p2* represent the known randomization probabilities, with $p2 = 1 - p1$.

MultiTree equations

MultiTree equations for the estimation of π (Pi) in a multinomial model. Parameter p1 denotes the known probability of being born in November or December (p1 = .158) and parameter p2 denotes the known probability of being born between January and October (p2 = .842, Pötzsch, 2012). ECWM = extended crosswise model, DQ = direct questioning.

```
ECWM_p1     ECWM_p1_bothnone      Pi_ECWM*p1
ECWM_p1     ECWM_p1_one           Pi_ECWM*(1-p1)
ECWM_p1     ECWM_p1_one           (1-Pi_ECWM)*p1
ECWM_p1     ECWM_p1_bothnone      (1-Pi_ECWM)*(1-p1)
ECWM_p2     ECWM_p2_bothnone      Pi_ECWM*p2
ECWM_p2     ECWM_p2_one           Pi_ECWM*(1-p2)
ECWM_p2     ECWM_p2_one           (1-Pi_ECWM)*p2
ECWM_p2     ECWM_p2_bothnone      (1-Pi_ECWM)*(1-p2)
DQ          DQ_agree              Pi_DQ
DQ          DQ_disagree           (1-Pi_DQ)
```

Data

Empirically observed answer frequencies used for parameter estimation in multiTree (Moshagen, 2010). ECWM = extended crosswise model, DQ = direct questioning.

ECWM_p1_bothnone	315
ECWM_p1_one	140
ECWM_p2_bothnone	136
ECWM_p2_one	320
DQ_agree	49
DQ_disagree	401

Can detailed instructions and comprehension checks increase the validity of crosswise model estimates?

Julia Meisters[1¶*], Adrian Hoffmann[1¶], Jochen Musch[1]

[1] Department of Experimental Psychology, University of Duesseldorf, Duesseldorf, Germany.

* Corresponding author

E-mail: julia.meisters@uni-duesseldorf.de (JM)

[¶]These authors contributed equally to this work.

Abstract

The crosswise model is an indirect questioning technique designed to control for socially desirable responding. Although the technique has delivered promising results in terms of improved validity in survey studies of sensitive issues, recent studies have indicated that the crosswise model may sometimes produce false positives. Hence, we investigated whether an insufficient understanding of the crosswise model instructions might be responsible for these false positives and whether ensuring a deeper understanding of the model and surveying more highly educated respondents reduces the problem of false positives. Our results indicate that false positives among highly educated respondents can be reduced when detailed instructions and comprehension checks are employed. Since false positives can also occur in direct questioning, they do not appear to be a specific flaw of the crosswise model, but rather a more general problem of self-reports on sensitive topics. False negatives were found to occur for all questioning techniques, but were less prevalent in the crosswise model than in the direct questioning condition. We highlight the importance of comprehension checks when applying indirect questioning and emphasize the necessity of developing instructions suitable for lower-educated respondents.

Introduction

Direct self-reports on sensitive personal attributes are susceptible to socially desirable responding. Specifically, some respondents may respond in line with social norms, rather than truthfully, leading to an overestimation of the prevalence of socially desirable and an

underestimation of the prevalence of socially undesirable attributes. This threatens the validity of direct self-reports [1-3].

Indirect questioning techniques such as the randomized response technique (RRT [4]) have been proposed to control for social desirability bias. In the original RRT, respondents are presented with two statements: a sensitive statement A (e.g. *I have used cocaine*) and its opposite B (*I have never used cocaine*). Respondents are instructed to employ a randomization procedure, e.g. throwing a die, whose outcome is only known to the respondent, but concealed from the interviewer. Depending on the outcome of this randomization procedure, respondents are asked to respond to either statement A or statement B by indicating whether the respective statement is "true" or "false". Since the interviewer does not know which statement an answer refers to, respondents' privacy is protected. However, the distribution of randomization outcomes is known; therefore, the proportion of respondents carrying the sensitive attribute can be deduced on the sample level.

So-called "weak" validation studies compare prevalence estimates obtained via RRTs with prevalence estimates obtained via a direct question. A meta-analysis of 32 "weak" validation studies [5] found that RRTs generally lead to higher and thus presumably more valid prevalence estimates than direct questioning (DQ). However, the "more-is-better" criterion employed in weak validation studies does not allow definite conclusions to be drawn regarding the validity of RRTs. Rather, definite conclusions result from "strong" validation studies, in which prevalence estimates obtained via RRTs are compared with the ground truth, that is, the known prevalence of a sensitive attribute in a given sample [6]. A meta-analysis of 6 strong validation studies [5] found that RRTs still notably underestimated known prevalences. Moreover, because they add random noise to the estimator, RRTs are generally less efficient than

DQ [7]. Therefore, the application of RRTs is only justified when the topic under investigation is sensitive in nature and an RRT can help to avoid response distortions due to socially desirable responding [5].

The crosswise model: A promising alternative to conventional RRT

Nonrandomized response techniques [8, 9], such as the crosswise model (CWM), represent recent advancements of the RRT. Questions in nonrandomized response format do not require an external randomization device and employ simpler instructions, supposedly making them easier to administer for the experimenter and easier to understand for the respondents. Likely on the basis of these favorable properties, significantly higher and thus presumably more valid prevalence estimates have been obtained via the CWM as compared to DQ for sensitive attributes such as xenophobia [10], plagiarism [11], tax evasion [12, 13], distrust in the Trust Game [13], crossing the street on a red light in plain view of children [14], the use of anabolic steroids among bodybuilders [15], intention to vote for the far-right German party Alternative for Germany [16], and prejudice against female leaders [17]. Moreover, in one strong validation study, the CWM accurately estimated the prevalence of experimentally induced cheating behavior, while DQ led to a severe underestimation [18]. Furthermore, the CWM is easier to understand than other RRT models and is perceived as significantly more confidential than DQ [19].

Cautionary evidence of false positives in the CWM

However, the results of two recent studies by Höglinger and Diekmann [20] and Höglinger and Jann [21] indicate that the CWM may sometimes produce false positives, that is,

some non-carriers of the sensitive attribute are falsely classified as carriers. Höglinger and Diekmann [20] asked respondents whether they had ever received a donated organ and whether they had ever suffered from Chagas disease, both of which are attributes with a prevalence close to zero. As expected, DQ provided estimates that did not significantly differ from zero. In the CWM condition, however, the prevalence estimates for the two zero-prevalence items – and thus false positive rates – were 8% and 5%, respectively. In an additional individual-level validation, the authors asked about a somewhat sensitive control attribute (i.e. whether respondents had completed the German general university entrance qualification). Again, DQ provided a prevalence estimate close to zero; for the CWM, a false positive rate of 7% was observed. Remarkably, the CWM also produced a substantial number of false negatives. As the false positives and false negatives cancelled each other out on the aggregate level, the overall prevalence estimates accurately reflected the known prevalence. However, the interpretability of this individual-level validation is limited because the relevant question was presented as a practice question in the CWM but not in the DQ condition, and because the prevalence estimates were compared with an external criterion that had been collected up to five years earlier and in a different response format. Finally, the authors found that the rate of false positives was moderated by the choice of the unrelated questions used for randomization. This finding implies that researchers using indirect questioning techniques must make a well-informed decision about which unrelated question to use.

In the second study, Höglinger and Jann [21] conducted individual-level validations via an online experiment in which participants had to play one of two dice games: In the *prediction game*, they had to predict the outcome of a die roll in private and were then asked to roll the die. Afterwards, to determine whether they qualified for a payout, respondents were asked to indicate

whether they had rolled the predicted outcome. Since the predictions were made in private, cheating was observable only on the group level; an individual-level validation could only be computed by making two strong assumptions. First, it had to be assumed that all respondents whose predictions were correct actually claimed the payout; second, the false positive rate among respondents whose predictions were correct and who claimed the payout had to be assumed to be equal to the false positive rate among respondents whose predictions were incorrect and who did not claim the payout. In the *roll-a-six game*, participants had to roll a die and were then asked to indicate whether they had rolled a six, in which case they would receive a financial reward. In this second game, the outcomes were tracked, making cheating directly observable on the individual level. After each of the two dice games, participants had to answer a sensitive question about whether they had cheated in the respective game. On the aggregate level, the CWM estimates of cheating were significantly higher than the DQ estimates for both games, thus satisfying the "more-is-better" criterion. However, in both individual-level validations, the CWM produced more than 10% false positives, whereas the false positive rate in the DQ condition did not significantly differ from zero.

At this point, it is not yet understood whether false positives only occur under certain circumstances, or whether they pose a general threat to the validity of the CWM and of indirect questioning techniques as a whole. Höglinger and Diekmann [20] exploratively examined potential causes and correlates of false positives, but did not find a consistent pattern. Respondents who sped through the CWM instructions and may therefore not have understood them properly produced descriptively, but not significantly, more false positives. However, the reverse pattern emerged when only the sensitive questions were examined: here, speeders tended to produce fewer false positives. The authors hypothesized that the problem of false positives

might be less severe in “better designed C[W]M implementations” (p. 5). Consequently, identifying conditions under which respondents show high levels of understanding and trust in the method could help to improve CWM implementation. Trust and understanding are necessary prerequisites for RRTs to yield valid results [22], but are often not achieved [19, 22-27]. Although the comprehensibility of the CWM, operationalized in terms of correct responses to scenario-based questions testing understanding of the model, was shown to be comparatively higher than the comprehensibility of other indirect questioning techniques, more than 16% incorrect responses were still observed [19]. Accordingly, Hoffmann et al. [19] suggested employing detailed instructions and comprehension checks to ensure respondents’ understanding of and trust in indirect questioning techniques. Building upon these recommendations, the present study sought to investigate whether the validity of results obtained via the CWM can be improved by providing respondents with more detailed instructions.

The present study

We sought to obtain a deeper understanding of the conditions under which false positives and false negatives occur in CWM surveys, and how they affect measurement validity. To this end, we conducted a strong validation based on a known external criterion by employing the anagram paradigm introduced by Hoffmann et al. [18]. This paradigm induces cheating to generate a sensitive attribute with known prevalence in the sample. It allowed us to compare all prevalence estimates with a known true value, and to conduct separate analyses of false negatives among carriers and false positives among non-carriers of the sensitive attribute. Based on the results of Höglinger and Diekmann [20] and Höglinger and Jann [21], we hypothesized that prevalence estimates based on self-reports would suffer from both false positives and false

159 negatives. Moreover, we expected false positives to occur more frequently in the CWM
160 condition compared to the DQ condition [cf. 20, 21]. In contrast, we expected false negatives to
161 occur more frequently in the DQ condition compared to the CWM condition due to the influence
162 of socially desirable responding [cf. 10, 11, 12].

163 Most importantly, the current study sought to identify potential means of reducing false
164 positives and false negatives in order to maximize the validity of prevalence estimates obtained
165 via indirect questioning techniques such as the CWM. We therefore tested the assumption that an
166 insufficient understanding of and trust in the method are major causes of false positives in the
167 CWM. To this end, we experimentally manipulated the amount of information respondents
168 received in the CWM instructions. Specifically, we compared two groups, one of which received
169 detailed instructions combined with several questions assessing comprehension (CWM detailed),
170 and the other of which received only brief instructions and no comprehension questions (CWM
171 brief). We expected that false positives were less likely when respondents had a better
172 understanding of the CWM (CWM detailed) than when they had only a superficial understanding
173 of the method (CWM brief).

174 Since comprehension of CWM instructions has been shown to be positively associated
175 with education [19], and lower-educated respondents have been found to disobey RRT
176 instructions more often [28], we additionally compared the false positive and false negative rates
177 between a higher-educated (at least 12 years of education, the German *Abitur*) and a lower-
178 educated subgroup (at most 10 years of education, the German *Realschule*). We expected a
179 higher false positive rate among lower educated than among highly educated respondents.
180

Methods

Participants

Respondents were recruited by a commercial German online panel provider. To avoid a lack of understanding of the instructions due to language difficulties, a necessary prerequisite for participation was that respondents were German native speakers. Moreover, to avoid confounding education with age, we restricted the age range of respondents to 30 to 40 years. This homogeneity with respect to age helped maximize the statistical power for testing our main hypotheses because it reduced the variance in education that would have been present in a more age-diverse sample due to a general trend towards higher educational attainment among younger cohorts in Germany [29].

The survey was carried out in accordance with the revised Declaration of Helsinki [30] and the ethical guidelines of the German Society for Psychology [31]. In Germany, there is no binding obligation that research projects can only be carried out after approval by an ethics committee. Participation in the present study could not have any negative consequences for the respondents, and anonymity was ensured at all times. The respondents participated voluntarily and after informed consent was obtained. There was no risk that participation could cause any physical or mental damage or discomfort to participants beyond their normal everyday experiences. Therefore, ethics committee approval was not required according to the "Ethical Research Principles and Test Methods in the Social and Economic Sciences" formulated by the Ethics Research Working Group of the German Data Forum [32] and the "Ethical Recommendations of the German Psychological Society" [33].

The initial sample consisted of 3060 respondents, with an equal distribution regarding education (higher-educated: at least 12 years of education, the German *Abitur*; lower-educated:

at most 10 years of education, the German *Realschule*) and gender (male vs female). Due to incomplete data, 347 respondents had to be excluded from the analysis (11.34% of the initial sample). This dropout was nonselective in terms of cheating on the anagram task, $\chi^2(1, N = 2934) = 2.75$, $p = .098$, *Cramer's V* = .03. Dropout rates were slightly lower among higher-educated respondents (8.25%) compared with lower-educated respondents (13.13%). $\chi^2(1, N = 3040) = 18.87$, $p < .001$, *Cramer's V* = .08. However, this effect was small and thus considered negligible. Respondents in the CWM detailed condition were more likely to drop out (19.20%) than respondents in the other conditions (CWM brief: 3.00%; DQ: 3.67%), $\chi^2(2, N = 3002) = 211.75$, $p < .001$, *Cramer's V* = .27.

The final sample consisted of 2713 respondents (50.31% female) with a mean age of $M = 34.73$ years ($SD = 3.15$). Half of the respondents (49.98%) were lower-educated, while the other half were higher-educated (50.02%). Overall, 972 respondents (35.83%) were assigned to the CWM detailed condition, 1164 (42.90%) to the CWM brief condition, and 577 (21.27%) to the DQ condition. Respondents in the three conditions did not differ with regard to education, $\chi^2(2) = 0.92$, $p = .632$, *Cramer's V* = .02.

Measures

Anagram Cheating Task

To enable a strong validation, we experimentally induced a sensitive attribute with known prevalence in the sample using the anagram paradigm established by Hoffmann et al. [18]. This paradigm consists of two parts: the anagram task itself and a subsequent opportunity for respondents to overreport their performance – that is, to cheat on the task. In the first part of the anagram task, respondents are presented with three scrambled words ("anagrams"). Instead of

227 directly reporting the solutions to these anagrams, respondents are instructed to solve the
228 anagrams in their head. The anagrams are presented for a maximum of 20 seconds each;
229 respondents can continue to the next anagram anytime by pressing a button. Unknown to the
230 respondents, the first two anagrams are very easy to solve (solved by > 99% of the respondents
231 in a pilot study [18]), while the third anagram is virtually impossible to solve (solved by ca. 1%
232 [18]). In the second part of the anagram task, respondents are presented with the solutions and
233 are given the opportunity to participate in a lottery for 100€, 50€ and 30€ under the condition
234 that they were able to solve all three anagrams. Respondents are asked whether they were able to
235 solve all three anagrams in time. The two available answer options are: "No, I solved fewer than
236 three anagrams" and "Yes, I solved all three anagrams (opportunity to participate in the lottery at
237 the end of the survey)". These answer options are explicitly designed to motivate respondents to
238 overreport their performance. Due to the indirect query of the number of solved anagrams,
239 respondents should feel safe that they will not be exposed as cheaters. However, because solving
240 all three anagrams is virtually impossible, all respondents claiming to have found all solutions
241 are categorized as cheaters.
242

243 ## Sensitive question

244 The sensitive question read: "On the anagram task, I claimed that I had solved more
245 anagrams than I had actually solved". It was asked in either the CWM detailed, CWM short or
246 DQ format (between-subjects). In the DQ format, respondents simply had to indicate whether the
247 sensitive question was "true" or "false". In the CWM format, respondents had to answer two
248 statements simultaneously: the aforementioned sensitive statement and a non-sensitive statement
249 with known prevalence p: "I was born in November or December" (p = .158 according to official

250 birth statistics [34]). The answer options read: “Both statements are true or both statements are
251 false” versus “exactly one statement is true (irrespective of which one)”. Respondents in the
252 CWM brief condition received brief instructions on how to answer the question, and were
253 informed that the response format would protect their privacy as their birth month would remain
254 unknown to the researchers. In addition to the instructions provided in the CWM brief condition,
255 respondents in the CWM detailed condition were further informed that the researchers would use
256 the relative probability of being born in November or December to compute the share of people
257 who agreed to the sensitive statement on the sample level, but that their individual answers
258 would remain confidential. Moreover, respondents were provided with four comprehension
259 questions capturing whether they had understood how to answer the CWM question. As an
260 example, the first comprehension question was as follows:

261 “Assuming you were born in February, and assuming you had *not* exaggerated on your
262 report of the number of solved anagrams. Which answer would you have to give?”

263 The answer options read: “I would have to answer ‘both statements are true or both statements
264 are false’” and “I would have to answer ‘exactly one statement is true (irrespective of which
265 one)’”, and were presented in randomized order. The four comprehension questions covered all
266 four combinations of respondents potentially holding or not holding the sensitive attribute
267 (exaggerating their report of the number of solved anagrams) as well as the non-sensitive
268 attribute used for randomization (being born in November or December). On the subsequent
269 page, respondents received feedback on their responses. If the respondents failed to provide
270 correct responses to any of the comprehension questions, the detailed instructions and those
271 comprehension questions that were not answered correctly were repeated up to two times. The
272 presentation ended when respondents had provided correct responses to all four comprehension

questions, or when they failed to provide a correct response to at least one of the questions three times. Subsequently, respondents were presented with two additional questions capturing whether they had understood how the CWM protected their privacy. The first of these questions read:

> "Imagine you had chosen the option 'Both statements are true or both statements are false'. What could someone who does not know your birth month infer from your choice?"

The answer options read: "He could infer that you had exaggerated your report of the number of solved anagrams", "He could infer that you reported the number of solved anagrams truthfully", and "He could infer neither of the above", presented in fixed order. The second question covered the case in which respondents had hypothetically chosen the second answer option with respect to the sensitive statement ("exactly one of the statements is true (irrespective of which one)"). Again, respondents received feedback on their answers, and the detailed instructions and questions were repeatedly presented up to two times if not solved correctly. After the comprehension questions, the respondents were presented with the actual sensitive question. While the detailed instructions on the questioning technique were still available, respondents were explicitly told that they should now choose the response that aligns with their own personal behavior, and that their birth month would remain unknown to the researchers.

Self-reported response behavior and perception of the questioning technique

Respondents were asked to evaluate the sensitive question by indicating how strongly they agreed with several statements. These statements read: "The question was comprehensible" (perceived comprehensibility), "The question guaranteed the confidentiality of my response"

(perceived confidentiality), “I clearly knew which answer to pick” (perceived clarity), and “I just ticked anything” (random response; this variable was then reverse-coded, with higher values indicating less random responding). All statements were rated on a 7-point Likert-type scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

Procedure

Respondents filled in an online questionnaire that began with a short introduction, followed by demographic questions asking about their gender, age, native language and highest school-leaving qualification. They were then given the instructions for the anagram task and had the opportunity to practice the task with two example anagrams. Next, respondents were informed that the actual task would now start and that if they could solve all three anagrams, they could take part in a lottery for 100€, 50€ and 30€. After the anagram task, they were given the opportunity to cheat on reporting the number of solved anagrams as described above. Subsequently, they were queried with regard to their cheating behavior in the anagram task in either the DQ, the CWM detailed, or the CWM brief format (between-subjects). After the sensitive question, the respondents were asked to evaluate the questioning technique, were debriefed and were then given the opportunity to participate in the lottery. In order to avoid discriminating against honest respondents, all respondents were given the opportunity to participate in the lottery regardless of whether they had answered honestly or dishonestly.

Statistical Analyses

For parameter estimation and comparison, we formulated multinomial processing tree (MPT) models [35, 36] following the procedure outlined in previous studies [37-39]. The

parameter π represents the prevalence of the sensitive attribute (cheating on the anagram task) and the parameter p represents the known prevalence of the non-sensitive attribute used for randomization (birth month, $p = .158$ according to official birth statistics [34]). Maximum likelihood estimates were obtained using the expectation maximization algorithm [40, 41] implemented in the software multiTree [42], version 0.46. Parameter comparisons and restrictions were assessed via differences in the asymptotically χ^2-distributed log-likelihood statistic G^2 between an unrestricted baseline model and a restricted alternative model (e.g. $\pi_{CWM\_detailed} = \pi_{DQ}$ or $\pi_{CWM\_detailed} = .00$).

To more thoroughly investigate the validity of the obtained estimates, we transferred the approach of analyzing false positives and false negatives detailed in Höglinger and Jann [21] to the multinomial framework. To this end, we first split the sample into two parts: respondents who claimed to have solved all three anagrams in the anagram task were categorized as cheaters, while respondents who reported having solved fewer than three anagrams were considered honest respondents. This categorization is justified by the fact that solving all three anagrams has been shown to be virtually impossible in a previous study [18]. We then formulated separate multinomial processing trees for cheaters and honest respondents, and within these subsamples, for the DQ and CWM conditions. Hence, the false positive rate was estimated as the proportion of carriers of the sensitive attribute (π) within the sub-sample of honest respondents for the respective questioning technique, and the false negative rate was estimated as the proportion of non-carriers ($1- \pi$) in the sub-sample of cheaters.

Results

Parameter Estimates, False Positives and False Negatives

Our analyses revealed significantly higher prevalence estimates in both CWM conditions (detailed: $\hat{\pi} = 25.48\%$, $SE = 2.21\%$; brief: $\hat{\pi} = 30.78\%$, $SE = 2.07\%$) as compared to the DQ condition ($\hat{\pi} = 11.79\%$, $SE = 1.34\%$); CWM detailed vs. DQ: $\Delta\hat{\pi} = 13.69\%$, $\Delta G^2(1) = 27.94$, $p < .001$; CWM brief vs. DQ: $\Delta\hat{\pi} = 18.99\%$, $\Delta G^2(1) = 56.74$, $p < .001$; CWM detailed vs. brief: $\Delta\hat{\pi} = 5.30\%$, $\Delta G^2(1) = 3.06$, $p = .080$. The known prevalence of the sensitive attribute (DQ: 58.93%, CWM brief: 56.70%, CWM detailed: 59.26%) did not differ across conditions, $\chi^2(2) = 1.63$, $p = .442$, *Cramer's V* = .03, and was underestimated by all questioning techniques; CWM detailed vs. known prevalence: $\Delta\hat{\pi} = 33.78\%$, $\Delta G^2(1) = 210.75$, $p < .001$; CWM brief vs. known prevalence: $\Delta\hat{\pi} = 25.92\%$, $\Delta G^2(1) = 147.47$, $p < .001$; DQ vs. known prevalence: $\Delta\hat{\pi} = 47.14\%$, $\Delta G^2(1) = 558.69$, $p < .001$. Thus, the CWM met a weak ("more is better"), but not a strong validation criterion, as it still substantially underestimated the known prevalence. Moreover, we detected substantial rates of false positives in all experimental groups (see Table 1 and S1 Appendix), with the highest rates in both CWM conditions (CWM detailed: 13.08%, CWM brief: 14.32%, DQ: 2.53%; this corresponds to a specificity of CWM detailed: 86.92%, CWM brief: 85.68%, DQ: 97.47%). Comparatively high rates of false negatives were observed in all conditions (see Table 1 and S2 Appendix). While the highest rate was found in the DQ condition (81.77%; sensitivity: 18.23%), the false negative was also substantial in both CWM conditions (detailed: 65.99%, sensitivity: 34.01%; brief: 56.65%; sensitivity: 43.35%).

Table 1. False positives and false negatives in the total sample and split by randomness of responses, perceived comprehensibility, perceived confidentiality and perceived clarity of the questioning technique (standard errors in parentheses).

