Indirekte Befragungstechniken zur Kontrolle sozialer Erwünschtheit in Umfragen

Inaugural-Dissertation

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

vorgelegt von
Adrian Hoffmann
geboren in Hamburg

Düsseldorf, Oktober 2014

Aus dem Institut für experimentelle Psychologie der
Heinrich-Heine-Universität Düsseldorf

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

Referent: Prof. Dr. Jochen Musch
Korreferent: Prof. Dr. Axel Buchner

Tag der mündlichen Prüfung: 28.11.2014

Inhaltsverzeichnis

Zusammenfassung

Die Validität von Prävalenzschätzungen für sensible Merkmale, die auf der Basis direkter Selbstauskünfte erhoben werden, wird durch den Einfluss sozialer Erwünschtheit bedroht. So antworten manche Merkmalsträger eher in Einklang mit gültigen sozialen Normen als mit ihrem wahren Merkmalsstatus, was zu einer Unterschätzung der Prävalenz sozial unerwünschter Einstellungen und Verhaltensweisen führen kann. Mit der Randomized-Response-Technik (RRT) hat Warner (1965) eine indirekte Befragungstechnik vorgestellt, die durch eine Zufallsverschlüsselung die Vertraulichkeit individueller Antworten garantiert und so zu ehrlicherem Antwortverhalten führen soll. Das Crosswise-Modell (CWM; Yu, Tian & Tang, 2008) ist eine aktuelle Weiterentwicklung der RRT. Im Vergleich mit Vorgängermodellen verfügt das CWM über vereinfachte Instruktionen und bietet symmetrische Antwortkategorien, also keine „sichere“ Antwortalternative, die eine Merkmalsträgerschaft ausschließt. Durch diese Eigenschaften könnte das CWM zu noch valideren Prävalenzschätzungen führen als konkurrierende Modelle. In der vorliegenden Arbeit wurde die Validität des CWM einer eingehenden empirischen Überprüfung unterzogen. Hierzu wurde das CWM zunächst in einem Methodenvergleich einem aktuellen RRT-Modell, dem Stochastischen Lügendetektor (Moshagen, Musch & Erdfelder, 2012), gegenübergestellt. Für zwei sensible Merkmale konnten beide indirekten Befragungstechniken höhere und damit potentiell validere Prävalenzschätzungen als eine konventionelle direkte Frage erzielen. Die bekannte Prävalenz eines nichtsensiblen Kontrollmerkmals wurde durch das CWM punktgenau und sogar genauer als durch das konkurrierende RRT-Modell geschätzt. In einem zweiten Experiment wurden gezielt die Verständlichkeit der Instruktionen und die subjektiv empfundene Vertraulichkeit erhoben, die zwei zentrale Variablen beim Einsatz indirekter Befragungstechniken darstellen. Im Vergleich mit drei konkurrierenden indirekten Befragungstechniken zeigte das CWM die höchste Verständlichkeit. Auch in der subjektiv

empfundenen Vertraulichkeit schnitt das CWM mit einem der höchsten Werte ab, der außerdem gegenüber einer direkten Frage substantiell erhöht war. Schließlich wurde das CWM in einem dritten Experiment einer „starken“ Validierung unterzogen, in welcher die bekannte Prävalenz eines sensiblen Merkmals in der Stichprobe als objektives Außenkriterium für die Validität der ermittelten Schätzung diente. Während eine direkte Befragung erwartungsgemäß zu einer substantiellen Unterschätzung der Prävalenz führte, konnte mit Hilfe des CWM ein akkurater Schätzer gewonnen werden. Die ermittelten Befunde legen nahe, dass das CWM bezüglich der Anwendbarkeit und der Validität der gewonnenen Ergebnisse konkurrierenden RRT-Modellen überlegen ist. Als Ergebnis seiner Evaluation kann festgehalten werden, dass das CWM eines der vielversprechendsten Verfahren zur Schätzung der Prävalenz sensibler Merkmale darstellt.

Abstract

The validity of prevalence estimates of sensitive attributes obtained through direct questions is threatened by the influence of social desirability bias. Especially carriers of sensitive attributes tend to refrain from answering truthfully in order to present themselves in a socially desirable light, possibly resulting in an underestimation of the prevalence of these attributes. Warner (1965) introduced the Randomized-Response-Technique (RRT) as a means of keeping individual answers confidential by making use of a randomization procedure, presumably leading to more honest responding. As a promising advancement of the RRT, the Crosswise Model (CWM; Yu et al., 2008) implements simpler instructions than competing RRT approaches. Moreover, the CWM offers response symmetry in the sense that none of the answer options offers a "safe" alternative to which respondents might turn to dispel any potential connection to the sensitive attribute. These features might lead to a higher validity of CWM prevalence estimates when compared to estimates based on competing models. The present thesis aimed to empirically evaluate the validity of the CWM. First, the CWM was experimentally compared to an established variant of the RRT. For two sensitive attributes, both techniques met the *more is better*-criterion; i.e., their application resulted in higher, and thus presumably more valid, prevalence estimates than a direct question. The CWM, however, produced a more exact estimate of the known prevalence of a nonsensitive control attribute. Second, two psychological aspects of the application of indirect questioning techniques, the comprehensibility and perceived privacy protection, were assessed. When compared to three competing approaches, the CWM was found to be the most comprehensible format. Moreover, the CWM was perceived to increase privacy protection as compared to a direct question. Third, a "strong" validation study of the CWM was conducted, using the known prevalence of an experimentally induced sensitive attribute as an external criterion. While a conventional direct question resulted in a substantial underesti-

mate of the known prevalence, a CWM question resulted in an accurate estimate. These results suggest that the CWM might be superior to competing approaches with regards to its applicability and validity. Therefore, the CWM is evaluated as one of the most promising techniques for assessing the prevalence of sensitive attributes.

1 Einleitung

In sozialwissenschaftlichen Studien, die die Erfassung der Prävalenz von persönlichen Einstellungen und Verhaltensweisen zum Ziel haben, stellen Selbstauskünfte von Befragten die am häufigsten verwendete Datenquelle dar. Besonders bei sensiblen Fragestellungen, deren wahrheitsgemäße Beantwortung negative Konsequenzen für die Befragten haben könnte, weichen Selbstauskünfte jedoch mitunter vom wahren Merkmalsstatus der Befragten ab. Für eine positive Selbstdarstellung oder um befürchtete soziale Missbilligung, Schamgefühle und rechtliche Sanktionen zu vermeiden, passen einige Befragte ihr Antwortverhalten den gültigen sozialen Normen an und antworten unehrlich. Dieser Effekt *sozialer Erwünschtheit* gefährdet die Validität von Befragungsergebnissen, da er zu einer Überschätzung der Prävalenz sozial erwünschter und zu einer Unterschätzung der Prävalenz sozial unerwünschter Einstellungen und Verhaltensweisen führen kann (Krumpal, 2013; Paulhus, 1991; Phillips & Clancy, 1972; Sudman & Bradburn, 1974; Tourangeau & Yan, 2007).

Um dem Einfluss sozial erwünschten Antwortverhaltens zu begegnen, wurden verschiedene Ansätze vorgestellt. So kann einerseits anhand spezialisierter Skalen ermittelt werden, wie stark einzelne Befragte unter dem Einfluss sozialer Erwünschtheit stehen (z.B. Musch, Brockhaus & Bröder, 2002; Paulhus, 1994; Paulhus, 1998; Stöber, 1999). Wenngleich die differentielle Messung sozialer Erwünschtheit Rückschlüsse auf die individuelle Prädisposition zur positiven Selbstdarstellung zulässt, können bezüglich der übergeordneten Befragungsergebnisse lediglich Hinweise auf eine mögliche Gefährdung der Validität gewonnen werden; unehrliche Antworten werden hierdurch nicht vermieden. Andererseits existieren Methoden, die durch einen erhöhten Druck in der Befragungssituation versuchen, ehrlichere Antworten zu provozieren. Als zwei prominente Vertreter dieser Klasse sind jedoch sowohl die *psychophysiologische Lügendetektion* (Iacono, 2000) als auch die *Bogus-Pipeline-Technik* (Jones & Sigall, 1971) mit hohem technischem Aufwand und teilweise erheb-