	False Positives (in %)			False Negatives (in %)		
	DQ	CWM brief	CWM detailed	DQ	CWM brief	CWM detailed
Total sample	2.53	14.32	13.08	81.77	56.65	65.99
(*N* = 2713)	(1.02)	(2.84)	(3.17)	(2.09)	(2.83)	(2.97)
Randomness of responses						
non-random	1.40	13.69	6.14	84.97	57.76	70.94
(*N* = 2194)	(0.80)	(2.98)	(3.35)	(2.11)	(3.14)	(3.51)
random	13.64	19.90	36.34	64.82	51.78	54.93
(*N* = 519)	(7.32)	(9.33)	(7.53)	(6.50)	(6.59)	(5.47)
Perceived comprehensibility						
comprehensible	1.10	15.89	8.83	85.20	58.17	72.07
(*N* = 1244)	(0.77)	(3.94)	(6.48)	(2.25)	(3.89)	(6.77)
incomprehensible	7.27	12.51	14.28	72.22	54.94	64.62
(*N* = 1469)	(3.50)	(4.10)	(3.63)	(4.72)	(4.14)	(3.30)
Perceived confidentiality						
confidential	1.42	15.97	6.36	87.03	55.70	70.77
(*N* = 1163)	(1.00)	(4.25)	(5.07)	(2.47)	(4.24)	(5.28)
not confidential	4.17	12.91	16.52	75.48	57.41	63.89
(*N* = 1550)	(2.04)	(3.82)	(4.02)	(3.46)	(3.81)	(3.59)
Perceived clarity						
clear	1.14	14.35	7.40	85.06	53.96	67.48
(*N* = 1551)	(0.80)	(3.33)	(5.04)	(2.21)	(3.54)	(5.23)
unclear	6.45	14.25	16.15	70.89	61.51	65.29
(*N* = 1162)	(3.12)	(5.45)	(4.04)	(5.11)	(4.71)	(3.61)

DQ = direct questioning, CWM brief = crosswise model with brief instructions, CWM detailed = crosswise model with detailed instructions and comprehension questions.

Effects of education

A split by level of education (high vs. low) revealed that false positives were particularly prevalent among lower-educated respondents (see Fig 1). In both CWM conditions, false positive rates were significantly lower for higher-educated respondents than for lower-educated

respondents, CWM brief: $\Delta\hat{\pi} = 11.91\%$, $\Delta G^2(1) = 4.33$, $p = .038$; CWM detailed: $\Delta\hat{\pi} = 18.16\%$, $\Delta G^2(1) = 8.13$, $p = .004$. In the DQ condition, this tendency was not significant, $\Delta\hat{\pi} = 3.53\%$, $\Delta G^2(1) = 3.23$, $p = .072$. False negative rates (see Fig 1) were not affected by level of education, DQ: $\Delta\hat{\pi} = 2.36\%$, $\Delta G^2(1) = 0.32$, $p < .574$; CWM brief: $\Delta\hat{\pi} = 6.98\%$, $\Delta G^2(1) = 1.51$, $p = .220$; CWM detailed: $\Delta\hat{\pi} = 3.90\%$, $\Delta G^2(1) = 0.43$, $p = .511$.

Fig 1. False positives and false negatives as a function of level of education and condition. DQ = direct questioning, CWM brief = crosswise model with brief instructions, CWM detailed = crosswise model with detailed instructions and comprehension questions.

Selection by comprehension questions

To more thoroughly evaluate respondents' objective comprehension of the CWM instructions, we analyzed the rates of correct responses to the comprehension questions in the CWM detailed condition. Comprehension Questions 1 to 6 were passed by a total of 68.83% of respondents; 12.45% provided correct responses to all comprehension questions in the first attempt, 56.38% in the second or third attempt. Higher-educated respondents were more likely to provide correct answers to all comprehension questions in the first attempt (18.71%) as well as in the second or third attempt (59.76%) than lower-educated respondents (first attempt: 5.89%, second or third attempt: 52.84%), $\chi^2(2) = 64.46$, $p < .001$, *Cramer's V* = .26.

To determine whether comprehension questions can be used to improve overall data quality, we repeated the analyses of false positives and false negatives in the CWM detailed condition including only those respondents who were eventually able to correctly answer all comprehension questions (hereinafter referred to as *respondents with high understanding*,

394 N = 669, 68.8% of respondents in the CWM detailed condition). For higher-educated
395 respondents, the false positive rate dropped from 4.78% (SE = 3.92%) when including all
396 respondents in the CWM detailed condition to 0.00% (SE = 4.35%) in the subgroup of
397 respondents with high understanding. For lower-educated respondents, however, the false
398 positive rate slightly increased from 22.94% (SE = 5.05%) when including all respondents in the
399 CWM detailed condition to 25.17% (SE = 6.68%) when considering only the subgroup of
400 respondents with high understanding. Moreover, among the subgroup of respondents with high
401 understanding, the false positive rate was significantly lower for higher-educated respondents
402 compared to lower-educated respondents, $\Delta\hat{\pi} = 25.17\%$, $\Delta G^2(1) = 14.15$, $p < .001$. In both
403 educational groups, the false negative rate was higher in the subsample of respondents with high
404 understanding compared to an analysis without sample constraints (higher education: all
405 respondents in the CWM detailed condition: 64.00%, SE = 4.27%, respondents with high
406 understanding: 70.79%, SE = 4.75%; low education: all respondents in the CWM detailed
407 condition: 67.90%, SE = 4.13%, respondents with high understanding: 73.24%, SE = 5.27%).
408 Within the subgroup of respondents in the CWM detailed condition with high understanding,
409 false negative rates did not differ with regard to education, $\Delta\hat{\pi} = 2.36\%$, $\Delta G^2(1) = 0.12$, $p = .730$.
410 Overall, these results suggest that false positives can be effectively reduced by
411 implementing detailed instructions and comprehension questions, but only among higher-
412 educated samples. Moreover, this comes at the expense of an increase in false negatives.
413

Effects of self-reported response behavior and perception of the questioning technique

Table 2 reports descriptive statistics for self-ratings of randomness of responses, perceived comprehensibility, perceived confidentiality, and perceived clarity of the questioning techniques. All of these variables were significantly intercorrelated (see Table 3); a Cronbach's alpha of .70 indicated that they measured a homogeneous construct. ANOVAs and Bonferroni-corrected post-hoc tests indicated that the CWM detailed condition was evaluated as less understandable, less confidential and less clear than the CWM brief condition, which in turn was evaluated as worse than the DQ condition on all of these variables. Moreover, respondents in the CWM detailed condition indicated significantly more random responses than respondents in the CWM brief or DQ conditions (see Table 2).

Table 2. Descriptive statistics and results of ANOVAs for self-reported response behaviors and perceptions of the questioning technique split by condition.

	CWM detailed *M (SD)*	CWM brief *M (SD)*	DQ *M (SD)*	F (2,2703)	*p*	η_p
Randomness of responses	6.29* (1.41)	6.54 (1.30)	6.60 (1.23)	13.44	< .001	.01
Perceived comprehensibility	4.47* (1.87)	6.01* (1.43)	6.53* (1.04)	409.99	< .001	.23
Perceived confidentiality	5.27* (1.64)	5.70* (1.58)	6.07* (1.39)	49.69	< .001	.04
Perceived clarity	4.55* (1.75)	5.82* (1.48)	6.12* (1.33)	249.92	< .001	.16

All variables were assessed on a 7-point Likert-type scale with higher values indicating more favorable evaluations. Randomness of responses was originally reverse-coded, but was inverted to facilitate the interpretability of means.
* Bonferroni-corrected post-hoc tests revealed that these conditions significantly differed from all other conditions (all $p < .001$).

Table 3. Correlations between self-reported response behaviors and perceptions of the questioning technique.

	Randomness of responses	Perceived comprehensibility	Perceived confidentiality	Perceived clarity
Randomness of responses	-	.21*	.15*	.17*
Perceived comprehensibility		-	.50*	.64*
Perceived confidentiality			-	.42*
Perceived clarity				-

All variables were assessed on a 7-point Likert-type scale with higher values indicating more favorable evaluations. Randomness of responses was originally reverse-coded, but was inverted to facilitate the interpretability of means.
* $p < .001$

Spearman rank correlations revealed that respondents who performed better on the comprehension questions (1 = 'failed at least one comprehension question in the third attempt', 2 = 'comprehension questions solved in second or third attempt', 3 = 'comprehension questions solved in first attempt'), indicated lower rates of random responses ($r_s = -.31$, $p < .001$, $n = 972$) as well as higher perceived comprehensibility ($r_s = .30$, $p < .001$, $n = 972$), confidentiality ($r_s = .21$, $p < .001$, $n = 972$) and subjective clarity of the questioning technique ($r_s = .35$, $p < .001$, $n = 972$).

To determine whether respondents' self-assessment of the randomness of their responses was associated with the validity of the results obtained, we identified respondents who had indicated that they *strongly disagreed* with the statement "I simply ticked anything" (80.9% of the sample). These respondents were classified as having provided "non-random responses", while all other respondents were considered to have provided "random responses". A split by this moderator variable revealed that false positive rates were substantially lower among respondents

who indicated having provided non-random responses; this pattern was observed in both the CWM detailed and the DQ conditions, but not in the CWM brief condition. However, this decrease in false positive rates came at the expense of an increase in false negative rates in both the CWM detailed and the DQ condition. Similar results were observed for median splits of perceived comprehensibility, perceived confidentiality and perceived clarity of the questioning technique: Higher values on these variables were associated with reduced false positives, but also increased false negatives in the CWM detailed and the DQ conditions. However, these tendencies were only significant in the DQ condition, and only for splits with reference to perceived comprehensibility and perceived clarity. For detailed statistics on these analyses, see Table 1 and Appendices A and B.

Completion times

A Kruskal-Wallis test showed that completion times for the experimental section of the questionnaire differed significantly across the three experimental conditions, $\chi^2(2) = 2232.53$, $p < .001$. Dunn-Bonferroni corrected post-hoc tests revealed that the detailed CWM instructions were associated with higher completion times (median: 380 seconds) than the brief CWM instructions (median: 43 seconds), and the brief instructions with higher completion times than the DQ instructions (median: 9 seconds), DQ vs. CWM brief: $z = 20.14$, $p < .001$; DQ vs CWM detailed: $z = 45.61$, $p < .001$; CWM brief vs. CWM detailed: $z = 31.56$, $p < .001$.

Discussion

In the present study, we investigated an apparent contradiction in the scientific literature regarding the validity of the crosswise model (CWM [8]), an indirect questioning technique designed to control for socially desirable responding. While numerous studies suggest that prevalence estimates obtained via the CWM are highly valid [11-13, 17, 18], recent work by Höglinger and Diekmann [20] and Höglinger and Jann [21] suggests that the model tends to produce false positives in certain situations. Building upon these findings, we sought to identify conditions under which false positives occur in applications of the CWM and investigated what measures can be taken to effectively reduce the false positive rate to a minimum. The core idea was that false positives might be caused by an insufficient understanding of the instructions. To test this idea, we conducted a strong validation and compared the validity of estimates obtained via conventional direct questions (DQ) with the validity of estimates obtained via the CWM in two groups, one of which received only brief instructions on how to answer the sensitive question (CWM brief), and the other of which received more detailed information on the procedure and had to answer several comprehension questions (CWM detailed).

Overall, the CWM led to significantly higher prevalence estimates than DQ, thus meeting the “more is better” criterion on the aggregate level. However, both DQ and the CWM severely underestimated the known prevalence of the sensitive attribute, thus failing a strong validation. Moreover, in line with our hypotheses, we found higher rates of false positives for both CWM conditions as compared to the DQ condition. In contrast, false negatives were significantly more common in the DQ condition as compared to both CWM conditions. The hypotheses that false positives occur less frequently in CWM applications when respondents have a deep understanding of the method (detailed CWM) compared to a superficial understanding (CWM

brief), and that false positives occur more frequently in lower-educated than in higher-educated respondents, were only partially confirmed. As expected, detailed instructions combined with comprehension questions led to lower rates of false positives, but only within the subgroup of higher-educated respondents. However, neither detailed instructions and comprehension questions nor higher education completely eliminated false positives in the CWM at the individual level.

The results of our study generally support the findings of Höglinger and Diekmann [20] and Höglinger and Jann [21] that the CWM in its original form tends to produce false positives. However, in contrast to previous studies that did not experimentally investigate potential moderators of the false positive rate [20, 21], the present study showed that satisfactorily low rates of false positives can be achieved by the use of extensive instructions in combination with comprehension questions. However, this was only true for the sub-sample of higher-educated respondents and among participants who indicated that they did not provide random answers and who perceived the questions as easily comprehensible and as protecting their confidentiality. False positives were completely eliminated (0.0%) among the higher-educated respondents who passed all comprehension questions. The positive association between education level and CWM performance is consistent with the results of a recent study showing that higher-educated respondents are better at understanding CWM instructions [19]. However, in the present study, the beneficial effect of comprehension checks on the validity of prevalence estimates came at the expense of higher dropout rates and higher completion times.

Interestingly, in the current study, substantial rates of false positives were also observed in the DQ condition. This observation is striking given that DQ does not include complex instructions, but only requires respondents to make a rather simple decision of agreeing or

519 disagreeing with a statement. Hence, this finding seems to indicate that the issue of false
520 positives is not a specific drawback of indirect questioning techniques such as the CWM, but
521 extends to situations in which prevalence estimates are based on self-reports of any kind. In line
522 with this, Hoffmann et al. [19] found that the rate of incorrect answers in a DQ condition was
523 about 10%. In another study by Bishop et al. [43], a substantial number of respondents took a
524 clear stance on a purely fictional issue, which impressively illustrates that self-reports must be
525 interpreted cautiously. Such response patterns may be due to careless responding, straightlining,
526 or non-serious participation, which are common phenomena in self-reports and have been shown
527 to impair data quality [44-48]. These concepts are closely related to the 'randomness' of
528 responses in the present study. The finding that the false positive rate was lower among
529 respondents who indicated having provided non-random responses than among respondents who
530 indicated having provided random responses lends further support to the assumption that false
531 positives are a product of careless responding and non-serious participation.

532 One point that has received little attention in the recently published literature on indirect
533 questioning techniques is the problem of false negatives in surveys on sensitive topics. False
534 negatives refer to the proportion of carriers of a sensitive attribute that are incorrectly categorized
535 as being non-carriers. While false positives can lead to an undesired overestimation, false
536 negatives carry the risk of underestimating the prevalence of sensitive attributes. It was precisely
537 to avoid this problem that indirect questioning techniques were introduced in the first place. In
538 our study, significantly lower rates of false negatives were observed in both CWM conditions
539 compared to the DQ condition. This finding provides clear evidence of an advantage of CWM
540 questions over conventional direct questions, namely a higher proportion of correctly identified
541 carriers of the sensitive attribute. Remarkably, for the CWM, the rate of false negatives was

considerably higher than the rate of false positives. Moreover, the rates of false positives and false negatives were interdependent: A reduction in false positives (e.g. by selecting only those respondents who passed the comprehension checks) led to an increase in false negatives, presumably due to the application of a more conservative criterion.

Overall, with regard to the prevalence estimates obtained, the deflating influence of false negatives clearly outweighed the inflating influence of false positives. This led to a severe underestimation of the known prevalence of the sensitive attribute in all conditions. In light of this, prevalence estimates for sensitive personal attributes obtained in previous studies using the CWM (e.g. xenophobia [10]; prejudice against female leaders [17]; plagiarism [11]) or other indirect questioning techniques (e.g. doping [49]) were most likely underestimations rather than overestimates of the population values.

In summary, our results are in line with the findings of two meta-analyses on RRT studies [5]: Prevalence estimates for sensitive attributes obtained via indirect questioning techniques such as the CWM demonstrably underestimate the true value due to substantial rates of false negatives; nevertheless, CWM estimates are clearly superior to estimates obtained via a conventional DQ approach as they more closely reflect the ground truth.

Limitations and future research directions

While the current study will hopefully contribute to a better understanding of the conditions under which false positives and false negatives occur in applications of the CWM, some questions cannot be answered on the basis of our data and should therefore be addressed in future research.

First, it would be interesting to gain a deeper understanding of the cognitive processes involved in the formation of false positives and false negatives in the CWM. As corroborated by our data, false positives are most likely a product of inadvertent instruction non-adherence. It seems rather unlikely that non-carriers try to make themselves appear to be carriers of the sensitive attribute by deliberately choosing an answer that does not correspond to their actual status. False negatives, however, could be a mixture of carriers inadvertently providing a false response due to instruction miscomprehension, and carriers deliberately choosing the answer that minimizes the probability of them being identified as a carrier. While a particular advantage of the CWM is that none of the answer options clearly excludes the possibility of carrying the sensitive attribute, one of the answer options is still associated with a lower risk of being identified as a carrier, depending on the randomization probability p. Given our data, we cannot answer whether and to what extent carriers pursued this strategy. Hence, future research should address this question via methods such as personal interviews and open-ended questions about how respondents arrived at their specific answers.

Second, our data cannot uncover the processes responsible for the large share of inaccurate responses provided by lower-educated respondents. While false positives could be reduced among higher-educated respondents when detailed instructions and comprehension questions were included, the false positive rate among lower-educated respondents was not affected by such measures. Future research projects should therefore continue to optimize conditions until both higher- and lower-educated respondents are willing and able to provide accurate responses. This is of particular importance when the attribute under investigation is moderated by respondents' level of education (e.g., negative attitudes towards foreigners [28]), because differential comprehension levels might lead to erroneous conclusions.

Third, the present study highlights that respondents' thorough comprehension of indirect questions is a necessary prerequisite for obtaining valid results. For this reason, the exact implementation of the questioning technique seems crucial. However, implementation details are often unknown due to insufficient documentation, and a considerable amount of research focuses exclusively on the development of new statistical models and ignores questions of feasibility and implementation. We therefore recommend that future research focus more on the procedural implementation and comprehensibility of indirect questioning techniques. In addition, we encourage researchers to contribute to the improvement of tools that capture respondents' understanding of indirect questions, such as the comprehension checks employed in the present study. It would be desirable to design these measures in a way that ensures a thorough comprehension of indirect questioning techniques even among lower-educated samples.

Fourth, it remains unclear why the CWM was perceived as less confidential overall than DQ. This finding contrasts with the objective confidentiality guaranteed by the CWM and also contradicts a recent study in which the CWM's subjective privacy protection was rated significantly higher than the protection provided by DQ [19]. Possible reasons might include the perceived high complexity of the CWM instructions as well as the between-subjects design of the current study, which could have prevented the respondents from establishing common reference frames [cf. 50]. Moreover, it is unclear why the CWM detailed format, which was supposed to enhance understanding, was perceived as less comprehensible than the CWM brief format, and why respondents were less sure of what to do in the CWM detailed than in the CWM brief condition. It seems likely that the comprehension questions in the CWM detailed condition raised respondents' awareness of the complexity of the CWM instructions, leading them to subjectively perceive the questioning format as rather complicated, whereas respondents in the

CWM brief condition received no feedback on their understanding of the instructions and thus might not have realized that they did not understand the procedure properly. Once again, the between-subjects design of the current study may have also prevented respondents from establishing common reference frames [cf. 50].

Finally, the harmful influence of false negatives was substantially more pronounced in the present study than the influence of false positives. Overall, this led to an underestimation of the known prevalence. Future research should thus also try to identify conditions under which false negatives can be avoided. To this end, we recommend that studies employing a strong validation approach, comprehensive instructions and comprehension checks also be conducted for other indirect questioning techniques (e.g. cheating detection models). Such studies should ideally compare the validity of different models across sensitive attributes with varying prevalence in order to explore the potential influence of the population value on the validity of the prevalence estimates obtained.

Practical implications

In light of the current results, the CWM can be recommended for application if–and only if–the sample under investigation is highly educated, and detailed instructions and comprehension questions are used. As the present results also show, however, the desirable positive effect of detailed instructions and comprehension questions on the validity of the prevalence estimates obtained comes at the expense of higher dropout rates and higher completion times. Moreover, our results call into question the application of the CWM in its current format among lower-educated samples. In order to obtain valid answers among lower-educated respondents, more research seems needed to find ways of improving such respondents'

633 instruction comprehension. The present results also highlight the importance of including
634 measures of instruction comprehension as well as specific aspects of respondents' subjective
635 experience (such as perceived confidentiality and randomness of responses) in surveys of
636 sensitive personal attributes. Moreover, the present study underscores the importance of strong
637 validations, since only individual-level data allow for the detection of false positives and false
638 negatives, and thus for a comprehensive assessment of a method's validity [6, 21].
639

640 # Conclusion

641 The present study confirmed the assumption that the CWM tends to produce false
642 positives. It also showed that the problem of false positives is not specific to indirect questioning
643 techniques, but rather seems to be a drawback of self-reports of any kind, including conventional
644 DQ. On the aggregate level, there were many more false negatives than false positives, resulting
645 in severe underestimations of the prevalence of the sensitive attribute across all questioning
646 techniques. However, taking both false positives and false negatives into account, the CWM
647 clearly outperformed DQ in terms of aggregate validity. Our findings therefore further suggest
648 that CWM estimates in previous studies were more likely to be underestimates rather than
649 overestimates of the true prevalence of sensitive attributes.

References

1. Paulhus DL. Measurement and Control of Response Bias. In: Robinson JP, Shaver PR, Wrightsman LS, editors. Measures of personality and social psychological attitudes, Vol 1. San Diego, CA: Academic Press; 1991. p. 17-59.
2. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull. 2007;133:859-83. doi: 10.1037/0033-2909.133.5.859 17723033. PubMed PMID: 2007-12463-007.
3. Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. Qual Quant. 2013;47:2025-47. doi: 10.1007/s11135-011-9640-9. PubMed PMID: ISI:000316267500014.
4. Warner SL. Randomized-Response - a Survey Technique for Eliminating Evasive Answer Bias. J Am Stat Assoc. 1965;60:63-9. PubMed PMID: ISI:A1965CKX1300005.
5. Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM, Maas CJM. Meta-analysis of randomized response research: thirty-five years of validation. Sociol Method Res. 2005;33:319-48. doi: 10.1177/0049124104268664. PubMed PMID: ISI:000226871800001.
6. Umesh UN, Peterson RA. A Critical Evaluation of the Randomized-Response Method - Applications, Validation, and Research Agenda. Sociol Method Res. 1991;20:104-38. doi: 10.1177/0049124191020001004. PubMed PMID: ISI:A1991GA53200004.
7. Ulrich R, Schröter H, Striegel H, Simon P. Asking Sensitive Questions: A Statistical Power Analysis of Randomized Response Models. Psychol Methods. 2012;17:623-41. doi: 10.1037/A0029314. PubMed PMID: ISI:000312113600010.
8. Yu J-W, Tian G-L, Tang M-L. Two new models for survey sampling with sensitive characteristic: design and analysis. Metrika. 2008;67:251-63. doi: 10.1007/s00184-007-0131-x. PubMed PMID: ISI:000254204100001.
9. Tian G-L, Tang M-L. Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2014.
10. Hoffmann A, Musch J. Assessing the validity of two indirect questioning techniques: a Stochastic Lie Detector versus the Crosswise Model. Behav Res Methods. 2016;48:1032-46. doi: 10.3758/s13428-015-0628-6. PubMed PMID: WOS:000382653900017.
11. Jann B, Jerke J, Krumpal I. Asking Sensitive Questions Using the Crosswise Model. Public Opin Q. 2012;76:32-49. doi: 10.1093/Poq/Nfr036. PubMed PMID: ISI:000301068300002.
12. Kundt TC, Misch F, Nerré B. Re-assessing the merits of measuring tax evasion through business surveys: an application of the crosswise model. Int Tax Public Finan. 2017;24:112-33. doi: 10.1007/s10797-015-9373-0.
13. Korndörfer M, Krumpal I, Schmukle SC. Measuring and explaining tax evasion: Improving self-reports using the crosswise model. J Econ Psychol. 2014;45:18-32. doi: 10.1016/j.joep.2014.08.001.
14. Hoffmann A, Meisters J, Musch J. Nothing but the truth? Effects of faking on the validity of the crosswise model. 2019.
15. Nakhaee MR, Pakravan F, Nakhaee N. Prevalence of Use of Anabolic Steroids by Bodybuilders Using Three Methods in a City of Iran. Addict Health. 2013;5:1-6.

16. Waubert de Puiseau B, Hoffmann A, Musch J. How indirect questioning techniques may promote democracy: A pre-election polling experiment. Basic And Applied Social Psychology. 2017;39:209-17. doi: 10.1080/01973533.2017.1331351.
17. Hoffmann A, Musch J. Prejudice against Women Leaders: Insights from an Indirect Questioning Approach. Sex Roles. 2019;80:681–92. doi: 10.1007/s11199-018-0969-6.
18. Hoffmann A, Diedenhofen B, Verschuere BJ, Musch J. A strong validation of the Crosswise Model using experimentally induced cheating behavior. Exp Psychol. 2015;62:403-14. doi: 10.1027/1618-3169/a000304.
19. Hoffmann A, Waubert de Puiseau B, Schmidt AF, Musch J. On the comprehensibility and perceived privacy protection of indirect questioning techniques. Behav Res Methods. 2017;49:1470-83. doi: 10.3758/s13428-016-0804-3.
20. Höglinger M, Diekmann A. Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. Polit Anal. 2017;25:131-7. doi: 10.1017/pan.2016.5. PubMed PMID: WOS:000398071200008.
21. Höglinger M, Jann B. More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. PLoS One. 2018;13. doi: 10.1371/journal.pone.0201770.
22. Landsheer JA, van der Heijden PGM, van Gils G. Trust and understanding, two psychological aspects of randomized response - A study of a method for improving the estimate of social security fraud. Qual Quant. 1999;33:1-12. doi: 10.1023/A:1004361819974. PubMed PMID: ISI:000079006700001.
23. Coutts E, Jann B. Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). Sociol Method Res. 2011;40:169-93. doi: 10.1177/0049124110390768. PubMed PMID: ISI:000286103900008.
24. Edgell SE, Himmelfarb S, Duchan KL. Validity of Forced Responses in a Randomized-Response Model. Sociol Method Res. 1982;11:89-100. doi: 10.1177/0049124182011001005. PubMed PMID: ISI:A1982PF08700005.
25. I-Cheng C, Chow LP, Rider RV. Randomized Response Technique as Used in Taiwan Outcome of Pregnancy Study. Stud Family Plann. 1972;3:265-9. PubMed PMID: ISI:A1972N976300002.
26. Hejri SM, Zendehdel K, Asghari F, Fotouhi A, Rashidian A. Academic disintegrity among medical students: a randomised response technique study. Med Educ. 2013;47:144-53. doi: 10.1111/Medu.12085. PubMed PMID: ISI:000313752400006.
27. van der Heijden PGM, van Gils G, Bouts J, Hox JJ. A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of social security fraud. Kwantitatieve Methoden. 1998;19:15-34.
28. Ostapczuk M, Musch J, Moshagen M. A randomized-response investigation of the education effect in attitudes towards foreigners. Eur J Soc Psychol. 2009;39:920-31. doi: 10.1002/ejsp.588. PubMed PMID: 2009-17720-004.
29. German Federal Statistical Office. Bildungsstand der Bevölkerung - Ergebnisse des Mikrozensus 2017 [Educational status of the population - Results of the microcensus 2017]. 2018.
30. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310:2191-4. doi: 10.1001/jama.2013.281053. PubMed PMID: 24141714.

31. Berufsethische Richtlinien des Berufsverbandes Deutscher Psychologinnen und Psychologen e.V. und der Deutschen Gesellschaft für Psychologie e.V. [Professional ethical guidelines of the Berufsverband Deutscher Psychologinnen und Psychologen e.V. and the Deutsche Gesellschaft für Psychologie e.V.] [Internet]. 2016 [cited Sep 10th, 2018]. Available from: https://www.dgps.de/fileadmin/documents/Empfehlungen/berufsethische_richtlinien_dgps.pdf.
32. RatSWD. Forschungsethische Grundsätze und Prüfverfahren in den Sozial- und Wirtschaftswissenschaften [Ethical research principles and test methods in the social and economic sciences]. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD) 2017.
33. DGPs. In welchen Fällen auf einen Ethikantrag verzichtet werden kann [In which cases an ethics application is not needed]. Ethisches Handeln in der psychologischen Forschung - Empfehlungen der Deutschen Gesellschaft für Psychologie für Forschende und Ethikkommissionen. Göttingen: Hogrefe; 2018
34. Geburten in Deutschland [Births in Germany] [Internet]. German Federal Statistical Office. 2012 [cited Jun 6, 2012]. Available from: https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/Bevoelkerungsbewegung/BroschuereGeburtenDeutschland0120007129004.pdf.
35. Batchelder WH. Multinomial processing tree models and psychological assessment. Psychol Assessment. 1998;10:331-44. doi: 10.1037/1040-3590.10.4.331. PubMed PMID: ISI:000077959900003.
36. Batchelder WH, Riefer DM. Theoretical and empirical review of multinomial process tree modeling. Psychon B Rev. 1999;6:57-86. doi: 10.3758/Bf03210812. PubMed PMID: ISI:000079575700004.
37. Moshagen M, Hilbig BE, Musch J. Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. Eur J Soc Psychol. 2011;41:638-44. doi: 10.1002/Ejsp.793. PubMed PMID: ISI:000293687800012.
38. Moshagen M, Musch J, Erdfelder E. A stochastic lie detector. Behav Res Methods. 2012;44:222-31. doi: 10.3758/s13428-011-0144-2 21858604. PubMed PMID: 2012-04194-018.
39. Ostapczuk M, Musch J, Moshagen M. Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. Stat Methods Med Res. 2011;20:489-503. doi: 10.1177/0962280210372843. PubMed PMID: ISI:000296245700003.
40. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via Em Algorithm. J R Stat Soc Series B Stat Methodol. 1977;39:1-38. PubMed PMID: ISI:A1977DM46400001.
41. Hu X, Batchelder WH. The Statistical-Analysis of General Processing Tree Models with the Em Algorithm. Psychometrika. 1994;59:21-47. doi: 10.1007/Bf02294263. PubMed PMID: ISI:A1994NA79500002.
42. Moshagen M. multiTree: A computer program for the analysis of multinomial processing tree models. Behav Res Methods. 2010;42:42-54. doi: 10.3758/BRM.42.1.42.
43. Bishop GF, Oldendick RW, Tuchfarber AJ. Experiments in filtering political opinions. Political Behavior. 1980;2:339-69.

44. Aust F, Diedenhofen B, Ullrich S, Musch J. Seriousness checks are useful to improve data validity in online research. Behav Res Methods. 2013;45:527-35. doi: 10.3758/s13428-012-0265-2.
45. Meade AW, Craig SB. Identifying Careless Responses in Survey Data. Psychol Methods. 2012;17:437-55. doi: 10.1037/a0028085. PubMed PMID: WOS:000308679400010.
46. Maniaci MR, Rogge RD. Caring about carelessness: Participant inattention and its effects on research. J Res Pers. 2014;48:61-83. doi: 10.1016/j.jrp.2013.09.008. PubMed PMID: WOS:000331023900006.
47. Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: Detecting satisficing to increase statistical power. J Exp Soc Psychol. 2009;45:867-72. doi: 10.1016/j.jesp.2009.03.009. PubMed PMID: WOS:000269278800029.
48. Woods CM. Careless responding to reverse-worded items: Implications for confirmatory factor analysis. J Psychopathol Behav. 2006;28:189-94. doi: 10.1007/s10862-005-9004-7. PubMed PMID: WOS:000240054800007.
49. Ulrich R, Pope HG, Jr., Cleret L, Petroczi A, Nepusz T, Schaffer J, et al. Doping in Two Elite Athletics Competitions Assessed by Randomized-Response Surveys. Sports Med. 2018;48:211-9. Epub 2017/08/30. doi: 10.1007/s40279-017-0765-4. PubMed PMID: 28849386.
50. Birnbaum MH. How to show that 9 > 221: Collect judgments in a between-subjects design. Psychol Methods. 1999;4:243-9. doi: 10.1037/1082-989x.4.3.243. PubMed PMID: ISI:000082696900001.

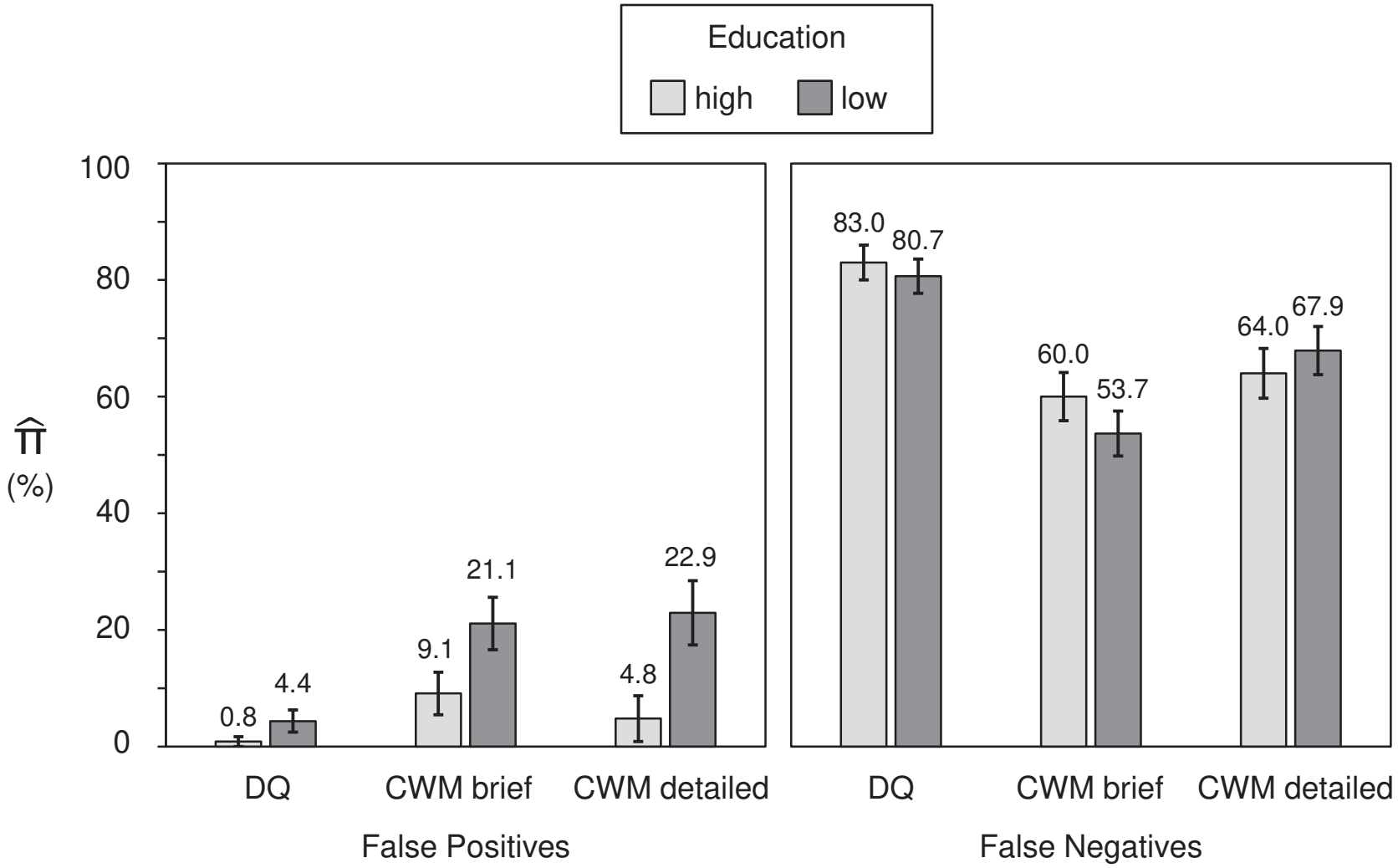
Supporting Information

S1 Appendix. Parameter comparisons of false positives.

S2 Appendix. Parameter comparisons of false negatives.

S1 MultiTree Equations. MultiTree equations for the estimation of π, false positives and false negatives in a multinomial model.

S1 Data. Empirically observed answer frequencies for the attributes used for parameter estimation in multiTree.


Education
high
low
100
80
60
40
20
0
$\hat{\pi}$
(%)
0.8
4.4
9.1
21.1
4.8
22.9
83.0
80.7
60.0
53.7
64.0
67.9
DQ
CWM brief
CWM detailed
DQ
CWM brief
CWM detailed
False Positives
False Negatives

Supporting Information File: Appendix

Parameter comparisons of false positives for the total sample and split by randomness of responses, perceived comprehensibility, perceived confidentiality and perceived clarity of the questioning technique

Parameter Comparisons: False Positives						
					Model fit	
		Parameter 1 (in %)	Parameter 2 (in %)	|*Difference*| (in %)	ΔG^2 ($df = 1$)	p
Total Sample						
	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	13.08	14.32	1.24	0.08	= .771
	$FP_{CWM\ detailed} = FP_{DQ}$	13.08	2.53	10.55	10.82	= .001*
	$FP_{CWM\ brief} = FP_{DQ}$	14.32	2.53	11.79	15.86	< .001*
	$FP_{CWM\ detailed} = 0\%$	13.08	0.00	13.08	20.99	< .001*
	$FP_{CWM\ brief} = 0\%$	14.32	0.00	14.32	31.70	< .001*
	$FP_{DQ} = 0\%$	2.53	0.00	2.53	165.09	< .001*
Randomness of responses						
Non-random	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	6.14	13.69	7.55	2.77	= .096
	$FP_{CWM\ detailed} = FP_{DQ}$	6.14	1.40	4.74	2.03	= .154
	$FP_{CWM\ brief} = FP_{DQ}$	13.69	1.40	12.29	17.17	< .001*
	$FP_{CWM\ detailed} = 0\%$	6.14	0.00	6.14	3.79	= .052
	$FP_{CWM\ brief} = 0\%$	13.69	0.00	13.69	26.18	< .001*
	$FP_{DQ} = 0\%$	1.40	0.00	1.40	78.93	< .001*
Random	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	36.34	19.90	16.44	1.81	= .178
	$FP_{CWM\ detailed} = FP_{DQ}$	36.34	13.64	22.70	3.87	= .049*
	$FP_{CWM\ brief} = FP_{DQ}$	19.90	13.64	6.26	0.27	= .601
	$FP_{CWM\ detailed} = 0\%$	36.34	0.00	36.34	32.16	< .001*
	$FP_{CWM\ brief} = 0\%$	19.90	0.00	19.90	5.95	= .015*
	$FP_{DQ} = 0\%$	13.64	0.00	13.64	93.00	< .001*

CWM detailed	$FP_{non\text{-}random} = FP_{random}$	6.14	36.34	30.2	14.96	< .001*
CWM brief	$FP_{non\text{-}random} = FP_{random}$	13.69	19.90	6.21	0.42	= .516
DQ	$FP_{non\text{-}random} = FP_{random}$	1.40	13.64	12.24	6.85	= .009*
Perceived comprehensibility						
Comprehensible	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	8.83	16.03	7.20	0.83	= .363
	$FP_{CWM\ detailed} = FP_{DQ}$	8.83	1.10	7.73	1.70	= .192
	$FP_{CWM\ brief} = FP_{DQ}$	15.89	1.10	14.79	15.79	< .001*
	$FP_{CWM\ detailed} = 0\%$	8.83	0.00	8.83	2.18	= .140
	$FP_{CWM\ brief} = 0\%$	15.89	0.00	15.89	20.65	< .001*
	$FP_{DQ} = 0\%$	1.10	0.00	1.10	51.66	< .001*
Incomprehensible	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	14.28	12.51	1.77	0.10	= .748
	$FP_{CWM\ detailed} = FP_{DQ}$	14.28	7.27	7.01	1.61	= .205
	$FP_{CWM\ brief} = FP_{DQ}$	12.51	7.27	5.24	0.88	= .347
	$FP_{CWM\ detailed} = 0\%$	14.28	0.00	14.28	19.33	< .001*
	$FP_{CWM\ brief} = 0\%$	12.51	0.00	12.51	11.40	< .001*
	$FP_{DQ} = 0\%$	7.27	0.00	7.27	118.70	< .001*
CWM detailed	$FP_{comprehensible} = FP_{incomprehensible}$	8.83	14.28	5.45	0.52	= .472
CWM brief	$FP_{comprehensible} = FP_{incomprehensible}$	15.89	12.51	3.38	0.35	= .553
DQ	$FP_{comprehensible} = FP_{incomprehensible}$	1.10	7.27	6.17	5.27	= .022*
Perceived confidentiality						
Confidential	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	6.36	15.97	9.61	2.03	= .154
	$FP_{CWM\ detailed} = FP_{DQ}$	6.36	1.42	4.94	0.99	= .319
	$FP_{CWM\ brief} = FP_{DQ}$	15.97	1.42	14.55	12.55	< .001*
	$FP_{CWM\ detailed} = 0\%$	6.36	0.00	6.36	1.78	= .182
	$FP_{CWM\ brief} = 0\%$	15.97	0.00	15.97	17.92	< .001*
	$FP_{DQ} = 0\%$	1.42	0.00	1.42	52.69	< .001*
Not confidential	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	16.52	12.91	3.61	0.42	= .515
	$FP_{CWM\ detailed} = FP_{DQ}$	16.52	4.17	12.35	7.20	= .007*
	$FP_{CWM\ brief} = FP_{DQ}$	12.91	4.17	8.74	3.96	= .046*
	$FP_{CWM\ detailed} = 0\%$	16.52	0.00	16.52	21.56	< .001*
	$FP_{CWM\ brief} = 0\%$	12.91	0.00	12.91	14.07	< .001*

	$FP_{DQ} = 0\%$	4.17	0.00	4.17	114.11	< .001*
CWM detailed	$FP_{confidential} = FP_{not\ confidential}$	6.36	16.52	10.16	2.35	= .125
CWM brief	$FP_{confidential} = FP_{not\ confidential}$	15.97	12.91	3.06	0.29	= .592
DQ	$FP_{confidential} = FP_{not\ confidential}$	1.42	4.17	2.75	1.71	= .191
Perceived clarity						
Clear	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	7.40	14.35	6.95	1.26	= .261
	$FP_{CWM\ detailed} = FP_{DQ}$	7.40	1.14	6.26	1.67	= .196
	$FP_{CWM\ brief} = FP_{DQ}$	14.35	1.14	13.21	16.26	< .001*
	$FP_{CWM\ detailed} = 0\%$	7.40	0.00	7.40	2.48	= .115
	$FP_{CWM\ brief} = 0\%$	14.35	0.00	14.35	23.16	< .001*
	$FP_{CWM\ brief} = 0\%$	1.14	0.00	1.14	51.82	< .001*
Unclear	$FP_{CWM\ detailed} = FP_{CWM\ brief}$	16.15	14.25	1.90	0.08	= .780
	$FP_{CWM\ detailed} = FP_{DQ}$	16.15	6.45	9.70	3.21	= .073
	$FP_{CWM\ brief} = FP_{DQ}$	14.25	6.45	7.80	1.53	= .215
	$FP_{CWM\ detailed} = 0\%$	16.15	0.00	16.15	20.28	< .001*
	$FP_{CWM\ brief} = 0\%$	14.25	0.00	14.25	8.54	= .003*
	$FP_{DQ} = 0\%$	6.45	0.00	6.45	117.70	< .001
CWM detailed	$FP_{clear} = FP_{unclear}$	7.40	16.15	8.75	1.77	= .184
CWM brief	$FP_{clear} = FP_{unclear}$	14.35	14.25	0.10	0.00	= .988
DQ	$FP_{clear} = FP_{unclear}$	1.14	6.45	5.31	4.44	= .035*

Note. FP = False positives.
* $p < .05$

Supporting Information File: Appendix B

Parameter comparisons of false negatives for the total sample and split by randomness of responses, perceived comprehensibility, perceived confidentiality and perceived clarity of the questioning technique

Parameter Comparisons: False Negatives						
					Model fit	
		Parameter 1 (in %)	Parameter 2 (in %)	|*Difference*| (in %)	ΔG^2 ($df = 1$)	p
Total Sample						
	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	65.99	56.65	9.34	5.15	= .023*
	$FN_{CWM\ detailed} = FN_{DQ}$	65.99	81.77	15.78	18.47	< .001*
	$FN_{CWM\ brief} = FN_{DQ}$	56.65	81.77	25.12	47.71	< .001*
	$FN_{CWM\ detailed} = 0\%$	65.99	0.00	65.99	601.97	< .001*
	$FN_{CWM\ brief} = 0\%$	56.65	0.00	56.65	522.21	< .001*
	$FN_{DQ} = 0\%$	81.77	0.00	81.77	9918.94	< .001*
Randomness of responses						
Non-random	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	70.94	57.76	13.18	7.73	= .005*
	$FN_{CWM\ detailed} = FN_{DQ}$	70.94	84.97	14.03	11.79	< .001*
	$FN_{CWM\ brief} = FN_{DQ}$	57.76	84.97	27.21	48.32	< .001*
	$FN_{CWM\ detailed} = 0\%$	70.94	0.00	70.94	93.447	< .001*
	$FN_{CWM\ brief} = 0\%$	57.76	0.00	57.76	440.19	< .001*
	$FN_{DQ} = 0\%$	84.97	0.00	84.97	8710.31	< .001*
Random	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	54.93	51.78	3.15	0.13	= .713
	$FN_{CWM\ detailed} = FN_{DQ}$	54.93	64.82	9.89	1.33	= .250
	$FN_{CWM\ brief} = FN_{DQ}$	51.78	64.82	13.04	1.94	= .163
	$FN_{CWM\ detailed} = 0\%$	54.93	0.00	54.93	133.18	< .001*
	$FN_{CWM\ brief} = 0\%$	51.78	0.00	51.78	82.69	< .001*
	$FN_{DQ} = 0\%$	64.82	0.00	64.82	1219.40	< .001*

CWM detailed	$FN_{non\text{-}random} = FN_{random}$	70.94	54.93	16.01	6.14	= .013*
CWM brief	$FN_{non\text{-}random} = FN_{random}$	57.76	51.78	5.98	0.67	= .412
DQ	$FN_{non\text{-}random} = FN_{random}$	84.97	64.82	20.15	10.78	= .001*
Perceived comprehensibility						
Comprehensible	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	72.07	58.17	13.90	3.06	= .080
	$FN_{CWM\ detailed} = FN_{DQ}$	72.07	85.20	13.13	3.65	= .056
	$FN_{CWM\ brief} = FN_{DQ}$	58.17	85.20	27.03	35.44	< .001*
	$FN_{CWM\ detailed} = 0\%$	72.07	0.00	72.07	130.22	< .001*
	$FN_{CWM\ brief} = 0\%$	58.17	0.00	58.17	289.79	< .001*
	$FN_{DQ} = 0\%$	85.20	0.00	85.20	7637.60	< .001*
Incomprehensible	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	64.62	54.94	9.68	3.35	= .067
	$FN_{CWM\ detailed} = FN_{DQ}$	64.62	72.22	7.60	1.67	= .197
	$FN_{CWM\ brief} = FN_{DQ}$	54.94	72.22	17.28	7.12	= .008*
	$FN_{CWM\ detailed} = 0\%$	64.62	0.00	64.62	472.70	< .001*
	$FN_{CWM\ brief} = 0\%$	54.94	0.00	54.94	232.75	< .001*
	$FN_{DQ} = 0\%$	72.22	0.00	72.22	2288.34	< .001*
CWM detailed	$FN_{comprehensible} = FN_{incomprehensible}$	72.07	64.62	7.45	0.95	= .329
CWM brief	$FN_{comprehensible} = FN_{incomprehensible}$	58.17	54.94	3.23	0.32	= .569
DQ	$FN_{comprehensible} = FN_{incomprehensible}$	85.20	72.22	12.98	7.00	= .008*
Perceived confidentiality						
Confidential	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	70.77	55.70	15.07	4.84	= .028*
	$FN_{CWM\ detailed} = FN_{DQ}$	70.77	87.03	16.26	8.14	= .004*
	$FN_{CWM\ brief} = FN_{DQ}$	55.70	87.03	31.33	38.50	< .001*
	$FN_{CWM\ detailed} = 0\%$	70.77	0.00	70.77	209.09	< .001*
	$FN_{CWM\ brief} = 0\%$	55.70	0.00	55.70	226.38	< .001*
	$FN_{DQ} = 0\%$	87.03	0.00	87.03	5788.69	< .001*
Not confidential	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	63.89	57.41	6.48	1.53	= .216
	$FN_{CWM\ detailed} = FN_{DQ}$	63.89	75.48	11.59	5.23	= .022*
	$FN_{CWM\ brief} = FN_{DQ}$	57.41	75.48	18.07	11.77	< .001*
	$FN_{CWM\ detailed} = 0\%$	63.89	0.00	63.89	394.03	< .001*
	$FN_{CWM\ brief} = 0\%$	57.41	0.00	57.41	295.92	< .001*
	$FN_{DQ} = 0\%$	75.48	0.00	75.48	4137.78	< .001*

CWM detailed	$FN_{confidential} = FN_{not\ confidential}$	70.77	63.89	6.88	1.14	= .285
CWM brief	$FN_{confidential} = FN_{not\ confidential}$	55.70	57.41	1.71	0.09	= .764
DQ	$FN_{confidential} = FN_{not\ confidential}$	87.03	75.48	11.55	7.53	= .006*
Perceived clarity						
Clear	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	67.48	53.96	13.52	4.48	= .034*
	$FN_{CWM\ detailed} = FN_{DQ}$	67.48	85.06	17.58	10.16	= .001*
	$FN_{CWM\ brief} = FN_{DQ}$	53.96	85.06	31.10	52.26	< .001*
	$FN_{CWM\ detailed} = 0\%$	67.48	0.00	67.48	200.32	< .001*
	$FN_{CWM\ brief} = 0\%$	53.96	0.00	53.96	307.84	< .001*
	$FN_{DQ} = 0\%$	85.06	0.00	85.06	7958.65	< .001*
Unclear	$FN_{CWM\ detailed} = FN_{CWM\ brief}$	65.29	61.51	3.78	0.41	= .523
	$FN_{CWM\ detailed} = FN_{DQ}$	65.29	70.89	5.60	0.78	= .378
	$FN_{CWM\ brief} = FN_{DQ}$	61.51	70.89	9.38	1.77	= .183
	$FN_{CWM\ detailed} = 0\%$	65.29	0.00	65.29	401.77	< .001*
	$FN_{CWM\ brief} = 0\%$	61.51	0.00	61.51	216.01	< .001*
	$FN_{DQ} = 0\%$	70.89	0.00	70.89	1967.82	< .001*
CWM detailed	$FN_{clear} = FN_{unclear}$	67.48	65.29	2.19	0.12	= .731
CWM brief	$FN_{clear} = FN_{unclear}$	53.96	61.51	7.55	1.63	= .201
DQ	$FN_{clear} = FN_{unclear}$	85.06	70.89	14.17	7.53	= .006*

Note. * $p < .05$

Supporting Information File: Multitree Equations

MultiTree equations for the estimation of π (Pi), false positives (FP) and false negatives (FN) in a multinomial model. Parameter p1 denotes the known probability of being born in November or December (p1 = .158) and parameter p2 denotes the known probability of being born between January and October (p2 = .842, Pötzsch, 2012). CWM_detailed = crosswise model with detailed instructions and comprehension questions, CWM_brief = crosswise model with brief instructions, DQ = direct questioning.