lichen rechtlichen und ethischen Konflikten verbunden. Ein vielversprechendes alternatives Verfahren zur Kontrolle sozialer Erwünschtheit in Umfragen stellt die Anwendung indirekter Befragungstechniken wie der *Randomized-Response-Technik* (RRT; Warner, 1965) dar. Diese garantiert durch eine Zufallsverschlüsselung die Vertraulichkeit individueller Antworten und erhöht so die Bereitschaft von Befragten, zu sensiblen Themen ehrlich Stellung zu beziehen. Da Prävalenzschätzungen, die mit Hilfe von RRT-Fragen gewonnen wurden, potentiell weniger durch den Einfluss sozialer Erwünschtheit verzerrt sind, weisen diese in der Regel eine höhere Validität auf als Prävalenzschätzungen auf der Basis konventioneller direkter Selbstauskünfte (Antonak & Livneh, 1995; Fox, J. A. & Tracy, 1986; Horvitz, Greenberg & Abernathy, 1976; Lensvelt-Mulders, Hox, van der Heijden & Maas, 2005). Besonders mit Hilfe weiterentwickelter Modelle wie dem *Cheating Detection Model* (Clark & Desharnais, 1998) oder dem *Stochastischen Lügendetektor* (Moshagen et al., 2012), die ein Nichtbefolgen der Instruktionen durch einen Teil der Befragten mit in die Modellannahmen einbeziehen, konnten in der Vergangenheit validere Prävalenzschätzungen für sensible Merkmale gewonnen werden als durch eine direkte Befragung (z.B. Moshagen, Hilbig, Erdfelder & Moritz, 2014; Ostapczuk & Musch, 2011; Ostapczuk, Musch & Moshagen, 2011).

Das Crosswise-Modell (CWM; Yu et al., 2008) stellt eine aktuelle Weiterentwicklung der RRT dar. Im Vergleich zu Vorgängermodellen bietet das CWM stark vereinfachte Instruktionen, was die praktische Anwendung erleichtert sowie potentiell das Verständnis und die subjektiv empfundene Vertraulichkeit auf Seiten der Befragten erhöht. Zusätzlich verfügt das CWM über symmetrische Antwortkategorien, die keine Möglichkeit bieten, die Trägerschaft eines sensiblen Merkmals sicher auszuschließen. Hierdurch soll die Motivation zu ehrlichen Antworten gefördert werden. Diese Eigenschaften könnten dazu führen, dass eine Berücksichtigung des Nichtbefolgens der Instruktionen entbehrlich wird und Prävalenzschätzungen auf Basis des CWM eine hohe Validität aufweisen. Die vorliegende Dissertation hatte

zum Ziel, die Validität des CWM einer eingehenden empirischen Prüfung zu unterziehen. Zu diesem Zweck wurde das CWM zunächst in einem Methodenvergleich einem konkurrierenden RRT-Modell, dem SLD (Moshagen et al., 2012), gegenübergestellt (Experiment 1). In einem zweiten Schritt wurden die objektive Verständlichkeit und die subjektiv empfundene Vertraulichkeit von Fragen im Format des CWM, drei konkurrierenden indirekten Befragungstechniken und einer konventionellen direkten Frage erhoben und verglichen (Experiment 2). Schließlich wurde das CWM einer „starken" Validierung unterzogen, in welcher die Prävalenz des sensiblen Merkmals in der Stichprobe bekannt war und als objektives Außenkriterium für die Validität der CWM-Prävalenzschätzung herangezogen werde konnte (Experiment 3).

Die folgende Darstellung gliedert sich zunächst in eine Einführung in das Themengebiet der RRT (Kapitel 2), die Beschreibung der Modellannahmen des CWM und des diesbezüglichen Forschungsstands (Kapitel 3) sowie eine Herleitung der zentralen Forschungsfragen der vorliegenden Dissertation (Kapitel 4). Anschließend werden die Einzelarbeiten zusammengefasst (Kapitel 5), die ermittelten Ergebnisse diskutiert und ein Ausblick auf zukünftige Fragestellungen gegeben (Kapitel 6). Die vollständigen Einzelarbeiten sind dem Anhang der Arbeit zu entnehmen.