Estimation of π in the total sample:

```
CWM_detailed      CWM_detailed_both_true_both_false      1_Pi_CWM_detailed*p
CWM_detailed      CWM_detailed_one_true                  1_Pi_CWM_detailed*(1-p)
CWM_detailed      CWM_detailed_one_true                  (1-1_Pi_CWM_detailed)*p
CWM_detailed      CWM_detailed_both_true_both_false      (1-1_Pi_CWM_detailed)*(1-p)
CWM_brief         CWM_brief_both_true_both_false         2_Pi_CWM_brief*p
CWM_brief         CWM_brief_one_true                     2_Pi_CWM_brief*(1-p)
CWM_brief         CWM_brief_one_true                     (1-2_Pi_CWM_brief)*p
CWM_brief         CWM_brief_both_true_both_false         (1-2_Pi_CWM_brief)*(1-p)
DQ                DQ_true                                3_Pi_DQ
DQ                DQ_false                               (1-3_Pi_DQ)
```

Estimation of false positives among subsample of non-carriers of the sensitive attribute:

```
CWM_detailed      CWM_detailed_both_true_both_false      1_FP_CWM_detailed*p
CWM_detailed      CWM_detailed_one_true                  1_FP_CWM_detailed*(1-p)
CWM_detailed      CWM_detailed_one_true                  (1-1_FP_CWM_detailed)*p
CWM_detailed      CWM_detailed_both_true_both_false      (1-1_FP_CWM_detailed)*(1-p)
CWM_brief         CWM_brief_both_true_both_false         2_FP_CWM_brief*p
CWM_brief         CWM_brief_one_true                     2_FP_CWM_brief*(1-p)
CWM_brief         CWM_brief_one_true                     (1-2_FP_CWM_brief)*p
CWM_brief         CWM_brief_both_true_both_false         (1-2_FP_CWM_brief)*(1-p)
DQ                DQ_true                                3_FP_DQ
DQ                DQ_false                               (1-3_FP_DQ)
```

Estimation of false negatives among subsample of carriers of the sensitive attribute:

CWM_detailed	CWM_detailed_both_true_both_false	(1-1_FN_CWM_detailed)*p
CWM_detailed	CWM_detailed_one_true	(1-1_FN_CWM_detailed)*(1-p)
CWM_detailed	CWM_detailed_one_true	1_FN_CWM_detailed*p
CWM_detailed	CWM_detailed_both_true_both_false	1_FN_CWM_detailed*(1-p)
CWM_brief	CWM_brief_both_true_both_false	(1-2_FN_CWM_brief)*p
CWM_brief	CWM_brief_one_true	(1-2_FN_CWM_brief)*(1-p)
CWM_brief	CWM_brief_one_true	2_FN_CWM_brief*p
CWM_brief	CWM_brief_both_true_both_false	2_FN_CWM_brief*(1-p)
DQ	DQ_true	(1-3_FN_DQ)
DQ	DQ_false	3_FN_DQ

Supporting Information File: Data

Empirically observed answer frequencies for the attributes used for parameter estimation in multiTree (Moshagen, 2010) version 0.46. CWM_detailed = crosswise model with detailed instructions and comprehension questions, CWM_brief = crosswise model with brief instructions, DQ = direct questioning.

Total Sample

CWM_detailed_both_true_both_false	649
CWM_detailed_one_true	323
CWM_brief_both_true_both_false	735
CWM_brief_one_true	429
DQ_true	68
DQ_false	509

Only non-carriers of the sensitive attribute (required for estimation of false positives)

CWM_detailed_both_true_both_false	298
CWM_detailed_one_true	98
CWM_brief_both_true_both_false	375
CWM_brief_one_true	129
DQ_true	6
DQ_false	231

Only carriers of the sensitive attribute (required for estimation of false negatives)

CWM_detailed_both_true_both_false	351

CWM_detailed_one_true	225
CWM_brief_both_true_both_false	360
CWM_brief_one_true	300
DQ_true	62
DQ_false	278

A new approach to detecting cheating in sensitive surveys: The Cheating Detection Triangular Model

Julia Meisters*, Adrian Hoffmann*, and Jochen Musch

University of Duesseldorf

«fn» *J. Meisters and A. Hoffmann contributed equally to this work.

Author Note

Julia Meisters, Adrian Hoffmann, and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Grant number 393108549.

Correspondence concerning this article should be addressed to Julia Meisters, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, Floor 00, Room 26, 40225 Duesseldorf, Germany. E-mail: julia.meisters@uni-duesseldorf.de

Abstract

By granting full confidentiality, indirect questioning techniques such as the randomized response technique aim to control social desirability bias in surveys of sensitive topics. To improve upon previous indirect questioning techniques, we propose the new Cheating Detection Triangular Model (CDTRM). Similar to the Cheating Detection Model (CDM; Clark & Desharnais, 1998), it includes a mechanism for detecting instruction non-adherence, and similar to the Triangular Model (TRM; Yu, Tian, & Tang, 2008), it uses simplified instructions to improve respondents' understanding of the procedure. Based on a comparison with the known prevalence of a sensitive attribute serving as external criterion, we report the first strong validation of the CDM, the TRM and the CDTRM. All three questioning techniques underestimated the known prevalence of the sensitive attribute; however, underestimation was more severe when a direct question was used. When questioned indirectly, more carriers of the sensitive attribute were detected as such. However, indirect questions also more frequently misidentified non-carriers as carriers. The CDTRM and the CDM estimated the known prevalence of the sensitive attribute equally well, but respondents evaluated the CDTRM more positively. Problematically, TRM prevalence estimates were found to be moderated by randomization probability. This finding contradicts a central assumption of the model and calls the validity of conventional TRM estimates into question. Based on our results, the CDTRM appears to be the best choice among the competing indirect questioning techniques; nevertheless, specificity was superior when respondents were questioned directly.

Keywords: Randomized Response, Cheating Detection Model, Triangular Model, validity

A new approach to detecting cheating in sensitive surveys: The Cheating Detection Triangular Model

Estimating the prevalence of sensitive attitudes and behaviors such as xenophobia, drug use, doping, or tax evasion is of high societal relevance. However, self-reports on such sensitive topics often lack validity because respondents fear negative consequences when answering honestly. Instead of providing accurate answers, some respondents choose to present themselves in a more socially acceptable light by under-reporting socially undesirable and over-reporting socially desirable attributes (Krumpal, 2013; Paulhus, 1991; Phillips & Clancy, 1972; Tourangeau & Yan, 2007).

The randomized response technique (RRT; Warner, 1965) was developed to control for the influence of social desirability bias by guaranteeing the confidentiality of individual responses. In the original RRT, respondents are presented with two opposing sensitive statements A and B (e.g. Statement A: 'I have taken cocaine' and Statement B: 'I have never taken cocaine'). Instead of answering to any of these statements directly, respondents are instructed to use an external randomization device (e.g. a die) that determines the statement they have to respond to with either 'true' or 'false'. The randomization outcome is known only to the respondent and is kept secret from the interviewer. Therefore, the interviewer does not know which statement was answered by any individual respondent, and therefore cannot tell whether a given respondent is a carrier of the sensitive attribute. However, prevalence estimates for the sensitive attribute can be computed on the sample level based on the known distribution of the randomization outcomes. Various variants of the RRT have been proposed and are reviewed in Chaudhuri (2011) and Chaudhuri and Christofides (2013).

Meta-analyses have shown that prevalence estimates obtained via RRTs are more valid than estimates obtained via a conventional direct question (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). However, while RRT estimates were usually closer to the true value, they still often underestimated the prevalence of sensitive attributes with known prevalence (Lensvelt-Mulders et al., 2005; Wolter & Preisendörfer, 2013). In some studies, RRTs even yielded prevalence estimates for socially undesirable attributes below those obtained with a conventional direct question and therefore presumably less valid (Holbrook & Krosnick, 2010; John, Loewenstein, Acquisti, & Vosgerau, 2018). One possible explanation for such inconsistent findings is that some respondents do not follow the instructions (Holbrook & Krosnick, 2010; Krumpal, 2012), perhaps because they fail to understand the questioning format, lack trust in the randomization procedure, or fear being falsely associated with the sensitive attribute (Edgell, Himmelfarb, & Duchan, 1982; John et al., 2018; Krumpal, 2012; Landsheer, van der Heijden, & van Gils, 1999).

Two main approaches have been proposed to address the problem of instruction non-adherence in RRTs. The first approach aims to make instruction non-adherence detectable. Examples of this approach include the Cheating Detection Model (Clark & Desharnais, 1998) and the Stochastic Lie Detector (Moshagen, Musch, & Erdfelder, 2012). The second approach aims to overcome instruction non-adherence by providing simplified instructions to increase understanding and perceived confidentiality among respondents. The Triangular Model and the Crosswise Model (Yu et al., 2008) are examples of this second approach.

The Cheating Detection Model

The Cheating Detection Model (CDM; Clark & Desharnais, 1998) is an extension of a forced choice variant of the RRT (Dawes & Moore, 1980). In the CDM, all respondents are

presented with the sensitive statement (e.g. 'I have taken cocaine'). Based on the outcome of a randomization procedure with a known distribution (e.g. the respondent's month of birth), respondents are requested either with probability *p* to simply answer 'true' regardless of their true status with respect to the sensitive attribute, or to answer honestly with probability 1-*p*. Since the distribution of the randomization outcomes is known, prevalence estimates for the sensitive attribute can be determined on the sample level when using the CDM. The status of individuals responding 'true' remains confidential because depending on their month of birth, both carriers and non-carriers of the sensitive attribute may be asked to provide a 'true' response. However, a 'false' response can only stem from non-carriers of the sensitive attribute, thus providing a 'safe' answer option respondents can choose to unambiguously deny being a carrier of the sensitive attribute. The CDM accounts for the possibility that some respondents will make use of this option instead of following the instructions. To this end, the CDM identifies three non-overlapping groups of respondents: honest carriers of the sensitive attribute (π), honest non-carriers of the sensitive attribute (β), and respondents who do not adhere to the instructions, but instead choose the self-protective response alternative (γ). Within the CDM, the true status of these non-adherent respondents remains unknown. It is therefore possible that none, some or all of the respondents not adhering to the instructions are carriers of the sensitive attribute. Hence, only a lower (π) and an upper ($\pi+\gamma$) bound for the prevalence of the sensitive attribute can be determined.

Since a randomized response model is non-identifiable if the number of its parameters exceeds the number of the available answer categories, the CDM has to be applied to two non-overlapping samples with different randomization probabilities *p1* and *p2* to make the model identifiable and obtain estimates for π, β, and γ (Clark & Desharnais, 1998). A graphical

representation of the CDM for the group with randomization probability *p1* is provided in Figure 1; for the second group, parameter *p1* is simply replaced with *p2*.

---------- INSERT FIGURE 1 HERE ----------

In applied settings, the CDM has been used to investigate the prevalence of phenomena such as doping (Frenger, Pitsch, & Emrich, 2016; Pitsch & Emrich, 2011; Pitsch, Emrich, & Klein, 2007; Schröter et al., 2016). To validate the CDM, several studies have compared the prevalence estimates obtained with the CDM to prevalence estimates obtained with a conventional direct question (DQ). Significantly higher prevalence estimates for socially undesirable attributes or significantly lower prevalence estimates for socially desirable attributes were interpreted as more valid and less distorted by social desirability bias. Employing this so-called 'weak' validation criterion, the CDM was found to be superior to DQ for investigating sensitive attributes such as tax evasion (Musch, Bröder, & Klauer, 2001), insufficient dental hygiene (Moshagen, Musch, Ostapczuk, & Zhao, 2010), cooperation in the prisoner dilemma game (Moshagen, Hilbig, & Musch, 2011), medication non-adherence (Ostapczuk, Musch, & Moshagen, 2011), opposition to granting asylum to civil war refugees, and prejudice against gay people (Moshagen & Musch, 2012). Other studies, however, did not detect significant differences between prevalence estimates obtained via the CDM and DQ for sensitive attributes such as academic dishonesty (Ostapczuk, Moshagen, Zhao, & Musch, 2009), xenophobia (Ostapczuk, Musch, & Moshagen, 2009), negative attitudes towards people with disabilities (Ostapczuk & Musch, 2011), and the use of renewable energy (Moshagen & Musch, 2012). Moreover, a recent study comparing several indirect questioning techniques found CDM

questions to be less comprehensible than conventional direct questions (Hoffmann, Waubert de Puiseau, Schmidt, & Musch, 2017).

Importantly, ‘weak’ validation studies cannot determine whether prevalence estimates accurately reflect the true prevalence of sensitive attributes because under- or overestimates can never be identified as such without knowing the ground truth (cf. Umesh & Peterson, 1991). Strong validation studies in which the prevalence of a sensitive attribute is known can provide better evidence regarding the validity of indirect questioning techniques (Moshagen, Hilbig, Erdfelder, & Moritz, 2014; Umesh & Peterson, 1991). Even more conclusive evidence can be obtained by conducting strong validation studies in which individual-level data with respect to the sensitive attribute are collected, because they make it possible to determine false positive and false negative rates and thus sensitivity and specificity (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; Umesh & Peterson, 1991). However, no strong validation study has yet been conducted for the CDM.

The CDM does not make any assumptions about whether respondents who disregard the instructions are carriers or non-carriers of the sensitive attribute, and to the best of our knowledge, no study has ever tried to empirically determine the true status of non-adherent respondents. However, substantial rates of instruction non-adherence (γ) have been found in most studies applying the CDM. The CDM provides a lower bound of π for the prevalence of the sensitive attribute (assuming that no non-adherent respondent is a carrier) and an upper bound of $\pi+\gamma$ (assuming that all non-adherent respondents are carriers). Therefore, high proportions of non-adherent respondents lead to uninformatively large ranges between the lower and upper bounds for the prevalence of the sensitive attribute. In a study of lifetime medication non-adherence among patients of medical practices, the proportion of non-adherent respondents was

found to be as high as 47.1% (Ostapczuk et al., 2011), and in a study on physical doping in sports, non-adherent respondents accounted for up to 60% of the sample (Schröter et al., 2016). This resulted in very large ranges between the lower bound of 32.7% and the upper bound of 79.8% for the prevalence of medication non-adherence (Ostapczuk et al., 2011), and between the lower bound of 11.9% and the upper bound of 76.8% for the prevalence of illegal physical doping (Schröter et al., 2016). Arguably, these very broad intervals are of limited practical use. Hence, it is important to collect more information regarding the true status of non-adherent respondents. For this reason, we decided to conduct the first strong validation of the CDM based on individual-level data to assess the true status of non-adherent respondents for the first time.

The Triangular Model

As an alternative to models enabling the detection of instruction non-adherence, so-called nonrandomized response techniques (NRRTs; Yu et al., 2008) such as the Triangular Model (TRM; Yu et al., 2008) have been developed. These techniques provide simplified instructions to increase respondents' understanding and perceived confidentiality, and thereby try to minimize or eliminate instruction non-adherence. In the TRM, respondents are presented with two statements—a sensitive statement A with unknown prevalence π (e.g. 'I have taken cocaine') and a non-sensitive statement B with known prevalence p that is used for randomization (e.g. 'I was born in November or December'; $p = .158$ according to official birth statistics, Pötzsch, 2012). Respondents are asked to indicate whether 'at least one of the statements is true (no matter which one)' or 'none of the statements is true'. A graphical depiction of the TRM is given in the middle of Figure 1. Similar to the CDM, one of the answer options in the TRM can be considered self-protective, as it explicitly excludes being a carrier of the sensitive attribute ('none of the statements is true'). In contrast to the CDM, however, the original TRM does not include a

parameter estimating the proportion of non-adherent respondents in the sample. Instead, it assumes that despite the presence of a 'safe' answer option, all respondents adhere to the instructions. The existence of a self-protective response option in the TRM along with the lack of a cheating detection mechanism may be responsible for some of the rather unsatisfactory results obtained with the TRM. Three weak validation studies found that the TRM provided prevalence estimates for sensitive attributes that were not higher, and therefore presumably not more valid, than estimates obtained via DQ (Erdmann, 2019; Hoffmann, Meisters, & Musch, 2019; Jerke & Krumpal, 2013). On the other hand, the prevalence of a non-sensitive control attribute could accurately be recovered using the TRM (Hoffmann et al., 2019). Moreover, a strong validation study evaluating the TRM based on individual-level data has not yet been conducted. The present study attempted to close this gap.

The Cheating Detection Triangular Model

Ideally, an indirect questioning technique should both maximize the proportion of respondents who fully understand the instructions, and detect all respondents who do not follow them. However, no such model is currently available. We therefore propose the new Cheating Detection Triangular Model (CDTRM), which combines the strengths of its two predecessor models, the simplified instructions of the TRM and a CDM-like mechanism for detecting non-adherent respondents.

In the CDTRM, respondents receive the exact same instructions and answer options as in the TRM. However, like in the CDM—and in contrast to the original TRM—we assume that the sample can be divided into a proportion of honest carriers of the sensitive attribute (π), a proportion of honest non-carriers (β), and a proportion of respondents who completely ignore the instructions and always choose the self-protective alternative (γ). Just like the CDM, the

CDTRM is based on collecting two independent samples with different randomization probabilities *p1* (e.g. *p1* = .158) and *p2* (e.g. *p2* = .842) to accommodate the additional parameter and make the model identifiable. A graphical depiction of the CDTRM for the first group with randomization probability *p1* is given in Figure 1; like in the CDM, parameter *p1* is simply replaced with *p2* for the second group.

Aims of the study

In the present study, we wanted to conduct the first strong validations of 1) the Cheating Detection Model (CDM), which offers a mechanism to detect the proportion of non-adherent respondents; 2) the Triangular Model (TRM), which seeks to minimize instruction non-adherence through simplified instructions; and 3) the newly proposed Cheating Detection Triangular Model (CDTRM) that combines the strengths of the two approaches by providing both simplified instructions and a mechanism to detect instruction non-adherence. We wanted to compare all three questioning techniques to each other and to a direct questioning control condition. Using an experimentally induced sensitive attribute allowed us not only to compare the known prevalence of the sensitive attribute with the prevalence estimates of the competing models, but also to compute the questioning techniques' sensitivity and specificity based on individual-level data. We expected the indirect questioning techniques (CDM, TRM, CDTRM) to provide more valid prevalence estimates than a conventional direct question. We further expected the newly proposed CDTRM to provide more valid prevalence estimates than both the CDM and the TRM, since only the CDTRM combines the unique strengths of both of these approaches. Finally, the current study investigated for the first time whether non-adherent respondents in the CDM and the CDTRM are carriers of the sensitive attribute, non-carriers, or a mixture of both.

Methods

Participants

Respondents were recruited via a German commercial online panel provider. As a prerequisite for participation and to rule out language difficulties as a potential explanation for an insufficient understanding of the questions, respondents had to be at least 18 years old and had to indicate German as their native language. These criteria were met by 3308 respondents. Of these, 521 dropped out before answering the sensitive question (15.75% of the initial sample). The dropout rate differed significantly across the three questioning technique conditions, DQ: 0.36%, CDM: 19.74%; (CD)TRM: 18.02%, $\chi^2(2) = 121.90$, $p < .001$, *Cramer's V* = .19. In a pairwise comparison, dropout rates did not differ significantly between the CDM and (CD)TRM conditions, indicating that the dropout rate was generally lower in the direct than in the three indirect questioning conditions, $\chi^2(1) = 1.32$, $p = .250$, *Cramer's V* = .02.

The final sample consisted of the 2787 respondents (57.73% female) who answered the sensitive question. There were 558 respondents in the DQ condition, 554 respondents in the CDM condition with randomization probability $p1$, 556 respondents in the CDM condition with randomization probability $p2$, 563 respondents in the (CD)TRM condition with randomization probability $p1$, and 556 respondents in the (CD)TRM condition with randomization probability $p2$.

To ensure sufficient statistical power (> .80) for the planned prevalence comparisons, twice as many respondents were allocated to the indirect questioning conditions compared to the DQ condition to compensate for the generally lower efficiency of indirect questioning techniques (Moshagen et al., 2012; Ulrich, Schröter, Striegel, & Simon, 2012). The distribution of gender,

age group and educational achievement in the final sample did not significantly differ across questioning technique conditions (see Table 1).

---------- INSERT TABLE 1 HERE ----------

Measures

Anagram Cheating Task. To conduct a strong validation, we experimentally induced a sensitive attribute with known prevalence using the anagram paradigm established by Hoffmann, Diedenhofen, Verschuere, and Musch (2015). In this paradigm, respondents are first asked to solve some anagrams in private and are then given an opportunity to cheat by over-reporting their performance. To this end, we showed the respondents three scrambled words in sequential order, and asked them to solve these anagrams within 20 seconds each. Rather than typing in the solutions, we asked respondents to simply solve the anagrams in their head. Twenty seconds after the presentation of each word, the next anagram was presented. The first two anagrams were very easy and were solved by 99% of the participants in a pilot study (Hoffmann et al., 2015); the third anagram, however, was extremely difficult and solved by only 1% of the pilot study participants. After the third anagram, the solutions to all three anagrams were presented. Respondents were then given an opportunity to cheat by over-reporting their performance when answering the following question: ‘Did you solve all three anagrams in the available time?’. The two answer options read: ‘No, I solved less than three anagrams’ and ‘Yes, I solved all three anagrams’. An incentive to cheat was provided by announcing that only respondents who had solved all three anagrams would be allowed to participate in a lottery at the end of the survey. The identity of the cheaters was protected because the respondents did not report their anagram

solutions. However, because the pilot study had shown that only 1% of respondents could actually solve all three anagrams (cf. Hoffmann et al., 2015), respondents claiming to have solved all anagrams were considered cheaters. This allowed us to conduct a strong validation based on individual-level data, going beyond previous investigations employing the CDM and the TRM.

Sensitive question. The sensitive statement was formulated identically in all questioning technique conditions and read: ‘On the anagram task, I claimed that I had solved more anagrams than I had actually solved.’

Direct questioning. In the DQ condition, respondents were asked to evaluate the sensitive statement by selecting ‘true’ or ‘false’.

Cheating Detection Model. In the CDM condition with randomization probability *p1*, respondents were asked to answer ‘true’ if they were born in November or December, irrespective of whether they had actually cheated on the anagram task, or to answer the question honestly if they were born between January and October. The probability of being directed to answer ‘true’ due to being born in November or December was $p1 = .158$ according to official birth statistics (Pötzsch, 2012). In the second CDM condition, these instructions were reversed and a different randomization probability $p2 = 1 - p1 = .842$ was employed. Thus, respondents were instructed to answer ‘true’ irrespective of their status with respect to the sensitive attribute if they were born between January and October, and to respond honestly if they were born in November or December.

Cheating Detection Triangular Model. In the (CD)TRM conditions, respondents received the instructions for the Triangular Model. We presented them with a sensitive statement and a non-sensitive statement with known prevalence p_i ('I was born in November or December',

p1 = .158, or 'I was born between January and October', *p2* = .842; Pötzsch, 2012). The answer options read: '*None* of the statements is true' versus '*At least one* of the statements (no matter which one) is true'.