2 Die Randomized-Response-Technik (RRT)

Das grundlegende Funktionsprinzip der RRT besteht darin, die Vertraulichkeit individueller Antworten auf sensible Fragen durch eine Zufallsverschlüsselung zu garantieren, die keine Verbindung mehr zwischen den Antworten einzelner Umfrageteilnehmer[1] und ihrem wahren Merkmalsstatus zulässt. Im ursprünglichen RRT-Modell von Warner (1965) werden den Versuchsteilnehmern gleichzeitig zwei Fragen präsentiert: die sensible Frage A („Sind Sie Träger des sensiblen Merkmals?“) und deren Negation, Frage B („Sind Sie kein Träger des sensiblen Merkmals?“). Anhand eines Zufallsgenerators (z.B. eines Würfels) wird nun entschieden, zu welcher der beiden Fragen Stellung genommen werden soll. So könnten die Instruktionen beispielsweise lauten, auf Frage A zu antworten, wenn eine der Zahlen eins bis vier gewürfelt wurde (mit Randomisierungswahrscheinlichkeit p = 4/6 = .67), und auf Frage B zu antworten, wenn eine fünf oder sechs gewürfelt wurde (mit Wahrscheinlichkeit 1-p = 2/6 = .33). Hierbei wird das Zufallsexperiment (z.B. der Würfelwurf) von den Teilnehmern selbst durchgeführt und das Ergebnis gegenüber dem befragenden Versuchsleiter geheim gehalten. Aus den Antworten der Probanden können folglich keine Rückschlüsse über ihren Merkmalsstatus in Bezug auf das sensible Merkmal gezogen werden: Eine „ja“-Antwort kann ebenso von einem Merkmalsträger stammen, der auf Frage A geantwortet hat, wie von einem Nicht-Merkmalsträger, der auf Frage B geantwortet hat. Von der Zufallsverschlüsselung individueller Antworten wird erwartet, dass diese die Bereitschaft zu ehrlichem Antwortverhalten gegenüber einer konventionellen direkten Frage erhöht. Wenngleich durch die Zufallsverschlüsselung keine Informationen mehr über den Merkmalsstatus einzelner Probanden verfügbar sind, kann auf Stichprobenebene eine Schätzung für die Prävalenz π des sensiblen Merkmals gewonnen werden, da die

[1] Im Folgenden wird aus Gründen der Übersichtlichkeit für Begriffe wie Umfrageteilnehmer, Versuchsteilnehmer, Befragter, etc. auf die erweiternde Nennung der weiblichen Form verzichtet. Wenn nicht anders gekennzeichnet, gelten diese Begriffe äquivalent für Vertreter beider Geschlechter.

Randomisierungswahrscheinlichkeit p bekannt ist. Nach Warner (1965) ist der Maximum-Likelihood-Schätzer für π gegeben durch

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

Hierbei entspricht n' der absoluten Anzahl von „ja"-Antworten und n der Stichprobengröße. Die Varianz des Schätzers kann berechnet werden als

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

Wie aus Gleichung 2 hervorgeht, wird dem üblichen Varianzanteil einer binomial verteilten Zufallsvariablen (links vom Additionszeichen) hier ein weiterer Varianzanteil hinzugefügt, der sich aus der Zufallsverschlüsselung der individuellen Antworten ergibt (rechts vom Additionszeichen). Verglichen mit Schätzungen auf der Basis von konventionellen direkten Fragen sind RRT-Schätzungen somit ineffizienter; um Schätzer mit vergleichbarer Genauigkeit zu bestimmen, müssen in RRT-Studien deutlich höhere Stichprobenzahlen realisiert werden (Ulrich, Schröter, Striegel & Simon, 2012). Diese reduzierte Effizienz wird dann – und nur dann – als akzeptabel erachtet, wenn durch die Zufallsverschlüsselung und die damit einhergehende gesteigerte Kooperationsbereitschaft der Befragten tatsächlich ein Validitätszugewinn für die Prävalenzschätzung erzielt werden kann.