Comprehension questions in the CDM and CDTRM conditions. In both indirect questioning conditions (CDM and CDTRM), respondents were given detailed instructions on the respective model and had to answer four comprehension questions to ensure that they had properly understood the procedure (cf. Meisters, Hoffmann, & Musch, 2019). These questions covered all four possible cases of carrying vs. not carrying the sensitive and the non-sensitive attributes. For example, the first comprehension question read 'Assuming you were born in February, and assuming you had *not* exaggerated your report on the number of anagrams solved: Which answer would you have to give?' The response options were identical to those provided in the respective model, and were presented in random order. For each of the comprehension questions, respondents received feedback on their performance. Comprehension questions that had been answered incorrectly were presented again up to a maximum of three times (including the first presentation).

Self-reported response behavior and subjective evaluation of the questioning technique. At the end of the questionnaire, respondents were asked to evaluate their subjective perception of the sensitive question and provide information on their response behavior by indicating their agreement with nine statements on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The nine statements are listed in Table 5 and included questions referring to the subjectively perceived comprehensibility and privacy protection afforded by the sensitive question, as well as a question asking whether respondents had given random responses.

Procedure

After accessing the online questionnaire, respondents were asked to read a short introduction, provide informed consent, and answer some demographic questions regarding their age, gender and level of educational achievement. The subsequent pages contained an explanation of the anagram task and an opportunity to work on two example anagrams to familiarize oneself with the task. Before starting the main anagram task, respondents were informed that they would be given an opportunity to participate in a lottery of 100€, 50€ and 30€ if they were able to solve all three anagrams within the allotted time. After the anagram task, respondents had the opportunity to cheat by over-reporting their performance.

Employing one of the three questioning technique formats (DQ, CDM, [CD]TRM), they were then asked whether they had cheated in the anagram task. Respondents in the CDM and (CD)TRM conditions received detailed instructions on the questioning technique prior to this question, and had to answer the four comprehension questions first to measure whether they had understood the procedure (cf. Meisters et al., 2019). After answering the sensitive question, respondents were asked to provide information on their subjective experience and their response behavior. At the end of the survey, all respondents were given the opportunity to participate in the lottery, and were debriefed and thanked for their participation. Contrary to the initial announcement, participation in the lottery was not limited to respondents who had cheated on the anagram task in order to not discriminate against honest respondents.

Statistical Analyses

To obtain and compare parameter estimates, we established multinomial processing tree (MPT) models (Batchelder, 1998; Batchelder & Riefer, 1999) for all questioning techniques as detailed in, for example, Moshagen et al. (2011); Moshagen et al. (2012); and Ostapczuk et al.

(2011). In all models, parameter π represents the unknown prevalence of the sensitive attribute. Additional parameters in the CDM and CDTRM reflect the unknown proportions of non-carriers of the sensitive attribute (β) and respondents not adhering to the instructions (γ = 1-π-β). In both the CDM and the CDTRM, the status of non-adherent respondents with respect to the sensitive attribute is assumed to be unknown. Therefore, parameter π in these models represents only a lower bound for the prevalence of the sensitive attribute (assuming that all non-adherent respondents are non-carriers); the upper bound is given by π+γ (assuming that all non-adherent respondents are in fact carriers). The parameters *p1* and *p2* denote the known prevalence of the non-sensitive attribute used for randomization.

On the basis of the empirically observed answer frequencies, parameter estimates were derived using the expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977; Hu & Batchelder, 1994) as implemented in the software multiTree (Moshagen, 2010). The asymptotically χ^2-distributed log-likelihood statistic G^2 was used to assess model fit. For parameter comparisons, the model fit of a restricted model setting the parameters either equal or equal to a constant was compared to the model fit of a baseline model in which the parameters were estimated freely. Significant changes in model fit (ΔG^2) indicated that the restricted model had a worse fit than the baseline model, suggesting that the respective parameter restriction was inadmissible.

To obtain prevalence estimates based on the CDTRM, the TRM instructions were applied to two groups with different randomization probabilities *p1* and *p2*. Thus, as a by-product of employing the CDTRM, we also employed two TRM conditions that differed regarding their randomization probability, unlike traditional TRM applications. Comparing the resulting two estimates allowed us to conduct the first empirical test of an assumption inherent to the TRM that

TRM prevalence estimates are unaffected by randomization probability. No such test had previously been conducted because the original TRM (Yu et al., 2008) requires only one group of respondents. Therefore, it was not possible to detect a potential unwanted influence of randomization probability on the prevalence estimates in previous studies employing the TRM. If prevalence estimates differed as a function of randomization probability, this would contradict a central assumption underlying the TRM. Such a finding would call into question the applicability of the TRM and suggest that the obtained prevalence estimates are not trustworthy (cf. Heck, Hoffmann, & Moshagen, 2018).

Sensitivity and specificity of the models. To determine the sensitivity and specificity of the models, individual-level data on the respondents' status with respect to the sensitive attribute are required. Given that the probability of solving the third anagram in the anagram task was very low (1%, cf. Hoffmann et al., 2015), respondents who claimed to have solved all three anagrams were classified as cheaters and thus as carriers of the sensitive attribute. Accordingly, respondents who admitted that they had solved fewer than three anagrams were classified as non-carriers. This classification allowed us to compute separate prevalence estimates for carriers and non-carriers, respectively. The sensitivity of a questioning technique is defined as the proportion of carriers who are identified as such. In the DQ and TRM conditions, model sensitivity could therefore be determined by estimating the prevalence π among known carriers of the sensitive attribute. In the CDM and CDTRM conditions, π (assuming that no non-adherent respondent is a carrier) and $\pi+\gamma$ (assuming that all non-adherent respondents are carriers) provide a lower and upper bound for the share of respondents identified as carriers, respectively. Thus, in the CDM and CDTRM conditions, a lower and upper bound for sensitivity can be obtained by estimating π and $\pi+\gamma$ for known carriers of the sensitive attribute.

The specificity of a questioning technique is defined as the proportion of non-carriers who are identified as such. In the DQ and TRM conditions, model specificity could therefore be determined by estimating 1-π among known non-carriers of the sensitive attribute. In the CDM and CDTRM, a lower and upper bound for the share of respondents identified as non-carriers can be calculated as β (= 1-π-γ; assuming that all non-adherent respondents are actually carriers) and β+γ (= 1-π; assuming that all non-adherent respondents are actually non-carriers). Thus, in the CDM and CDTRM conditions, a lower and upper bound for specificity can be obtained by estimating β and β+γ among non-carriers.

Proportion of carriers among non-adherent respondents. To assess the share of carriers among non-adherent respondents (γ) in the CDM and CDTRM, we extended the original MPT models by generating separate branches for carriers and non-carriers of the sensitive attribute. The parameter *c_g* (carriers gamma) in these extended models reflects the proportion of carriers among non-adherent respondents. The resulting extended models had one degree of freedom each; their model fit could therefore be tested.

The multiTree equations for all models and the empirically observed answer frequencies - both for the overall sample and separately for carriers and non-carriers - are provided in Appendices A and B in the electronic supplement.

Results

Parameter estimates, sensitivity and specificity

Estimates of the prevalence, sensitivity and specificity along with their standard errors are given in Table 2. Comparisons of the prevalence estimates calculated using the competing

questioning techniques can be found in Table 3. Table 4 provides comparisons of the techniques' sensitivity and specificity.

---------- INSERT TABLES 2, 3, AND 4 HERE ----------

Exploratory analysis for the extended TRM. Contrary to a central assumption of the model, TRM prevalence estimates differed as a function of randomization probability ($\pi 1 = 22.58\%$, $SE = 2.38\%$; $\pi 2 = 0.00\%$, $SE = 8.36\%$; $\Delta G^2(1) = 45.79$, $p < .001$). Prevalence estimates could therefore not be pooled across conditions and were considered untrustworthy (cf. Heck et al., 2018). This precluded a meaningful comparison of the validity of the TRM and the validity of the other models.

Prevalence estimates for the sensitive attribute. In the anagram task, 59.13% of the respondents cheated. Cheating rates did not differ significantly across conditions (DQ: 57.71%, CDM: 59.73%, CDTRM: 59.25%, $\chi^2(2) = 0.64$, $p = .73$, *Cramer's V* = .02). Prevalence estimates $\hat{\pi}$ for cheating in the anagram task obtained via DQ, CDM, and CDTRM were 15.59%, 29.70%, 25.07%, respectively. Thus, all questioning techniques significantly underestimated the known prevalence of cheating (see Table 3). However, for the CDM and CDTRM, $\hat{\pi}$ only represents a lower bound for the prevalence of cheating on the anagram task under the assumption that all non-adherent respondents are non-carriers who did not cheat on the anagram task. Alternatively assuming that all non-adherent respondents are carriers results in upper bounds of 38.94% and 38.36% for the prevalence of cheating in the CDM and CDTRM conditions, respectively. These upper bounds are closer to the true prevalence but still underestimate it. Comparisons of the prevalence estimates provided by the different questioning techniques revealed that both the

lower and upper bound estimates for the prevalence of cheating in the CDM and CDTRM conditions significantly exceeded the estimates obtained via DQ. However, neither the lower nor upper bound estimates differed significantly between the CDM and CDTRM conditions. Estimates for the proportion of non-adherent respondents (γ) were 9.24% and 13.28% in the CDM and CDTRM conditions, respectively. Both γ parameters were significantly higher than 0%, indicating a substantial proportion of non-adherent respondents in both conditions. Neither the γ nor β parameters differed significantly between the CDM and the CDTRM conditions. Detailed inferential statistics for the above comparisons are provided in Table 3.

Sensitivity. In the DQ and TRM conditions, model sensitivity was assessed by estimating the prevalence π among carriers. In the CDM and CDTRM conditions, the lower and upper bounds for sensitivity were determined by computing π and π+γ for carriers. The sensitivity was significantly below 100% for all questioning techniques. In the DQ condition, sensitivity was estimated at 25.47%. In the CDM condition, sensitivity estimates ranged from 39.59% (lower bound) to 52.00% (upper bound); and in the CDTRM condition, they ranged from 36.18% (lower bound) to 51.34% (upper bound). Pairwise comparisons revealed that both the lower and upper bound estimates for sensitivity in the CDM and CDTRM conditions were significantly higher than the sensitivity of DQ. Neither the lower nor upper bound estimates differed significantly between the CDM and CDTRM conditions. Detailed inferential statistics for the parameter comparisons are given in Table 4.

Specificity. In the DQ and TRM conditions, model specificity was assessed by estimating 1-π among non-carriers. In the CDM and CDTRM, the lower and upper bounds for specificity were determined by estimating β (= 1-π-γ) and β+γ (= 1-π) for non-carriers. The specificity was significantly below 100% for all questioning techniques. In the DQ condition, specificity was

estimated at 97.88%. In the CDM condition, specificity estimates ranged from 79.13% (lower bound) to 83.69% (upper bound). In the CDTRM condition, the specificity ranged from 79.46% (lower bound) to 90.19% (upper bound). Pairwise comparisons revealed that the specificity of DQ was significantly higher than even the upper bounds of CDM and CDTRM. Neither the lower nor the upper bound estimates of specificity differed significantly between the CDM and CDTRM conditions. Detailed inferential statistics for the parameter comparisons are given in Table 4.

Proportion of carriers among non-adherent respondents. Both extended MPT models established to estimate the proportion of carriers among non-adherent respondents fit the data well, CDM: $G^2(1) = 2.50$, $p = .114$; CDTRM: $G^2(1) = 0.85$, $p = .357$. Consequently, the resulting parameter estimates were considered trustworthy (cf. Heck et al., 2018). The estimated proportion of carriers among non-adherent respondents (c_g) was 80.16% (SE = 11.94%) in the CDM and 67.25% (SE = 8.55%) in the CDTRM condition, respectively.

Comprehensibility and evaluation of the questioning techniques

Objective comprehensibility of the questioning techniques. To assess whether CDM and CDTRM differed with respect to their objective comprehensibility, we computed the proportion of respondents who correctly answered all comprehension questions in their first attempt (CDM: 13.42%, CDTRM: 26.99%), in their first or second attempt (CDM: 49.73%, CDTRM: 55.94%), or in any attempt (CDM: 76.16%, CDTRM: 76.14%). Three χ^2 tests employing a Bonferroni-corrected alpha level revealed that compared to respondents in the CDM condition, respondents in the CDTRM condition answered all comprehension questions correctly significantly more often in their first attempt, $\chi^2(1) = 63.53$, $p < .001$, *Cramer's V* = .17, or in their first or second attempt, $\chi^2(1) = 8.63$, $p = .003$, *Cramer's V* = .06. However, no difference

between the CDM and CDTRM conditions was observed regarding the proportion of respondents who answered all comprehension questions correctly by the third attempt, $\chi^2(1) < 0.001$, $p = .994$, *Cramer's V* < .001.

Self-reported response behavior and subjective evaluation of the questioning technique. Consistency analyses indicated that the nine items capturing respondents' subjective evaluation of the questioning techniques were highly intercorrelated (*Cronbach* $\alpha = .84$). Therefore, we computed the mean of all items and compared this mean across questioning techniques. Higher values indicated a more positive evaluation of the questioning technique. A one-way between-subjects ANOVA revealed a significant effect of questioning technique (DQ, CDM, [CD]TRM) on subjective evaluations, $F(2, 2784) = 336.49$, $p < .001$, $\eta_p = .20$.

Bonferroni-corrected post-hoc tests showed that DQ was evaluated significantly more positively than the (CD)TRM, and that the (CD)TRM was evaluated significantly more positively than the CDM. This pattern of results was observed for every item of the scale except for the item measuring adherence to the instructions ('I carefully read and followed all instructions'), for which Bonferroni-corrected post-hoc tests did not reveal a significant difference between the CDM and CDTRM conditions. Detailed statistics for these analyses can be obtained from Table 5.

---------- INSERT TABLE 5 HERE ----------

Exploratory analyses of respondents with high understanding of the CDM and CDTRM

To further explore how high respondent vigilance and a deeper understanding of the questioning techniques influenced the results, we conducted exploratory analyses on the validity of the CDM and CDTRM in the subsample of respondents who answered all comprehension

questions correctly on their first attempt. We expected that limiting the analysis to these respondents would increase the validity of the prevalence estimates. This subsample was quite small compared to the total sample (CDM: n = 149, CDTRM: n = 302), and was even smaller when performing separate analyses to assess sensitivity and specificity for carriers and non-carriers. This limited the power of some of the following analyses; however, their results nevertheless provide valuable insights into the functioning of the models and the validity of estimates obtained under optimized conditions.

Compared to the total sample (CDM: 29.70%, CDTRM: 25.07%), prevalence estimates were slightly lower in the subsample of respondents with a high level of understanding who solved all comprehension questions on their first attempt (CDM: 21.94%, CDTRM: 22.92%). The proportion of non-adherent respondents (γ) was also lower among respondents with a high level of understanding (CDM: 0.62%, CDTRM: 5.73%; total sample: CDM: 9.24%, CDTRM: 13.28%). The sensitivity of the CDM was slightly lower in the subsample of respondents with high understanding (lower bound: 34.29%, upper bound: 37.79%) than in the total sample (lower bound: 39.59%, upper bound: 52.00%). For the CDTRM, a lower bound of 39.17% and an upper bound of 47.28% for sensitivity were calculated in the subsample of respondents with high understanding. Thus, the lower bound was higher and the upper bound lower than in the total sample (lower bound: 36.18%, upper bound: 51.34%). This reduction in range was due to the reduced number of non-adherent respondents (i.e., the smaller γ) in the subsample with a high level of understanding.

The specificity of the CDM was slightly higher in the subsample of respondents with high understanding (lower bound: 86.74%, upper bound: 86.74%) than in the total sample (lower bound: 79.13%, upper bound: 83.69%). For the CDTRM, a lower bound of 91.28% and an upper

bound of 94.42% for the specificity were calculated in the subsample of respondents with high understanding. Thus, in the CDTRM, both the lower and upper bounds were higher than in the total sample (lower bound: 79.46%, upper bound: 90.19%).

In summary, limiting the analysis to respondents with a high level of understanding reduced instruction non-adherence and increased the specificity of both CDM and CDTRM estimates. However, it also slightly reduced the sensitivity of CDM estimates. Only the sensitivity of the CDTRM was largely unaffected. Detailed statistics for these analyses are provided in Appendix C in the electronic supplement.

We also tested whether there were any demographic differences between respondents correctly answering all comprehension questions on their first attempt (n = 451) and respondents failing to do so (n = 1778). The two groups of respondents did not differ significantly with regard to gender, $\chi^2(1) = 2.84$, $p = .092$, *Cramer's V* = .04. However, significant differences were observed with regard to age group, $\chi^2(5) = 38.03$, $p < .001$, *Cramer's V* = .13, and educational achievement, $\chi^2(8) = 68.80$, $p < .001$, *Cramer's V* = .18. Respondents answering all comprehension questions correctly in their first attempt were slightly younger and slightly higher educated than respondents who failed to correctly answer all comprehension questions in their first attempt. Detailed descriptive statistics can be obtained from Table 4 in Appendix C in the electronic supplement.

We also conducted a 2 (model comprehension: correctly answering all comprehension questions in the first attempt vs. failing to correctly answer at least one comprehension question in the first attempt) x 2 (questioning technique: CDM vs. CDTRM) between-subjects ANOVA with the scale mean for the nine items capturing the respondents' subjective evaluation of the questioning techniques as the dependent variable. This analysis revealed a significant main effect

of correctly answering the comprehension questions in the first attempt, $F(1, 2225) = 330.04$, $p < .001$, $\eta_p = .13$, and a significant main effect of questioning technique condition, $F(1, 2225) = 48.87$, $p < .001$, $\eta_p = .02$, but no interaction between these factors, $F(1, 2225) = 3.57$, $p = .059$, $\eta_p = .002$. Respondents correctly answering all comprehension questions in their first attempt evaluated the questioning technique more favorably ($M = 5.86$, $SD = 0.83$) than respondents failing to correctly answer at least one of the comprehension questions in their first attempt ($M = 4.76$, $SD = 1.05$). Overall, the CDTRM ($M = 5.21$, $SD = 1.08$) was rated more favorably than the CDM ($M = 4.75$, $SD = 1.08$) on almost all scale items. However, this main effect was qualified by an interaction for the item 'The question was cumbersome to answer'. Only among respondents who correctly answered all comprehension questions in their first attempt were CDM questions rated significantly more cumbersome to answer than CDTRM questions; among respondents failing to correctly answer at least one comprehension question in their first attempt, both questioning techniques were rated equally cumbersome to answer. For the two items 'I carefully read and followed all instructions' and 'I just randomly ticked one of the answers', only a significant main effect of correctly answering all comprehension questions on the first attempt was observed, with more favorable evaluations provided by respondents who correctly answered all comprehension questions in their first attempt. For these two items, the main effect of questioning technique and the interaction between the two factors were not significant. Detailed statistics for these ANOVAs can be obtained from Table 5 in Appendix C of the electronic supplement.

Discussion

The current study presents the first individual-level validation of three competing indirect questioning techniques: the Cheating Detection Model (CDM; Clark & Desharnais, 1998), the Triangular Model (TRM; Yu et al., 2008), and the newly proposed Cheating Detection Triangular Model (CDTRM), which combines the CDM's cheating detection mechanism and the TRM's simplified, easy-to-understand instructions. Going beyond previous studies, we applied an extended version of the TRM consisting of two groups with different randomization probabilities. This enabled a comparison of prevalence estimates across groups and uncovered a violation of the main assumption inherent to the TRM that prevalence estimates are not dependent on randomization probability. A potential reason for this unexpected finding could be instruction non-adherence and the fact that self-protective answers differentially affect the two conditions differing with regard to randomization probability. In particular, the TRM (see Figure 1) cannot account for a proportion of 'At least one statement is true'-responses that is below the randomization probability p. For this reason, model assumptions are more likely to be violated and prevalence estimates are more distorted by self-protective answer behavior in the condition with a high randomization probability. This is also a likely reason for why in the present study, a high randomization probability ($p2$ = .842) resulted in an implausible boundary prevalence estimate of 0%. Unfortunately, however, the TRM does not allow for the detection of instruction non-adherence, making it difficult to pinpoint the exact reasons for its failure. Nevertheless, given the demonstrated violation of one of its central assumptions, it seems difficult to justify further use of the TRM in its original format. Moreover, our findings regarding the TRM emphasize the importance of taking instruction non-adherence into account (cf. Wu & Tang, 2016), as the CDTRM we introduce in the present study is able to do.

The CDTRM uses the same set of instructions as the TRM, but employs the measurement model of the CDM. Therefore, unlike the TRM, the CDTRM can provide an estimate of the prevalence of instruction non-adherence. In the present study, the proportion of respondents who did not follow the CDTRM instructions but instead chose the self-protecting alternative was significantly higher than zero (13%). This finding provides direct evidence for the occurrence of instruction non-adherence in the CDTRM, and indirect evidence that instruction non-adherence likely also impaired the validity of the original TRM.

Both the CDTRM and the CDM provided significantly higher estimates for the prevalence of the sensitive attribute than a conventional direct question. Thus, they both met a weak validation criterion ("more is better"), according to which higher prevalence estimates can be considered less biased by socially desirable responding and therefore more valid. However, the CDTRM, CDM, and DQ all significantly underestimated the known prevalence of the sensitive attribute and thus failed to meet the strong validation criterion represented by the ground truth we established in the present study. Moreover, all questioning techniques had an estimated sensitivity and specificity considerably lower than 100%. With respect to sensitivity, CDTRM and CDM outperformed DQ; however, with respect to specificity, DQ performed better than the CDTRM and the CDM. A similar pattern of results was recently observed for the Crosswise Model (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018); it might therefore potentially be generalizable to other indirect questioning techniques as well. In this case, the decision for or against the application of indirect questioning techniques should be based on whether the aim is to maximize sensitivity or specificity in a given context. Maximizing sensitivity and therefore choosing indirect over direct questioning techniques might be preferable in situations in which underestimating the prevalence of a sensitive attribute can have serious

negative consequences. This is particularly the case when a study's ultimate goal is to assess the need for educational or preventive measures, such as in surveys on medication non-adherence (Ostapczuk et al., 2011), tax evasion (Korndörfer, Krumpal, & Schmukle, 2014; Kundt, Misch, & Nerré, 2017), or xenophobia (Hoffmann et al., 2019; Ostapczuk, Musch, et al., 2009). Maximizing specificity and thus choosing direct over indirect questioning techniques might be preferable if an overestimation of the prevalence of a sensitive attribute is associated with serious negative consequences such as reputational damage, as can be true for illegal or morally reprehensible behavior, where *in dubio pro reo* might be the necessary choice.

A comparison of the CDTRM and the CDM revealed that both models obtained comparable estimates for the prevalence of honest carriers (π), honest non-carriers (β), and non-adherent respondents (γ), as well as for sensitivity and specificity. However, the objective comprehensibility of the (CD)TRM exceeded the comprehensibility of the CDM, as evidenced by the fact that in the (CD)TRM condition, significantly more respondents correctly answered all comprehension questions in their first attempt. This finding corresponds with the basic objective of nonrandomized response techniques such as the TRM to facilitate a higher level of understanding among respondents than randomized response techniques such as the CDM (cf. Yu et al., 2008). In accordance with this finding, a previous study assessing the proportion of correct responses for several indirect questioning techniques (Hoffmann et al., 2017) also reported problems in understanding the relatively complex instructions of the CDM.

While CDTRM and CDM prevalence estimates were comparable, considerable differences between the two models were observed with respect to the respondents' self-reported evaluations. Overall, respondents evaluated the (CD)TRM more favorably than the CDM, and fewer respondents reported having answered randomly in the (CD)TRM compared to the CDM

condition. Respondents indicated the same degree of carefulness in reading and following the instructions for both models. However, respondents in the CDM condition reported a more negative attitude towards indirect questioning techniques than respondents in the (CD)TRM condition; they might therefore be less likely to participate in future surveys incorporating these techniques. Taking both the respondents' objective performance and their subjective evaluations of the questioning techniques into account, we recommend employing the CDTRM over the CDM.

The status of non-adherent respondents with respect to the sensitive attribute in the CDM (and thus also in the CDTRM) usually remains unknown. Other indirect questioning techniques, such as the Stochastic Lie Detector (Moshagen et al., 2012), make stronger assumptions and assume that only carriers have a motivation to disregard the instructions and choose self-protective responses. It has been difficult to test the validity of such assumptions. Employing a strong validation approach based on individual-level data, the present study was the first to empirically determine the proportion of carriers among non-adherent respondents. We found that this proportion was 80.16% and 67.25% in the CDM and CDTRM conditions, respectively, implying that the group of non-adherent respondents did not consist solely of carriers or non-carriers, but rather of a mixture of these two groups.

We also found that limiting the analysis to respondents with high comprehension who correctly answered all comprehension questions in their first attempt slightly reduced instruction non-adherence and increased the specificity of the CDTRM and CDM. However, analyzing only respondents with high levels of comprehension also slightly reduced the sensitivity of the CDM. For the CDTRM, the same analysis led to an increase in the lower bound estimate for sensitivity and a decrease in the upper bound estimate. This pattern of results indicates that for respondents

with a precise understanding of indirect questioning techniques, the proportion of false positives decreases, while the proportion of false negatives slightly increases. Similar patterns have been observed for another nonrandomized response technique, the Crosswise Model (Meisters et al., 2019), suggesting that such models are affected by a general tradeoff between false positives and false negatives, or sensitivity and specificity (Höglinger & Jann, 2018; Meisters et al., 2019). Future studies should therefore address the question of whether the instructions for indirect questioning techniques can be further optimized to increase both sensitivity and specificity.

Conclusions

The present results suggest that the newly proposed CDTRM, which combines the simplified instructions of the TRM with the cheating detection capability of the CDM, should be preferred over its two predecessor models. Unexpectedly and problematically, TRM prevalence estimates were also found to be influenced by the randomization probability, probably due to instruction non-adherence by some respondents, which unfortunately cannot be detected by this model. Compared to the CDM, the CDTRM obtained prevalence estimates of similar validity, but received substantially more positive evaluations with respect to the respondents' subjective experience of the questioning technique. We therefore recommend the CDTRM as a promising new indirect questioning technique that controls for the influence of socially desirable responding with easy-to-understand instructions, a mechanism to detect instruction non-adherence, and a high level of acceptance by respondents. However, we also note that all prevalence estimates obtained via CDTRM, CDM or conventional direct questions still underestimated the known prevalence of the sensitive attribute. Both the sensitivity and specificity of the models were far from perfect, with indirect questioning techniques achieving higher sensitivity but lower specificity compared to direct questions. We therefore recommend

choosing indirect over direct questioning techniques in situations in which maximizing sensitivity is of paramount importance; if, however, maximizing specificity is more important, choosing direct over indirect questioning might be advisable.

Open Practice Statement

All data and equation files necessary to reproduce the parameter estimates reported in the main analyses of the manuscript are provided in Appendices A and B in the electronic supplement.

References

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10*, 331-344. doi:10.1037/1040-3590.10.4.331

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86. doi:10.3758/Bf03210812

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton, Florida: Chapman & Hall, CRC Press, Taylor & Francis Group.

Chaudhuri, A., & Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys*. Berlin, Heidelberg: Springer.

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160-168.

Dawes, R. M., & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen [Guttman scaling of orthodox and randomized reactions]. In F. Petermann (Ed.), *Einstellungsmessung, Einstellungsforschung [Attitude measurement, attitude research]* (pp. 117–133). Göttingen: Hogrefe.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39*, 1-38.

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of Forced Responses in a Randomized-Response Model. *Sociological Methods & Research, 11*, 89-100. doi:10.1177/0049124182011001005

Erdmann, A. (2019). Non-Randomized Response Models: An Experimental Application of the Triangular Model as an Indirect Questioning Method for Sensitive Topics. *methods, data, analyses, 13*, 139-167. doi:10.12758/mda.2018.07

Frenger, M., Pitsch, W., & Emrich, E. (2016). Sport-Induced Substance Use - An Empirical Study to the Extent within a German Sports Association. *Plos One, 11*. doi:10.1371/journal.pone.0165103

Heck, D. W., Hoffmann, A., & Moshagen, M. (2018). Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior Research Methods, 50*, 1895-1905. doi:10.3758/s13428-017-0957-8

Hoffmann, A., Diedenhofen, B., Verschuere, B. J., & Musch, J. (2015). A strong validation of the Crosswise Model using experimentally induced cheating behavior. *Experimental Psychology, 62*, 403-414. doi:10.1027/1618-3169/a000304

Hoffmann, A., Meisters, J., & Musch, J. (2019). On the Validity of Non-Randomized Response Techniques: An Experimental Comparison of the Crosswise Model and the Triangular Model. [Manuscript submitted].

Hoffmann, A., Waubert de Puiseau, B., Schmidt, A. F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods, 49*, 1470-1483. doi:10.3758/s13428-016-0804-3

Höglinger, M., & Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis, 25*, 131-137. doi:10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *Plos One, 13*. doi:10.1371/journal.pone.0201770

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling into Question the Method's Validity. *Public Opinion Quarterly, 74*, 328-343. doi:10.1093/Poq/Nfq012

Hu, X., & Batchelder, W. H. (1994). The Statistical-Analysis of General Processing Tree Models with the Em Algorithm. *Psychometrika, 59*, 21-47. doi:10.1007/Bf02294263

Jerke, J., & Krumpal, I. (2013). Plagiate in studentischen Arbeiten [Plagiarism in Student Papers]. *methoden, daten, analysen, 7*, 347-368. doi:10.12758/mda.2013.017

John, L. K., Loewenstein, G., Acquisti, A., & Vosgerau, J. (2018). When and why randomized response techniques (fail to) elicit the truth. *Organizational Behavior and Human Decision Processes, 148*, 101-123. doi:10.1016/j.obhdp.2018.07.004

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology, 45*, 18-32. doi:10.1016/j.joep.2014.08.001

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research, 41*, 1387-1403. doi:10.1016/j.ssresearch.2012.05.015

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity, 47*, 2025-2047. doi:10.1007/s11135-011-9640-9

Kundt, T. C., Misch, F., & Nerré, B. (2017). Re-assessing the merits of measuring tax evasion through business surveys: an application of the crosswise model. *International Tax and Public Finance, 24*, 112-133. doi:10.1007/s10797-015-9373-0

Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response - A study of a method for improving the estimate of social security fraud. *Quality & Quantity, 33*, 1-12. doi:10.1023/A:1004361819974

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: thirty-five years of validation. *Sociological Methods & Research, 33*, 319-348. doi:10.1177/0049124104268664

Meisters, J., Hoffmann, A., & Musch, J. (2019). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? [Manuscript submitted].

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42-54. doi:10.3758/BRM.42.1.42

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An Experimental Validation Method for Questioning Techniques That Assess Sensitive Issues. *Experimental Psychology, 61*, 48-54. doi:10.1027/1618-3169/a000226

Moshagen, M., Hilbig, B. E., & Musch, J. (2011). Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *European Journal of Social Psychology, 41*, 638-644. doi:10.1002/Ejsp.793

Moshagen, M., & Musch, J. (2012). Surveying Multiple Sensitive Attributes using an Extension of the Randomized-Response Technique. *International Journal of Public Opinion Research, 24*, 508-523.

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44*, 222-231. doi:10.3758/s13428-011-0144-2 21858604

Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2010). Reducing Socially Desirable Responses in Epidemiologic Surveys. An Extension of the Randomized-response Technique. *Epidemiology, 21*, 379-382. doi:10.1097/Ede.0b013e3181d61dbc

Musch, J., Bröder, A., & Klauer, K. C. (2001). Improving Survey Research on the World-Wide Web using the Randomized Response Technique. In U. D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 179-192). Lengerich, Germany: Pabst.

Ostapczuk, M., Moshagen, M., Zhao, Z., & Musch, J. (2009). Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. *Journal of Educational and Behavioral Statistics, 34*, 267-287. doi:10.3102/1076998609332747

Ostapczuk, M., & Musch, J. (2011). Estimating the prevalence of negative attitudes towards people with disability: A comparison of direct questioning, projective questioning and randomised response. *Disability and Rehabilitation, 33*, 1-13. doi:10.3109/09638288.2010.492067

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology, 39*, 920-931. doi:10.1002/ejsp.588

Ostapczuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. *Statistical Methods in Medical Research, 20*, 489-503. doi:10.1177/0962280210372843

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes, Vol. 1* (pp. 17-59). San Diego, CA: Academic Press.

Phillips, D. L., & Clancy, K. J. (1972). Some Effects of Social Desirability in Survey Studies. *American Journal of Sociology, 77*, 921-940. doi:10.1086/225231

Pitsch, W., & Emrich, E. (2011). The frequency of doping in elite sport: Results of a replication study. *International Review for the Sociology of Sport, 47*, 559-580. doi:10.1177/1012690211413969

Pitsch, W., Emrich, E., & Klein, M. (2007). Doping in elite sports in Germany: results of a www survey. *European Journal of Sport and Society, 4*, 89-102.

Pötzsch, O. (2012). Geburten in Deutschland [Births in Germany]. Retrieved Jun 6, 2012, from German Federal Statistical Office

https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/Bevoelkerungsbewegung/BroschuereGeburtenDeutschland0120007129004.pdf

Schröter, H., Studzinski, B., Dietz, P., Ulrich, R., Striegel, H., & Simon, P. (2016). A Comparison of the Cheater Detection and the Unrelated Question Models: A Randomized Response Survey on Physical and Cognitive Doping in Recreational Triathletes. *Plos One, 11*(5), 1-11. doi:10.1371/journal.pone.0155765

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883. doi:10.1037/0033-2909.133.5.859 17723033

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking Sensitive Questions: A Statistical Power Analysis of Randomized Response Models. *Psychological Methods, 17*, 623-641. doi:10.1037/A0029314

Umesh, U. N., & Peterson, R. A. (1991). A Critical Evaluation of the Randomized-Response Method - Applications, Validation, and Research Agenda. *Sociological Methods & Research, 20*, 104-138. doi:10.1177/0049124191020001004

Warner, S. L. (1965). Randomized-Response - a Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*, 63-69.

Wolter, F., & Preisendörfer, P. (2013). Asking Sensitive Questions: An Evaluation of the Randomized Response Technique Versus Direct Questioning Using Individual Validation Data. *Sociological Methods & Research, 42*, 321-353. doi:10.1177/0049124113500474

Wu, Q., & Tang, M.-L. (2016). Non-randomized response model for sensitive survey with noncompliance. *Statistical Methods in Medical Research, 25*, 2827-2839. doi:10.1177/0962280214533022

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika, 67*, 251-263. doi:10.1007/s00184-007-0131-x

Tables

Table 1

Demographics by questioning technique.

	DQ (%)	CDM (%)	CDTRM (%)	
Gender				
female	54.30	59.46	57.73	$\chi^2(2) = 4.05$, $p = .132$, *Cramer's V* = .04
male	45.70	40.54	42.27	
Age (years)				
18-25	8.60	11.35	9.03	
26-35	20.61	23.60	21.63	
36-45	21.33	21.44	22.07	$\chi^2(10) = 11.51$, $p = .319$, *Cramer's V* = .05
46-55	23.30	22.43	23.95	
56-65	22.04	17.21	18.95	
> 65	4.12	3.96	4.38	
Educational achievement				
No school leaving certificate	0.18	0.18	0.36	
Lower secondary school leaving certificate	8.24	6.04	6.17	
Secondary school leaving certificate	19.89	16.31	16.35	
Subject-specific university entrance qualification	6.81	6.31	6.61	$\chi^2(16) = 15.32$, $p = .502$, *Cramer's V* = .05
Higher education entrance qualification	9.32	11.71	12.51	
Completed vocational training	30.47	31.17	30.47	
Bachelor's degree	8.06	7.66	7.95	
Master's degree	16.49	19.37	18.14	
PhD	0.54	1.26	1.43	

Note. DQ = Direct Questioning, CDM = Cheating Detection Model. CDTRM = Cheating Detection Triangular Model.

Table 2

Parameter estimates (standard errors in parentheses) for the prevalence of the sensitive attribute in the total sample (N = 2787).

Prevalence Estimates	DQ	CDM	CDTRM
$\hat{\pi}$ (honest carriers)	15.59	29.70	25.07
	(1.54)	(2.58)	(2.51)
$\hat{\beta}$ (honest non-carriers)	-	61.06	61.65
	-	(3.89)	(3.93)
$\hat{\gamma}$ (non-adherent respondents)	-	9.24	13.28
	-	(2.10)	(2.25)
$\hat{\pi} + \hat{\gamma}$ (upper bound for carriers)	-	38.94	38.36
	-	(3.89)	(3.93)
Sensitivity			
Lower bound	25.47	-	-
	(2.43)	-	-
	-	39.59	36.18
	-	(3.48)	(3.42)
Upper bound	-	52.00	51.34
	-	(5.16)	(5.23)
Specificity			
Lower bound	97.88	-	-
	(0.94)	-	-
	-	79.13	79.46
	-	(5.74)	(5.75)
Upper bound	-	83.69	90.19
	-	(3.68)	(3.40)

Note. DQ = Direct Questioning, CDM = Cheating Detection Model. CDTRM = Cheating Detection Triangular Model. TRM = Triangular Model. In the CDM and CDTRM conditions, the prevalence estimates for π form the lower bound and $\pi+\gamma$ the upper bound of the prevalence of the sensitive attribute.

Table 3

Parameter comparisons for the prevalence of the sensitive attribute in the total sample (N = 2787).

Parameter 1 = Parameter 2	Parameter 1 (in %)	Parameter 2 (in %)	|*Difference*| (in %)	Model fit ΔG^2 (df = 1)	p
$\hat{\pi}_{DQ} = \pi_{DQ}$	15.59	57.71	42.12	423.32	< .001*
$\hat{\pi}_{CDM} = \pi_{CDM}$	29.70	59.73	30.03	131.94	< .001*
$\hat{\pi}_{CDTRM} = \pi_{DQ}$	25.07	59.25	34.18	172.27	< .001*
$(\hat{\pi}_{CDM} + \hat{\gamma}_{CDM}) = \pi_{CDM}$	38.94	59.73	20.79	26.53	< .001*
$(\hat{\pi}_{CDTRM} + \hat{\gamma}_{CDTRM}) = \pi_{CDTRM}$	38.36	59.25	20.89	26.11	< .001*
$\hat{\pi}_{DQ} = \hat{\pi}_{CDM}$	15.59	29.70	14.11	22.17	< .001*
$\hat{\pi}_{DQ} = \hat{\pi}_{CDTRM}$	15.59	25.07	9.48	10.52	= .001*
$\hat{\pi}_{CDM} = \hat{\pi}_{CDTRM}$	29.70	25.07	4.63	1.65	= .199
$\hat{\pi}_{DQ} = (\hat{\pi}_{CDM} + \hat{\gamma}_{CDM})$	15.59	38.94	23.35	33.38	< .001*
$\hat{\pi}_{DQ} = (\hat{\pi}_{CDTRM} + \hat{\gamma}_{CDTRM})$	15.59	38.36	22.77	31.29	< .001*
$(\hat{\pi}_{CDM} + \hat{\gamma}_{CDM}) = (\hat{\pi}_{CDTRM} + \hat{\gamma}_{CDTRM})$	38.94	38.36	0.58	0.01	= .916
$\hat{\gamma}_{CDM} = 0$ %	9.24	0.00	9.24	24.28	< .001*
$\hat{\gamma}_{CDTRM} = 0$ %	13.28	0.00	13.28	45.79	< .001*
$\hat{\gamma}_{CDM} = \hat{\gamma}_{CDTRM}$	9.24	13.28	4.04	1.73	= .188
$\hat{\beta}_{CDM} = \hat{\beta}_{CDTRM}$	61.06	61.65	0.59	0.01	= .916

Note. DQ = Direct Questioning, CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model. In the DQ condition, $\hat{\pi}$ reflects the estimated proportion of carriers of the sensitive attribute. In the CDM and CDTRM conditions, $\hat{\pi}$ reflects the estimated proportion of honest carriers of the sensitive attribute, $\hat{\beta}$ reflects the estimated proportion of honest non-carriers, and $\hat{\gamma}$ (= 1 - $\hat{\pi}$ - $\hat{\beta}$) reflects the estimated proportion of respondents not adhering to the instructions. Consequently, in the CDM and CDTRM conditions, $\hat{\pi}$ reflects the lower bound for the proportion of

carriers and $\hat{\pi} + \hat{\gamma}$ reflects the upper bound. Parameter π represents the known prevalence of the sensitive attribute in the respective conditions.
*$p < .05$

Table 4

Comparisons of sensitivity and specificity by questioning technique.

	Parameter 1 (in %)	Parameter 2 (in %)	|*Difference*| (in %)	Model fit ΔG^2 (*df* = 1)	*p*
Sensitivity					
sen_{DQ} = 100%	25.47	100.00	74.53	8476.52	< .001*
$low\_b\_sen_{CDM}$ = 100 %	39.59	100.00	60.41	7946.30	< .001*
$low\_b\_sen_{CDTRM}$ = 100 %	36.18	100.00	63.82	8769.51	< .001*
$up\_b\_sen_{CDM}$ = 100 %	52.00	100.00	48.00	79.38	< .001*
$up\_b\_sen_{CDTRM}$ = 100 %	51.34	100.00	48.66	78.78	< .001*
sen_{DQ} = $low\_b\_sen_{CDM}$	25.47	39.59	14.12	10.94	< .001*
sen_{DQ} = $low\_b\_sen_{CDTRM}$	25.47	36.18	10.71	6.48	= .011*
$low\_b\_sen_{CDM}$ = $low\_b\_sen_{CDTRM}$	39.59	36.18	3.41	0.49	= .485
sen_{DQ} = $up\_b\_sen_{CDM}$	25.47	52.00	26.53	22.55	< .001*
sen_{DQ} = $up\_b\_sen_{CDTRM}$	25.47	51.34	25.87	21.14	< .001*
$up\_b\_sen_{CDM}$ = $up\_b\_sen_{CDTRM}$	52.00	51.34	0.66	0.01	= .928
Specificity					
$spec_{DQ}$ = 100 %	97.88	100.00	2.12	135.77	< .001*
$low\_b\_spec_{CDM}$ = 100 %	79.13	100.00	20.87	25.55	< .001*
$low\_b\_spec_{CDTRM}$ = 100 %	79.46	100.00	20.54	15.24	< .001*
$up\_b\_spec_{CDM}$ = 100 %	83.69	100.00	16.31	25.55	< .001*
$up\_b\_spec_{CDTRM}$ = 100 %	90.19	100.00	9.81	10.14	= .001*
$spec_{DQ}$ = $low\_b\_spec_{CDM}$	97.88	79.13	18.75	15.92	< .001*

$spec_{DQ}$ = $low\_b\_spec_{CDTRM}$	97.88	79.46	18.42	11.46	< .001*
$low\_b\_spec_{CDM}$ = $low\_b\_spec_{CDTRM}$	79.13	79.46	0.33	0.00	= .968
$spec_{DQ}$ = $up\_b\_spec_{CDM}$	97.88	83.69	14.19	15.91	= .001*
$spec_{DQ}$ = $up\_b\_spec_{CDTRM}$	97.88	90.19	7.69	5.26	= .022*
$up\_b\_spec_{CDM}$ = $up\_b\_spec_{CDTRM}$	83.69	90.19	6.50	1.68	= .195

Note. sen = sensitivity, spec = specificity; DQ = Direct Questioning, CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model. In the CDM and CDTRM conditions, low_b refers to the lower bounds for sensitivity and specificity, and up_b refers to the upper bounds for sensitivity and specificity.
* $p < .05$.

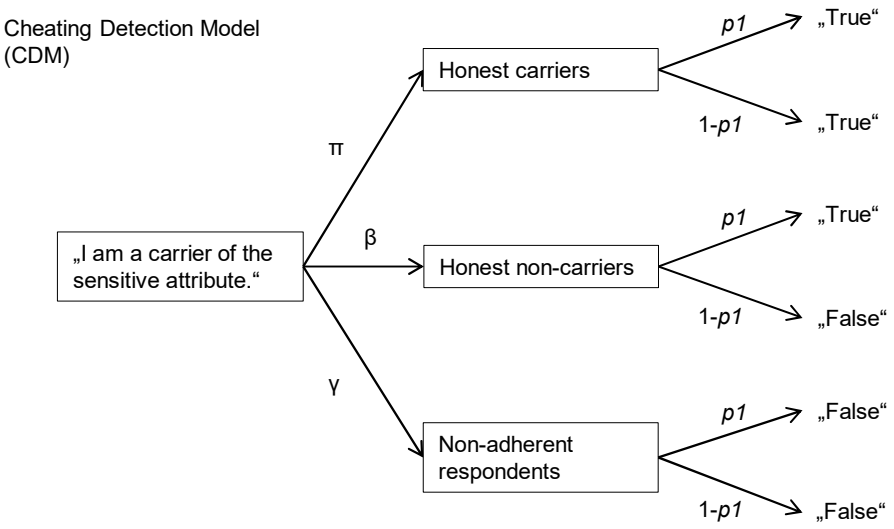
Table 5

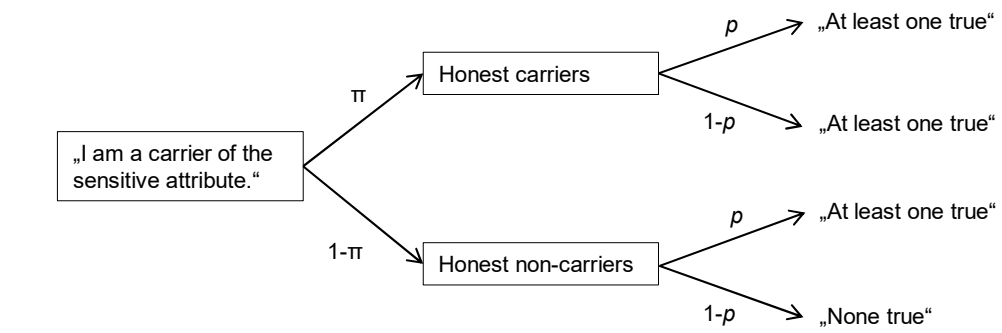
ANOVAs, means and standard deviations for subjective evaluations by questioning technique.

Item	DQ	CDM	CDTRM	$F(2, 2784)$	p	η_p
The question was comprehensible.	6.56* (0.98)	4.21* (1.94)	5.00* (1.85)	334.79	< .001	.19
The question guaranteed the confidentiality of my response.	6.05* (1.42)	5.05* (1.76)	5.49* (1.61)	71.32	< .001	.05
The way the question was asked was interesting.	5.72* (1.44)	5.02* (1.84)	5.36* (1.67)	32.60	< .001	.02
The way the question was asked was reasonable.	5.86* (1.38)	4.17* (1.79)	4.80* (1.69)	188.93	< .001	.12
The question was cumbersome to answer. (R)	5.64* (1.83)	3.12* (1.86)	3.74* (1.97)	331.86	< .001	.19
I carefully read and followed all instructions.	6.45* (1.08)	6.08 (1.25)	6.07 (1.29)	20.342	< .001	.01
I clearly knew which answer to pick.	6.16* (1.30)	4.38* (1.85)	4.99* (1.76)	199.85	< .001	.13
I felt overwhelmed by the question. (R)	6.12* (1.54)	4.36* (1.8)	4.92* (1.79)	188.21	< .001	.12
I just randomly ticked one of the answers. (R)	6.69* (1.09)	6.36* (1.37)	6.53* (1.18)	14.35	< .001	.01
Mean of the scale (*Cronbach* α = .84)	6.14* (.80)	4.75* (1.08)	5.21* (1.08)	336.49	< .001	.20

Note. DQ = Direct Questioning, CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model. All variables were assessed on a 7-point Likert-type scale with higher values indicating more favorable evaluations. Questions marked with an (R) have negative polarity and were reverse-coded prior to analysis to facilitate their interpretability and the computation of a scale mean.
* Bonferroni-corrected post-hoc tests revealed that these conditions significantly differed from all other conditions (all $p < .05$).