Seit der Einführung der RRT durch Warner (1965) wurde eine Vielzahl weiterentwickelter RRT-Modelle vorgestellt. Diese zielten vornehmlich auf eine verbesserte Effizienz (z.B. Boruch, 1971; Dawes & Moore, 1980; Eriksson, 1973; Mangat, 1994; Mangat & Singh, 1990; Moors, 1971), eine Berücksichtigung von Fragen mit mehr als zwei Antwortkategorien oder quantitativem Antwortformat (z.B. Abul-Ela, Greenberg & Horvitz, 1967; Himmelfarb & Edgell, 1980; Liu & Chow, 1976; Pollock & Bek, 1976), eine Steigerung der Kooperationsbereitschaft der Befragten (z.B. Greenberg, Abul-Ela, Simmons & Horvitz, 1969; Horvitz, Shah & Simmons, 1967;

Kuk, 1990; Ostapczuk, Moshagen, Zhao & Musch, 2009) sowie die Berücksichtigung von „Betrügern“, die vorsätzlich die Instruktionen missachten (z.B. Clark & Desharnais, 1998; Moshagen et al., 2012), ab. Fragen im RRT-Format wurden in Umfragen zu vielfältigen sensiblen Themenbereichen eingesetzt. Hierzu zählen unter anderem Drogenkonsum (Dietz et al., 2013; Goodstadt & Gruson, 1975), Doping (James, Nepusz, Naughton & Petroczi, 2013; Simon, Striegel, Aust, Dietz & Ulrich, 2006; Striegel, Ulrich & Simon, 2010), organisierte Kriminalität (IIT Research Institute and the Chicago Crime Commission, 1971; Wolter & Preisendörfer, 2013), uneheliche Schwangerschaft (Abul-Ela et al., 1967), Promiskuität (Liu, Chow & Mosley, 1975), Abtreibung (Abernathy, Greenberg & Horvitz, 1970; Greenberg, Kuebler, Abernathy & Horvitz, 1971), Vergewaltigung (Fidler & Kleinknecht, 1977; Soeken & Damrosch, 1986), Homosexualität (Clark & Desharnais, 1998), Steuerhinterziehung (Edgell, Himmelfarb & Duchan, 1982), Betrug (van der Heijden, van Gils, Bouts & Hox, 2000), akademisches Fehlverhalten (Fox, J. P. & Meijer, 2008; Hejri, Zendehdel, Asghari, Fotouhi & Rashidian, 2013; Ostapczuk, Moshagen, et al., 2009), Fremdenfeindlichkeit (Ostapczuk, Musch & Moshagen, 2009), Vorurteile gegenüber Menschen mit Behinderung (Ostapczuk & Musch, 2011), Zahnhygiene (Moshagen, Musch, Ostapczuk & Zhao, 2010) und häusliche Gewalt (Moshagen et al., 2012). Für eine umfassende Übersicht von RRT-Modellen und Anwendungen sei an dieser Stelle auf die einschlägigen Literaturreviews und Monographien verwiesen, wie beispielsweise Greenberg, Horvitz und Abernathy (1974), Horvitz et al. (1976), Fox und Tracy (1986), Chaudhuri und Mukerjee (1988), Umesh und Peterson (1991), Scheers (1992), Antonak und Livneh (1995), Tracy und Mangat (1996), Franklin (1998), Chaudhuri (2011) und Chaudhuri und Christofides (2013).

In zwei Metaanalysen konnten Lensvelt-Mulders et al. (2005) zeigen, dass die RRT insgesamt zu valideren Prävalenzschätzungen führte als direkte Befragungen. Zunächst ergab die Auswertung von 32 vergleichenden Validierungsstudien, dass RRT-Fragen im Mittel höhere Prävalenzschätzer für sensible Merkmale produzierten

als direkte Fragen. Dieses Ergebnis entspricht dem *more-is-better*-Kriterium (Umesh & Peterson, 1991), welches höhere Prävalenzschätzungen für sozial *un*erwünschte Merkmale als valider definiert, da eine Verzerrung durch soziale Erwünschtheit logisch nur zu einer *Unter*schätzung der Prävalenz führen sollte. Höhere Prävalenzschätzungen werden somit als Anzeichen für eine gelungene Kontrolle des Einflusses sozialer Erwünschtheit angesehen. Gleichwohl gelten Ergebnisse dieser Art nur als „schwache" Evidenz für die Validität von indirekten Befragungsmethoden, da diese höheren Schätzer immer noch eine Unterschätzung der Prävalenz darstellen könnten (Umesh & Peterson, 1991). In eine weitere metaanalytische Auswertung bezogen Lensvelt-Mulders et al. (2005) deshalb sechs „starke" Validierungsstudien ein, in welchen die Prävalenz des sensiblen Merkmals in der Stichprobe bekannt war und als objektives Außenkriterium für die Validität der Befragungsmethode diente. Auch diese Analyse ergab, dass die Validität von Prävalenzschätzungen auf Basis der RRT gegenüber direkten Befragungen höher ausfiel, weil die Abweichung der RRT-Schätzer vom wahren Wert deutlich geringer war als die Abweichung von Schätzern auf der Basis direkter Selbstauskünfte.