Figures


Cheating Detection Model (CDM)
„I am a carrier of the sensitive attribute.“
π
β
γ
Honest carriers
Honest non-carriers
Non-adherent respondents
p1
1-p1
„True“
„False“


Triangular Model (TRM)
„I am a carrier of the sensitive attribute.“
π
1-π
Honest carriers
Honest non-carriers
p
1-p
„At least one true“
„None true“

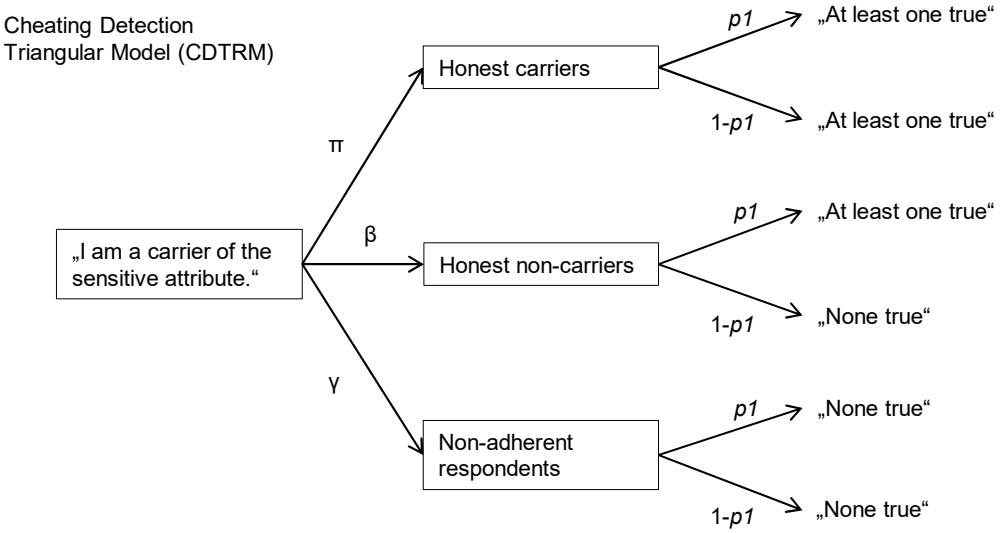

Cheating Detection Triangular Model (CDTRM)
„I am a carrier of the sensitive attribute.“
π
β
γ
Honest carriers
Honest non-carriers
Non-adherent respondents
p1
1-p1
„At least one true“
„None true“

Figure 1. Tree models of the Cheating Detection Model (CDM) for the group with randomization probability *p1*, the Triangular Model (TRM), and the Cheating Detection Triangular Model (CDTRM) for the group with randomization probability *p1*. In the CDM, respondents are instructed to answer 'yes' irrespective of their true status with probability *p1*, and to answer honestly with probability 1-*p1*. In the (CD)TRM, respondents are asked to provide a joint answer to a sensitive statement with unknown prevalence π and a non-sensitive statement with known randomization probability *p1*. For the second group in both the CDM and the CDTRM, parameter *p1* is simply replaced with *p2*.

Electronic supplementary information: Appendix A

MultiTree equations for the estimation of π (Pi), β (Beta), γ (Gamma), sensitivity (sen) and specificity (spec) in a multinomial model. In the CDM and CDTRM conditions, the prevalence estimate for π provides a lower bound for the prevalence of the sensitive attribute, and the estimate for π + γ provides an upper bound (upb). Likewise, lower bounds (lowb) and upper bounds (upb) are also estimated for sensitivity and specificity in the CDM and CDTRM conditions. The parameter p1 denotes the known probability of being born in November or December (p1 = .158) and the parameter p2 denotes the known probability of being born between January and October (p2 = .842, Pötzsch, 2012). Parameters containing 'cond' in their name are auxiliary parameters used for technical reasons that are not relevant for our research questions and are therefore not reported in the manuscript. CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model, DQ = direct questioning.

Equations for the prevalence estimates in the total sample:

```
TRM_p1          TRM_p1_at_least_one_true        1_Pi1_TRM
TRM_p1          TRM_p1_at_least_one_true        (1-1_Pi1_TRM)*p1
TRM_p1          TRM_p1_none_true                (1-1_Pi1_TRM)*(1-p1)
TRM_p2          TRM_p2_at_least_one_true        1_Pi2_TRM
TRM_p2          TRM_p2_at_least_one_true        (1-1_Pi2_TRM)*p2
TRM_p2          TRM_p2_none_true                (1-1_Pi2_TRM)*(1-p2)
1_CDM_p1        1_CDM_p1_true                   2_Pi_CDM
1_CDM_p1        1_CDM_p1_true                   (1-2_Pi_CDM)*Beta_cond_Pi_CDM*p1
1_CDM_p1        1_CDM_p1_false                  (1-2_Pi_CDM)*Beta_cond_Pi_CDM*(1-p1)
1_CDM_p1        1_CDM_p1_false                  (1-2_Pi_CDM)*(1-Beta_cond_Pi_CDM)
1_CDM_p2        1_CDM_p2_true                   2_Pi_CDM
1_CDM_p2        1_CDM_p2_true                   (1-2_Pi_CDM)*Beta_cond_Pi_CDM*p2
1_CDM_p2        1_CDM_p2_false                  (1-2_Pi_CDM)*Beta_cond_Pi_CDM*(1-p2)
1_CDM_p2        1_CDM_p2_false                  (1-2_Pi_CDM)*(1-Beta_cond_Pi_CDM)
2_CDM_p1        2_CDM_p1_true                   2_Beta_CDM*p1
2_CDM_p1        2_CDM_p1_false                  2_Beta_CDM*(1-p1)
```

2_CDM_p1	2_CDM_p1_true	(1-2_Beta_CDM)*Pi_cond_Beta_CDM
2_CDM_p1	2_CDM_p1_false	(1-2_Beta_CDM)*(1-Pi_cond_Beta_CDM)
2_CDM_p2	2_CDM_p2_true	2_Beta_CDM*p2
2_CDM_p2	2_CDM_p2_false	2_Beta_CDM*(1-p2)
2_CDM_p2	2_CDM_p2_true	(1-2_Beta_CDM)*Pi_cond_Beta_CDM
2_CDM_p2	2_CDM_p2_false	(1-2_Beta_CDM)*(1-Pi_cond_Beta_CDM)
3_CDM_p1	3_CDM_p1_false	2_Gamma_CDM
3_CDM_p1	3_CDM_p1_true	(1-2_Gamma_CDM)*Pi_cond_Gamma_CDM
3_CDM_p1	3_CDM_p1_true	(1-2_Gamma_CDM)*(1-Pi_cond_Gamma_CDM)*p1
3_CDM_p1	3_CDM_p1_false	(1-2_Gamma_CDM)*(1-Pi_cond_Gamma_CDM)*(1-p1)
3_CDM_p2	3_CDM_p2_false	2_Gamma_CDM
3_CDM_p2	3_CDM_p2_true	(1-2_Gamma_CDM)*Pi_cond_Gamma_CDM
3_CDM_p2	3_CDM_p2_true	(1-2_Gamma_CDM)*(1-Pi_cond_Gamma_CDM)*p2
3_CDM_p2	3_CDM_p2_false	(1-2_Gamma_CDM)*(1-Pi_cond_Gamma_CDM)*(1-p2)
4_CDM_p1	4_CDM_p1_true	(1-2_upb_CDM)*p1
4_CDM_p1	4_CDM_p1_false	(1-2_upb_CDM)*(1-p1)
4_CDM_p1	4_CDM_p1_true	2_upb_CDM*Pi_cond_Beta_CDM
4_CDM_p1	4_CDM_p1_false	2_upb_CDM*(1-Pi_cond_Beta_CDM)
4_CDM_p2	4_CDM_p2_true	(1-2_upb_CDM)*p2
4_CDM_p2	4_CDM_p2_false	(1-2_upb_CDM)*(1-p2)
4_CDM_p2	4_CDM_p2_true	2_upb_CDM*Pi_cond_Beta_CDM
4_CDM_p2	4_CDM_p2_false	2_upb_CDM*(1-Pi_cond_Beta_CDM)
1_CDTRM_p1	1_CDTRM_p1_at_least_one_true	3_Pi_CDTRM
1_CDTRM_p1	1_CDTRM_p1_at_least_one_true	(1-3_Pi_CDTRM)*Beta_cond_Pi_CDTRM*p1
1_CDTRM_p1	1_CDTRM_p1_none_true	(1-3_Pi_CDTRM)*Beta_cond_Pi_CDTRM*(1-p1)
1_CDTRM_p1	1_CDTRM_p1_none_true	(1-3_Pi_CDTRM)*(1-Beta_cond_Pi_CDTRM)
1_CDTRM_p2	1_CDTRM_p2_at_least_one_true	3_Pi_CDTRM
1_CDTRM_p2	1_CDTRM_p2_at_least_one_true	(1-3_Pi_CDTRM)*Beta_cond_Pi_CDTRM*p2
1_CDTRM_p2	1_CDTRM_p2_none_true	(1-3_Pi_CDTRM)*Beta_cond_Pi_CDTRM*(1-p2)
1_CDTRM_p2	1_CDTRM_p2_none_true	(1-3_Pi_CDTRM)*(1-Beta_cond_Pi_CDTRM)
2_CDTRM_p1	2_CDTRM_p1_at_least_one_true	3_Beta_CDTRM*p1
2_CDTRM_p1	2_CDTRM_p1_none_true	3_Beta_CDTRM*(1-p1)
2_CDTRM_p1	2_CDTRM_p1_at_least_one_true	(1-3_Beta_CDTRM)*Pi_cond_Beta_CDTRM
2_CDTRM_p1	2_CDTRM_p1_none_true	(1-3_Beta_CDTRM)*(1-Pi_cond_Beta_CDTRM)
2_CDTRM_p2	2_CDTRM_p2_at_least_one_true	3_Beta_CDTRM*p2
2_CDTRM_p2	2_CDTRM_p2_none_true	3_Beta_CDTRM*(1-p2)
2_CDTRM_p2	2_CDTRM_p2_at_least_one_true	(1-3_Beta_CDTRM)*Pi_cond_Beta_CDTRM
2_CDTRM_p2	2_CDTRM_p2_none_true	(1-3_Beta_CDTRM)*(1-Pi_cond_Beta_CDTRM)
3_CDTRM_p1	3_CDTRM_p1_none_true	3_Gamma_CDTRM
3_CDTRM_p1	3_CDTRM_p1_at_least_one_true	(1-3_Gamma_CDTRM)*Pi_cond_Gamma_CDTRM
3_CDTRM_p1	3_CDTRM_p1_at_least_one_true	(1-3_Gamma_CDTRM)*(1-Pi_cond_Gamma_CDTRM)*p1

THE CHEATING DETECTION TRIANGULAR MODEL

3_CDTRM_p1	3_CDTRM_p1_none_true	(1-3_Gamma_CDTRM)*(1-Pi_cond_Gamma_CDTRM)*(1-p1)
3_CDTRM_p2	3_CDTRM_p2_none_true	3_Gamma_CDTRM
3_CDTRM_p2	3_CDTRM_p2_at_least_one_true	(1-3_Gamma_CDTRM)*Pi_cond_Gamma_CDTRM
3_CDTRM_p2	3_CDTRM_p2_at_least_one_true	(1-3_Gamma_CDTRM)*(1-Pi_cond_Gamma_CDTRM)*p2
3_CDTRM_p2	3_CDTRM_p2_none_true	(1-3_Gamma_CDTRM)*(1-Pi_cond_Gamma_CDTRM)*(1-p2)
4_CDTRM_p1	4_CDTRM_p1_at_least_one_true	(1-3_upb_CDTRM)*p1
4_CDTRM_p1	4_CDTRM_p1_none_true	(1-3_upb_CDTRM)*(1-p1)
4_CDTRM_p1	4_CDTRM_p1_at_least_one_true	3_upb_CDTRM*Pi_cond_Beta_CDTRM
4_CDTRM_p1	4_CDTRM_p1_none_true	3_upb_CDTRM*(1-Pi_cond_Beta_CDTRM)
4_CDTRM_p2	4_CDTRM_p2_at_least_one_true	(1-3_upb_CDTRM)*p2
4_CDTRM_p2	4_CDTRM_p2_none_true	(1-3_upb_CDTRM)*(1-p2)
4_CDTRM_p2	4_CDTRM_p2_at_least_one_true	3_upb_CDTRM*Pi_cond_Beta_CDTRM
4_CDTRM_p2	4_CDTRM_p2_none_true	3_upb_CDTRM*(1-Pi_cond_Beta_CDTRM)
DQ	DQ_true	4_Pi_DQ
DQ	DQ_false	(1-4_Pi_DQ)

Equations for the estimation of sensitivity among carriers:

1_CDM_p1	1_CDM_p1_true	1_lowb_sen_CDM
1_CDM_p1	1_CDM_p1_true	(1-1_lowb_sen_CDM)*Beta_cond_Pi_CDM*p1
1_CDM_p1	1_CDM_p1_false	(1-1_lowb_sen_CDM)*Beta_cond_Pi_CDM*(1-p1)
1_CDM_p1	1_CDM_p1_false	(1-1_lowb_sen_CDM)*(1-Beta_cond_Pi_CDM)
1_CDM_p2	1_CDM_p2_true	1_lowb_sen_CDM
1_CDM_p2	1_CDM_p2_true	(1-1_lowb_sen_CDM)*Beta_cond_Pi_CDM*p2
1_CDM_p2	1_CDM_p2_false	(1-1_lowb_sen_CDM)*Beta_cond_Pi_CDM*(1-p2)
1_CDM_p2	1_CDM_p2_false	(1-1_lowb_sen_CDM)*(1-Beta_cond_Pi_CDM)
2_CDM_p1	2_CDM_p1_true	(1-1_upb_sen_CDM)*p1
2_CDM_p1	2_CDM_p1_false	(1-1_upb_sen_CDM)*(1-p1)
2_CDM_p1	2_CDM_p1_true	1_upb_sen_CDM*Pi_cond_Beta_CDM
2_CDM_p1	2_CDM_p1_false	1_upb_sen_CDM*(1-Pi_cond_Beta_CDM)
2_CDM_p2	2_CDM_p2_true	(1-1_upb_sen_CDM)*p2
2_CDM_p2	2_CDM_p2_false	(1-1_upb_sen_CDM)*(1-p2)
2_CDM_p2	2_CDM_p2_true	1_upb_sen_CDM*Pi_cond_Beta_CDM
2_CDM_p2	2_CDM_p2_false	1_upb_sen_CDM*(1-Pi_cond_Beta_CDM)
1_CDTRM_p1	1_CDTRM_p1_at_least_one_true	2_lowb_sen_CDTRM
1_CDTRM_p1	1_CDTRM_p1_at_least_one_true	(1-2_lowb_sen_CDTRM)*Beta_cond_Pi_CDTRM*p1
1_CDTRM_p1	1_CDTRM_p1_none_true	(1-2_lowb_sen_CDTRM)*Beta_cond_Pi_CDTRM*(1-p1)
1_CDTRM_p1	1_CDTRM_p1_none_true	(1-2_lowb_sen_CDTRM)*(1-Beta_cond_Pi_CDTRM)

THE CHEATING DETECTION TRIANGULAR MODEL

```
1_CDTRM_p2      1_CDTRM_p2_at_least_one_true        2_lowb_sen_CDTRM
1_CDTRM_p2      1_CDTRM_p2_at_least_one_true        (1-2_lowb_sen_CDTRM)*Beta_cond_Pi_CDTRM*p2
1_CDTRM_p2      1_CDTRM_p2_none_true                (1-2_lowb_sen_CDTRM)*Beta_cond_Pi_CDTRM*(1-p2)
1_CDTRM_p2      1_CDTRM_p2_none_true                (1-2_lowb_sen_CDTRM)*(1-Beta_cond_Pi_CDTRM)
2_CDTRM_p1      2_CDTRM_p1_at_least_one_true        (1-2_upb_sen_CDTRM)*p1
2_CDTRM_p1      2_CDTRM_p1_none_true                (1-2_upb_sen_CDTRM)*(1-p1)
2_CDTRM_p1      2_CDTRM_p1_at_least_one_true        2_upb_sen_CDTRM*Pi_cond_Beta_CDTRM
2_CDTRM_p1      2_CDTRM_p1_none_true                2_upb_sen_CDTRM*(1-Pi_cond_Beta_CDTRM)
2_CDTRM_p2      2_CDTRM_p2_at_least_one_true        (1-2_upb_sen_CDTRM)*p2
2_CDTRM_p2      2_CDTRM_p2_none_true                (1-2_upb_sen_CDTRM)*(1-p2)
2_CDTRM_p2      2_CDTRM_p2_at_least_one_true        2_upb_sen_CDTRM*Pi_cond_Beta_CDTRM
2_CDTRM_p2      2_CDTRM_p2_none_true                2_upb_sen_CDTRM*(1-Pi_cond_Beta_CDTRM)
3_DQ            3_DQ_true                           3_sen_DQ
3_DQ            3_DQ_false                          (1-3_sen_DQ)
```

Equations for the estimation of specificity among non-carriers:

```
1_CDM_p1        1_CDM_p1_true                       (1-1_upb_spec_CDM)
1_CDM_p1        1_CDM_p1_true                       1_upb_spec_CDM*Beta_cond_Pi_CDM*p1
1_CDM_p1        1_CDM_p1_false                      1_upb_spec_CDM*Beta_cond_Pi_CDM*(1-p1)
1_CDM_p1        1_CDM_p1_false                      1_upb_spec_CDM*(1-Beta_cond_Pi_CDM)
1_CDM_p2        1_CDM_p2_true                       (1-1_upb_spec_CDM)
1_CDM_p2        1_CDM_p2_true                       1_upb_spec_CDM*Beta_cond_Pi_CDM*p2
1_CDM_p2        1_CDM_p2_false                      1_upb_spec_CDM*Beta_cond_Pi_CDM*(1-p2)
1_CDM_p2        1_CDM_p2_false                      1_upb_spec_CDM*(1-Beta_cond_Pi_CDM)
2_CDM_p1        2_CDM_p1_true                       1_lowb_spec_CDM*p1
2_CDM_p1        2_CDM_p1_false                      1_lowb_spec_CDM*(1-p1)
2_CDM_p1        2_CDM_p1_true                       (1-1_lowb_spec_CDM)*Pi_cond_Beta_CDM
2_CDM_p1        2_CDM_p1_false                      (1-1_lowb_spec_CDM)*(1-Pi_cond_Beta_CDM)
2_CDM_p2        2_CDM_p2_true                       1_lowb_spec_CDM*p2
2_CDM_p2        2_CDM_p2_false                      1_lowb_spec_CDM*(1-p2)
2_CDM_p2        2_CDM_p2_true                       (1-1_lowb_spec_CDM)*Pi_cond_Beta_CDM
2_CDM_p2        2_CDM_p2_false                      (1-1_lowb_spec_CDM)*(1-Pi_cond_Beta_CDM)
1_CDTRM_p1      1_CDTRM_p1_at_least_one_true        (1-2_upb_spec_CDTRM)
1_CDTRM_p1      1_CDTRM_p1_at_least_one_true        2_upb_spec_CDTRM*Beta_cond_Pi_CDTRM*p1
1_CDTRM_p1      1_CDTRM_p1_none_true                2_upb_spec_CDTRM*Beta_cond_Pi_CDTRM*(1-p1)
1_CDTRM_p1      1_CDTRM_p1_none_true                2_upb_spec_CDTRM*(1-Beta_cond_Pi_CDTRM)
1_CDTRM_p2      1_CDTRM_p2_at_least_one_true        (1-2_upb_spec_CDTRM)
1_CDTRM_p2      1_CDTRM_p2_at_least_one_true        2_upb_spec_CDTRM*Beta_cond_Pi_CDTRM*p2
```

THE CHEATING DETECTION TRIANGULAR MODEL

1_CDTRM_p2	1_CDTRM_p2_none_true	2_upb_spec_CDTRM*Beta_cond_Pi_CDTRM*(1-p2)
1_CDTRM_p2	1_CDTRM_p2_none_true	2_upb_spec_CDTRM*(1-Beta_cond_Pi_CDTRM)
2_CDTRM_p1	2_CDTRM_p1_at_least_one_true	2_lowb_spec_CDTRM*p1
2_CDTRM_p1	2_CDTRM_p1_none_true	2_lowb_spec_CDTRM*(1-p1)
2_CDTRM_p1	2_CDTRM_p1_at_least_one_true	(1-2_lowb_spec_CDTRM)*Pi_cond_Beta_CDTRM
2_CDTRM_p1	2_CDTRM_p1_none_true	(1-2_lowb_spec_CDTRM)*(1-Pi_cond_Beta_CDTRM)
2_CDTRM_p2	2_CDTRM_p2_at_least_one_true	2_lowb_spec_CDTRM*p2
2_CDTRM_p2	2_CDTRM_p2_none_true	2_lowb_spec_CDTRM*(1-p2)
2_CDTRM_p2	2_CDTRM_p2_at_least_one_true	(1-2_lowb_spec_CDTRM)*Pi_cond_Beta_CDTRM
2_CDTRM_p2	2_CDTRM_p2_none_true	(1-2_lowb_spec_CDTRM)*(1-Pi_cond_Beta_CDTRM)
3_DQ	3_DQ_true	(1-3_spec_DQ)
3_DQ	3_DQ_false	3_spec_DQ

THE CHEATING DETECTION TRIANGULAR MODEL

Empirically observed answer frequencies for the sensitive attribute used for parameter estimation in multiTree (Moshagen, 2010) in version 0.46. TRM = Triangular Model, CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model, DQ = direct questioning. Note that TRM and CDTRM estimates are based on the same data, but obtained by applying different measurement models. To be able to estimate π, β and γ in separate multinomial trees, answer frequencies for the CDM and the CDTRM had to be entered three times each.

Total Sample

TRM_p1_none_true	367
TRM_p1_at_least_one_true	196
TRM_p2_none_true	128
TRM_p2_at_least_one_true	428
1_CDM_p1_true	218
1_CDM_p1_false	336
1_CDM_p2_true	451
1_CDM_p2_false	105
2_CDM_p1_true	218
2_CDM_p1_false	336
2_CDM_p2_true	451
2_CDM_p2_false	105
3_CDM_p1_true	218
3_CDM_p1_false	336
3_CDM_p2_true	451
3_CDM_p2_false	105
4_CDM_p1_true	218
4_CDM_p1_false	336
4_CDM_p2_true	451
4_CDM_p2_false	105
1_CDTRM_p1_none_true	367
1_CDTRM_p1_at_least_one_true	196
1_CDTRM_p2_none_true	128
1_CDTRM_p2_at_least_one_true	428
2_CDTRM_p1_none_true	367

2_CDTRM_p1_at_least_one_true	196
2_CDTRM_p2_none_true	128
2_CDTRM_p2_at_least_one_true	428
3_CDTRM_p1_none_true	367
3_CDTRM_p1_at_least_one_true	196
3_CDTRM_p2_none_true	128
3_CDTRM_p2_at_least_one_true	428
4_CDTRM_p1_none_true	367
4_CDTRM_p1_at_least_one_true	196
4_CDTRM_p2_none_true	128
4_CDTRM_p2_at_least_one_true	428
DQ_true	87
DQ_false	471

Only carriers (required for estimation of sensitivity)

1_CDM_p1_true	150
1_CDM_p1_false	168
1_CDM_p2_true	276
1_CDM_p2_false	69
2_CDM_p1_true	150
2_CDM_p1_false	168
2_CDM_p2_true	276
2_CDM_p2_false	69
1_CDTRM_p1_none_true	183
1_CDTRM_p1_at_least_one_true	143
1_CDTRM_p2_none_true	77
1_CDTRM_p2_at_least_one_true	260
2_CDTRM_p1_none_true	183
2_CDTRM_p1_at_least_one_true	143
2_CDTRM_p2_none_true	77
2_CDTRM_p2_at_least_one_true	260
3_DQ_true	82
3_DQ_false	240

THE CHEATING DETECTION TRIANGULAR MODEL

Only non-carriers (required for estimation of specificity)

1_CDM_p1_true	68
1_CDM_p1_false	168
1_CDM_p2_true	175
1_CDM_p2_false	36
2_CDM_p1_true	68
2_CDM_p1_false	168
2_CDM_p2_true	175
2_CDM_p2_false	36
1_CDTRM_p1_none_true	184
1_CDTRM_p1_at_least_one_true	53
1_CDTRM_p2_none_true	51
1_CDTRM_p2_at_least_one_true	168
2_CDTRM_p1_none_true	184
2_CDTRM_p1_at_least_one_true	53
2_CDTRM_p2_none_true	51
2_CDTRM_p2_at_least_one_true	168
3_DQ_true	5
3_DQ_false	231

Electronic supplementary information: Appendix B

The following section describes the multinomial processing tree models established to estimate the proportion of carriers of the sensitive attribute given instruction non-adherence. In this description, the CDM with randomization probability *p1* is used as an example, but the procedure can analogously be applied to the CDM with randomization probability *p2* and the CDTRM with randomization probabilities *p1* and *p2*.

In a preliminary step, the tree model for the CDM as depicted in Figure 1 of the main manuscript was converted into binary format (see Figure A1), as is necessary in the multinomial framework. In this binary model, the parameter π_cond denotes the conditional probability of being a carrier of the sensitive attribute given that a respondent honestly adheres to the instructions.

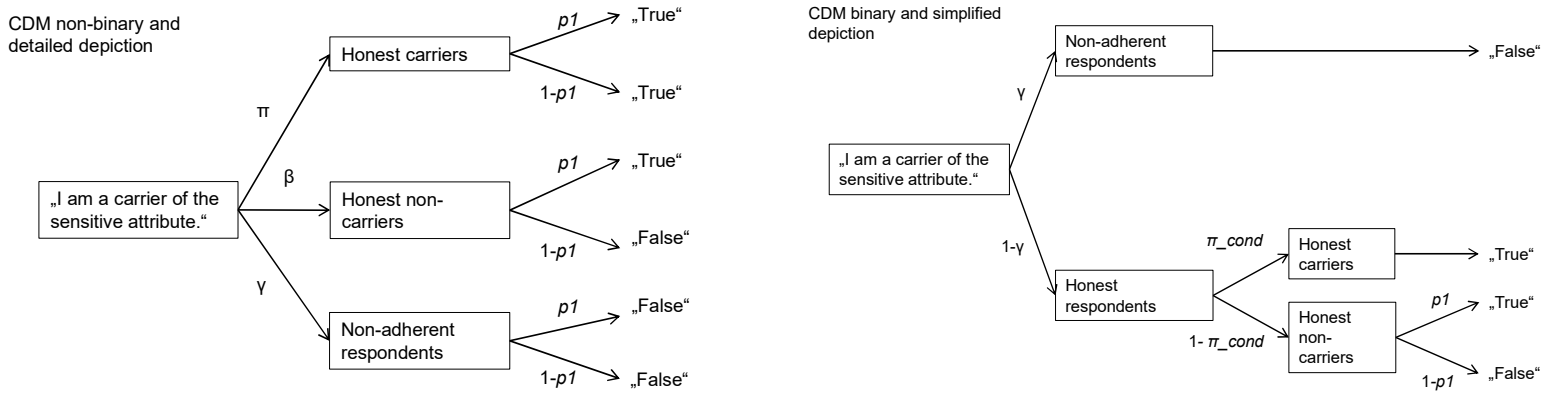


Figure A1. Non-binary model of the Cheating Detection Model (CDM; left panel), and binary reformulation of the CDM (right panel).

To estimate the proportion of carriers of the sensitive attribute given instruction non-adherence, we added the parameter *c_g* (carriers gamma) to the binary model; *c_g* denotes the probability of carrying the sensitive attribute given instruction non-adherence. Empirically observed answer frequencies for the respective answer options ('true' versus 'false') were included in the analysis separately for carriers (c) and non-carriers (nc) of the sensitive attribute, which was possible due to the availability of individual-level data on whether respondents carried the sensitive attribute or not. For a complete representation of all possible response paths, three auxiliary parameters were added to the model: *c_ng* (carriers not gamma) represents the probability of being a carrier given instruction adherence; π _cond_c represents the probability of being identified as a carrier for known carriers who adhered to the instructions; and π _cond_nc represents the probability of being identified as a carrier for known non-carriers who adhered to the instructions. Estimates for these auxiliary parameters are however not relevant, and are therefore not reported. Figure A2 shows the resulting multinomial processing tree model.

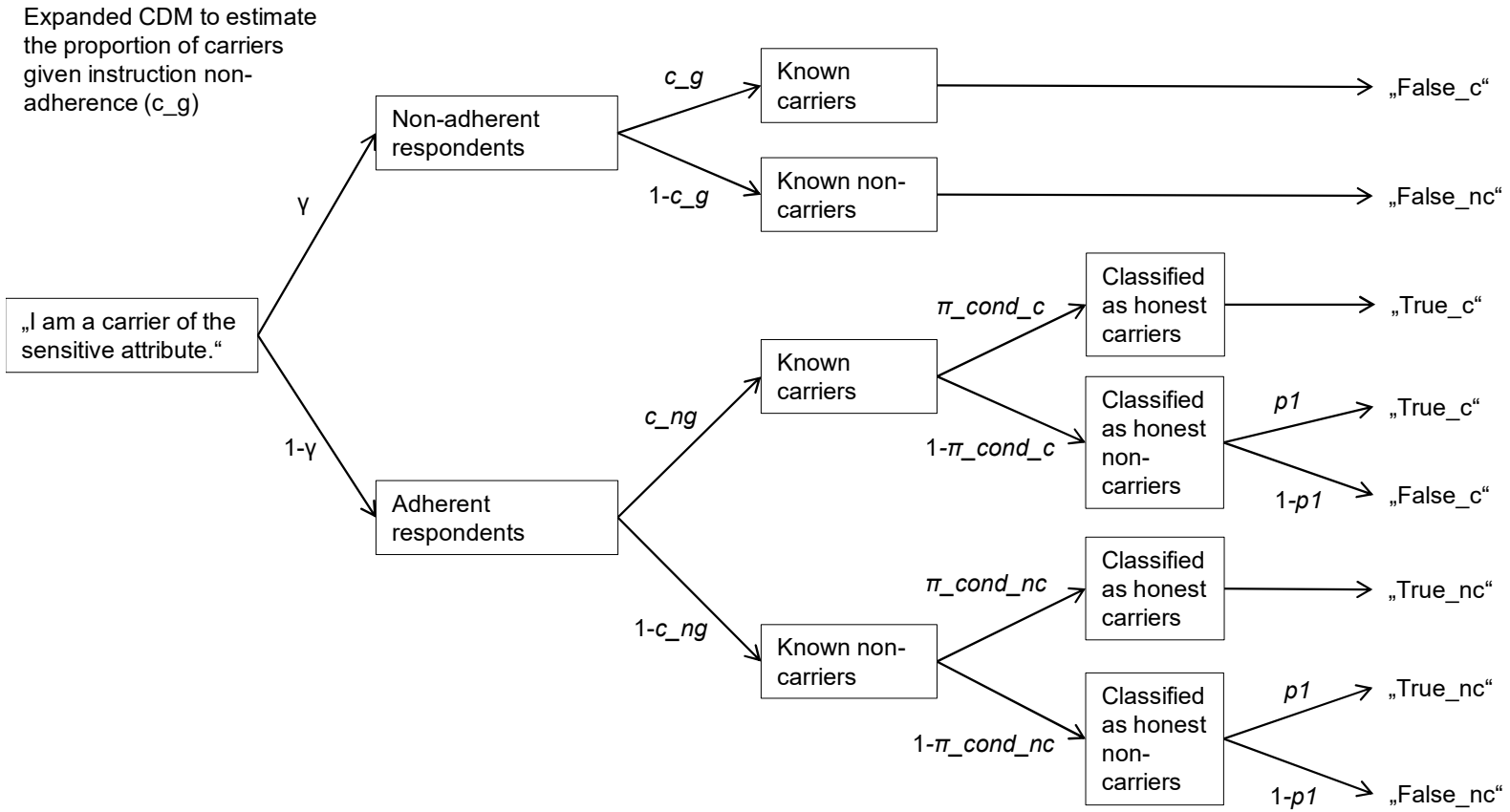


Figure A2. Extended multinomial processing tree model of the Cheating Detection Model (CDM) used to estimate the proportion of carriers given instruction non-adherence (*c_g*).

MultiTree equations used for estimating the proportion of carriers given instruction non-adherence are given below. Parameter γ denotes the probability of instruction non-adherence, parameters *p1* and *p2* denote the known probabilities of being born in November or December (*p1* = .158) or between January and October (*p2* = .842, Pötzsch, 2012), and parameter c_g denotes the probability of being a carrier of the sensitive attribute given instruction non-adherence. CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model.

```
CDM_p1      CDM_p1_false_c                  Gamma_CDM*c_g_CDM
CDM_p1      CDM_p1_false_nc                 Gamma_CDM*(1-c_g_CDM)
CDM_p1      CDM_p1_true_c                   (1-Gamma_CDM)*c_ng_CDM*Pi_cond_c_CDM
CDM_p1      CDM_p1_true_c                   (1-Gamma_CDM)*c_ng_CDM*(1-Pi_cond_c_CDM)*p1
CDM_p1      CDM_p1_false_c                  (1-Gamma_CDM)*c_ng_CDM*(1-Pi_cond_c_CDM)*(1-p1)
CDM_p1      CDM_p1_true_nc                  (1-Gamma_CDM)*(1-c_ng_CDM)*Pi_cond_nc_CDM
CDM_p1      CDM_p1_true_nc                  (1-Gamma_CDM)*(1-c_ng_CDM)*(1-Pi_cond_nc_CDM)*p1
CDM_p1      CDM_p1_false_nc                 (1-Gamma_CDM)*(1-c_ng_CDM)*(1-Pi_cond_nc_CDM)*(1-p1)
CDM_p2      CDM_p2_false_c                  Gamma_CDM*c_g_CDM
CDM_p2      CDM_p2_false_nc                 Gamma_CDM*(1-c_g_CDM)
CDM_p2      CDM_p2_true_c                   (1-Gamma_CDM)*c_ng_CDM*Pi_cond_c_CDM
CDM_p2      CDM_p2_true_c                   (1-Gamma_CDM)*c_ng_CDM*(1-Pi_cond_c_CDM)*p2
CDM_p2      CDM_p2_false_c                  (1-Gamma_CDM)*c_ng_CDM*(1-Pi_cond_c_CDM)*(1-p2)
CDM_p2      CDM_p2_true_nc                  (1-Gamma_CDM)*(1-c_ng_CDM)*Pi_cond_nc_CDM
CDM_p2      CDM_p2_true_nc                  (1-Gamma_CDM)*(1-c_ng_CDM)*(1-Pi_cond_nc_CDM)*p2
CDM_p2      CDM_p2_false_nc                 (1-Gamma_CDM)*(1-c_ng_CDM)*(1-Pi_cond_nc_CDM)*(1-p2)
CDTRM_p1    CDTRM_p1_none_true_c            Gamma_CDTRM*c_g_CDTRM
CDTRM_p1    CDTRM_p1_none_true_nc           Gamma_CDTRM*(1-c_g_CDTRM)
CDTRM_p1    CDTRM_p1_at_least_one_true_c    (1-Gamma_CDTRM)*c_ng_CDTRM*Pi_cond_c_CDTRM
CDTRM_p1    CDTRM_p1_at_least_one_true_c    (1-Gamma_CDTRM)*c_ng_CDTRM*(1-Pi_cond_c_CDTRM)*p1
CDTRM_p1    CDTRM_p1_none_true_c            (1-Gamma_CDTRM)*c_ng_CDTRM*(1-Pi_cond_c_CDTRM)*(1-p1)
CDTRM_p1    CDTRM_p1_at_least_one_true_nc   (1-Gamma_CDTRM)*(1-c_ng_CDTRM)*Pi_cond_nc_CDTRM
CDTRM_p1    CDTRM_p1_at_least_one_true_nc   (1-Gamma_CDTRM)*(1-c_ng_CDTRM)*(1-Pi_cond_nc_CDTRM)*p1
CDTRM_p1    CDTRM_p1_none_true_nc           (1-Gamma_CDTRM)*(1-c_ng_CDTRM)*(1-Pi_cond_nc_CDTRM)*(1-p1)
CDTRM_p2    CDTRM_p2_none_true_c            Gamma_CDTRM*c_g_CDTRM
CDTRM_p2    CDTRM_p2_none_true_nc           Gamma_CDTRM*(1-c_g_CDTRM)
CDTRM_p2    CDTRM_p2_at_least_one_true_c    (1-Gamma_CDTRM)*c_ng_CDTRM*Pi_cond_c_CDTRM
CDTRM_p2    CDTRM_p2_at_least_one_true_c    (1-Gamma_CDTRM)*c_ng_CDTRM*(1-Pi_cond_c_CDTRM)*p2
```

CDTRM_p2	CDTRM_p2_none_true_c	(1-Gamma_CDTRM)*c_ng_CDTRM*(1-Pi_cond_c_CDTRM)*(1-p2)
CDTRM_p2	CDTRM_p2_at_least_one_true_nc	(1-Gamma_CDTRM)*(1-c_ng_CDTRM)*Pi_cond_nc_CDTRM
CDTRM_p2	CDTRM_p2_at_least_one_true_nc	(1-Gamma_CDTRM)*(1-c_ng_CDTRM)*(1-Pi_cond_nc_CDTRM)*p2
CDTRM_p2	CDTRM_p2_none_true_nc	(1-Gamma_CDTRM)*(1-c_ng_CDTRM)*(1-Pi_cond_nc_CDTRM)*(1-p2)

THE CHEATING DETECTION TRIANGULAR MODEL

Empirically observed answer frequencies used for parameter estimation in multiTree (Moshagen, 2010) for estimating the proportion of carriers given instruction non-adherence. CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model.

CDM_p1_true_c	150
CDM_p1_false_c	168
CDM_p1_true_nc	68
CDM_p1_false_nc	168
CDM_p2_true_c	276
CDM_p2_false_c	69
CDM_p2_true_nc	175
CDM_p2_false_nc	36
CDTRM_p1_at_least_one_true_c	143
CDTRM_p1_none_true_c	183
CDTRM_p1_at_least_one_true_nc	53
CDTRM_p1_none_true_nc	184
CDTRM_p2_at_least_one_true_c	260
CDTRM_p2_none_true_c	77
CDTRM_p2_at_least_one_true_nc	168
CDTRM_p2_none_true_nc	51

Electronic supplementary information: Appendix C

Table C1

Parameter estimates (standard errors in parentheses) for the prevalence of the sensitive attribute in the subsample of respondents with high understanding of the indirect questioning techniques. To enable a comparison of estimates for this subsample with estimates in the DQ condition, estimates for the DQ condition are also displayed. Note that in the DQ condition, no comprehension questions were used; therefore, DQ estimates in this table are identical to those in the general sample (Table 2). N = 1009, 36.20% of the total sample.

Prevalence Estimates	DQ (n = 558)	CDM (n = 149)	CDTRM (n = 302)
$\hat{\pi}$ (honest carriers)	15.59	21.94	22.92
	(1.54)	(6.63)	(4.75)
$\hat{\beta}$ (honest non-carriers)	-	77.44	71.35
	-	(9.75)	(7.18)
$\hat{\gamma}$ (non-adherent respondents)	-	0.62	5.73
	-	(5.08)	(3.92)
$\hat{\pi} + \hat{\gamma}$ (upper bound for carriers)	-	22.57	28.66
	-	(9.75)	(7.18)
Sensitivity			
Lower bound	25.47	-	-
	(2.43)	-	-
	-	34.29	39.17
	-	(10.55)	(6.95)
Upper bound	-	37.79	47.28
	-	(14.49)	(10.34)
Specificity			
Lower bound	97.88	-	-
	(0.94)	-	-
	-	86.74	91.28

	-	(11.91)	(9.35)
Upper bound	-	86.74	94.42
	-	(8.42)	(5.78)

Note. DQ = Direct Questioning, CDM = Cheating Detection Model. CDTRM = Cheating Detection Triangular Model. TRM = Triangular Model. In the CDM and CDTRM conditions, the prevalence estimates for π represent the lower bound and $\pi+\gamma$ the upper bound of the prevalence of the sensitive attribute.

Table C2

Parameter comparisons for the prevalence of the sensitive attribute in the subsample of respondents with high understanding of the indirect questioning techniques. To enable a comparison of estimates for this subsample with estimates in the DQ condition, estimates for the DQ condition are also displayed. Note that in the DQ condition, no comprehension questions were used; therefore, DQ estimates in this table are identical to those in the general sample (Table 2). N = 1009, 36.20% of the total sample.

	Parameter 1 (in %)	Parameter 2 (in %)	|*Difference*| (in %)	Model fit ΔG^2 ($df = 1$)	Model fit p
$\hat{\pi}_{DQ} = \pi_{DQ}$	15.59	57.71	42.12	423.32	< .001*
$\hat{\pi}_{CDM} = \pi_{CDM}$	21.94	53.02	31.08	20.02	< .001*
$\hat{\pi}_{CDTRM} = \pi_{DQ}$	22.92	50.66	27.74	31.13	< .001*
$(\hat{\pi}_{CDM} + \hat{\gamma}_{CDM}) = \pi_{CDM}$	22.57	53.02	30.45	8.30	= .004*
$(\hat{\pi}_{CDTRM} + \hat{\gamma}_{CDTRM}) = \pi_{CDTRM}$	28.66	50.66	22.00	8.45	= .004*
$\hat{\pi}_{DQ} = \hat{\pi}_{CDM}$	15.59	21.94	6.35	0.97	= .324
$\hat{\pi}_{DQ} = \hat{\pi}_{CDTRM}$	15.59	22.92	7.33	2.27	= .132
$\hat{\pi}_{CDM} = \hat{\pi}_{CDTRM}$	21.94	22.92	0.98	0.01	= .905
$\hat{\pi}_{DQ} = (\hat{\pi}_{CDM} + \hat{\gamma}_{CDM})$	15.59	22.57	6.98	0.97	= .324
$\hat{\pi}_{DQ} = (\hat{\pi}_{CDTRM} + \hat{\gamma}_{CDTRM})$	15.59	28.66	13.07	3.41	= .065
$(\hat{\pi}_{CDM} + \hat{\gamma}_{CDM}) = (\hat{\pi}_{CDTRM} + \hat{\gamma}_{CDTRM})$	22.57	28.66	6.09	0.25	= .617
$\hat{\gamma}_{CDM} = 0$ %	0.62	0.00	0.62	0.02	= .901
$\hat{\gamma}_{CDTRM} = 0$ %	5.73	0.00	5.73	2.49	= .114
$\hat{\gamma}_{CDM} = \hat{\gamma}_{CDTRM}$	0.62	5.73	5.11	0.60	= .437
$\hat{\beta}_{CDM} = \hat{\beta}_{CDTRM}$	77.44	71.35	6.09	0.25	= .167

Note. DQ = Direct Questioning, CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model. In the DQ condition, $\hat{\pi}$ reflects the estimated proportion of carriers of the sensitive attribute. In the CDM and CDTRM conditions, $\hat{\pi}$ reflects the estimated proportion of honest carriers of the sensitive attribute, $\hat{\beta}$ reflect the estimated proportion of honest non-carriers, and $\hat{\gamma}$ (= 1 - $\hat{\pi}$ - $\hat{\beta}$) reflects the estimated proportion of respondents who did not adhere to the instructions. Consequently, in the CDM and CDTRM conditions, $\hat{\pi}$ reflects the lower bound for the proportion of carriers of the sensitive attribute and $\hat{\pi} + \hat{\gamma}$ reflects the upper bound.
$^{*}p < .05$

Table C3

Comparisons of sensitivity and specificity by questioning technique in the subsample of respondents with high understanding of the indirect questioning techniques. To enable a comparison of estimates for this subsample with estimates in the DQ condition, estimates for the DQ condition are also displayed. Note that in the DQ condition, no comprehension questions were used; therefore, DQ estimates in this table are identical to those in the general sample (Table 2). N = 1009, 36.20% of the total sample.

	Parameter 1 (in %)	Parameter 2 (in %)	|*Difference*| (in %)	Model fit ΔG^2 (*df* = 1)	*p*
Sensitivity					
$sen_{DQ} = 100\%$	25.47	100.00	74.53	8476.52	< .001*
$low\_b\_sen_{CDM} = 100\ \%$	34.29	100.00	65.71	839.03	< .001*
$low\_b\_sen_{CDTRM} = 100\ \%$	39.17	100.00	60.83	1813.49	< .001*
$up\_b\_sen_{CDM} = 100\ \%$	37.79	100.00	62.21	16.62	< .001*
$up\_b\_sen_{CDTRM} = 100\ \%$	47.28	100.00	52.72	22.72	< .001*
$sen_{DQ} = low\_b\_sen_{CDM}$	25.47	34.29	8.82	0.68	= .408
$sen_{DQ} = low\_b\_sen_{CDTRM}$	25.47	39.17	13.70	3.56	= .059
$low\_b\_sen_{CDM} = low\_b\_sen_{CDTRM}$	34.29	39.17	4.88	0.15	= .700
$sen_{DQ} = up\_b\_sen_{CDM}$	25.47	37.79	12.32	0.74	= .388
$sen_{DQ} = up\_b\_sen_{CDTRM}$	25.47	47.28	21.81	4.56	= .033*
$up\_b\_sen_{CDM} = up\_b\_sen_{CDTRM}$	37.79	47.28	9.49	0.28	= .596
Specificity					
$spec_{DQ} = 100\ \%$	97.88	100.00	2.12	135.77	< .001*

$low\_b\_spec_{CDM}$ = 100 %	86.74	100.00	13.26	3.67	= .055
$low\_b\_spec_{CDTRM}$ = 100 %	91.28	100.00	8.72	1.10	= .294
$up\_b\_spec_{CDM}$ = 100 %	86.74	100.00	13.26	3.67	= .055
$up\_b\_spec_{CDTRM}$ = 100 %	94.42	100.00	5.58	1.05	= .305
$spec_{DQ} = low\_b\_spec_{CDM}$	97.88	86.74	11.14	2.41	= .120
$spec_{DQ} = low\_b\_spec_{CDTRM}$	97.88	91.28	6.60	0.53	= .468
$low\_b\_spec_{CDM} = low\_b\_spec_{CDTRM}$	86.74	91.28	4.54	0.14	= .710
$spec_{DQ} = up\_b\_spec_{CDM}$	97.88	86.74	11.14	2.41	= .120
$spec_{DQ} = up\_b\_spec_{CDTRM}$	97.88	94.42	3.46	0.37	= .541
$up\_b\_spec_{CDM} = up\_b\_spec_{CDTRM}$	86.74	94.42	7.68	0.64	= .423

Note. sen = sensitivity, spec = specificity; DQ = Direct Questioning, CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model. In the CDM and CDTRM conditions, low_b refers to the lower bound for sensitivity and specificity, and up_b refers to the upper bound for sensitivity and specificity.
* $p < .05$.

Table C4

Demographics for the indirect questioning technique conditions by model comprehension (high: correctly answered all comprehension questions on the first attempt, low: failed to correctly answer at least one comprehension question on the first attempt).

	Understanding of the indirect questioning techniques		
	high (%)	low (%)	
Gender			
female	37.92	42.29	$\chi^2(1) = 2.84$, $p = .092$, *Cramer's V* = .04
male	62.08	57.71	
Age (years)			
18-25	12.86	9.51	
26-35	30.38	20.64	
36-45	21.06	21.93	$\chi^2(5) = 38.03$, $p < .001$*, *Cramer's V* = .13
46-55	19.51	24.13	
56-65	15.08	18.84	
> 65	1.11	4.95	
Educational achievement			
No school leaving certificate	0.22	0.28	
Lower secondary school leaving certificate	2.22	7.09	
Secondary school leaving certificate	9.76	18.00	
Subject-specific university entrance qualification	6.87	6.36	
Higher education entrance qualification	14.63	11.47	$\chi^2(8) = 68.80$, $p < .001$*, *Cramer's V* = .18
Completed vocational training	26.61	31.89	
Bachelor's degree	9.76	7.31	
Master's degree	27.05	16.65	
PhD	2.88	0.96	

Table C5

ANOVAS, means and standard deviations for subjective evaluation of the CDM and CDTRM by model comprehension (high: correctly answered all comprehension questions on the first attempt, low: failed at least one comprehension question on the first attempt).

	High comprehension		Low comprehension		Compre-hension	Questioning Technique	Interaction
Item	CDM	CDTRM	CDM	CDTRM	*F*	*F*	*F*
The question was comprehensible.	5.53 (1.56)	6.27 (1.09)	4.00 (1.91)	4.52 (1.85)	278.16*	41.41*	1.31
The question guaranteed the confidentiality of my response.	5.58 (1.68)	5.96 (1.36)	4.96 (1.76)	5.32 (1.67)	46.62*	15.58*	0.02
The way the question was asked was interesting.	5.63 (1.59)	6.00 (1.30)	4.93 (1.85)	5.13 (1.73)	68.09*	8.85*	0.78
The way the question was asked was reasonable.	4.89 (1.58)	5.52 (1.36)	4.06 (1.80)	4.54 (1.73)	91.45*	34.09*	0.59
The question was cumbersome to answer. (R)	3.30 (1.92)	4.54 (2.01)	3.10 (1.85)	3.44 (1.86)	38.49*	57.91*	18.48*
I carefully read and followed all instructions.	6.64 (0.91)	6.65 (0.70)	5.99 (1.27)	5.86 (1.39)	110.43*	0.78	1.21
I clearly knew which answer to pick.	5.97 (1.29)	6.28 (1.08)	4.14 (1.80)	4.51 (1.72)	382.40*	14.16*	0.09
I felt overwhelmed by the question. (R)	5.46 (1.62)	6.09 (1.31)	4.19 (1.77)	4.49 (1.75)	230.84*	24.55*	3.00
I just randomly ticked one of the answers. (R)	1.21 (0.89)	1.10 (0.55)	1.71 (1.42)	1.60 (1.31)	51.92*	2.32	0.00
Mean of the scale (*Cronbach* α = .84)	5.53 (0.84)	6.02 (0.78)	4.63 (1.07)	4.91 (1.02)	330.04*	48.87*	3.57

Note. CDM = Cheating Detection Model, CDTRM = Cheating Detection Triangular Model. All variables were assessed on a 7-point Likert-type scale with higher values indicating more favorable evaluations. Questions marked with an (R) have negative polarity and were reverse-coded prior to analysis to facilitate their interpretability and the computation of a scale mean.
* $p < .01$