Nicht alle Studien konnten jedoch bisher Belege für einen Validitätsvorteil der RRT finden. In manchen Fällen wurden Prävalenzschätzungen ermittelt, die in ihrer Höhe mit den Schätzungen aus einer direkten Befragung vergleichbar waren (z.B. Akers, Massey, Clarke & Lauer, 1983; Locander, Sudman & Bradburn, 1976; Wolter & Preisendörfer, 2013) oder sogar darunter lagen (z.B. Holbrook & Krosnick, 2010; Kulka, Weeks & Folsom, 1981). Darüber hinaus wurde in Studien mit bekanntem Merkmalsstatus der Teilnehmer gezeigt, dass einige Befragte sich nicht an die Instruktionen hielten und falsche Antworten gaben – besonders, wenn das untersuchte Merkmal hoch sensibel war (Edgell, Duchan & Himmelfarb, 1992; Edgell et al., 1982). In der Folge stellten Holbrook und Krosnick (2010) die Validität der gesamten RRT in Frage. Eine plausible Alternativerklärung für die kontroversen Befunde kann in der Motivation einiger Versuchsteilnehmer gefunden werden, sich explizit von einer

Trägerschaft vor allem höchst sensibler Merkmale zu distanzieren (Edgell et al., 1982). Aus einer übergeordneten Perspektive bezogen Antonak und Livneh (1995) diese Motivation in ihre Definition sogenannter *response hazards* (etwa „Bedrohungen für ein ehrliches Antwortverhalten“) mit ein, die sich in *respondent jeopardy* (etwa „Bedrohung für den Befragten“) und *risk of suspicion* (etwa „Risiko einer Verdächtigung“) unterteilen lassen. *Respondent jeopardy* steht hier für das von Trägern des sensiblen Merkmals empfundene Risiko, als solche identifiziert zu werden. *Risk of suspicion* wiederum bezieht sich auf die Befürchtung von Nicht-Merkmalsträgern, fälschlicherweise als Träger klassifiziert zu werden. Bei der Anwendung von RRT-Modellen, die eine „sichere“ Antwortalternative bieten, bei deren Wahl also eine Merkmalsträgerschaft ausgeschlossen ist (z.B. forced-choice-Designs; Boruch, 1971; Dawes & Moore, 1980), können diese *response hazards* zu unehrlichem Antwortverhalten führen, das wiederum in verzerrten Prävalenzschätzungen resultiert (Antonak & Livneh, 1995). Diesem Umstand wurde durch die Einführung von RRT-Modellen mit Verweigererdetektion Rechnung getragen.

Randomized-Response-Modelle mit Verweigererdetektion

Obwohl die Vertraulichkeit individueller Antworten durch die RRT garantiert wird, entscheiden sich manche Umfrageteilnehmer gegen ein Befolgen der Instruktionen und wählen – wenn möglich – die Antwortalternative, die eine Merkmalsträgerschaft sicher ausschließt. Aus diesem Grund berücksichtigten Clark und Desharnais (1998) nicht nur zwei (Merkmalsträger und Nicht-Merkmalsträger), sondern drei Klassen von Teilnehmern bei der Formulierung ihres *Cheating Detection Models* (CDM): Merkmalsträger (π), Nicht-Merkmalsträger (β) und Verweigerer (γ). Die Instruktionen des CDM entsprechen denen eines *Forced-Choice*-Paradigmas (Boruch, 1971; Dawes & Moore, 1980): Versuchsteilnehmern wird hier nur eine sensible Frage präsentiert. Der Ausgang eines Zufallsexperiments entscheidet nun, ob auf diese Frage

wahrheitsgemäß oder – völlig ungeachtet des wahren Merkmalsstatus – mit „ja“ geantwortet werden soll. Somit lässt eine „ja“-Antwort keine Rückschlüsse über den wahren Merkmalsstatus eines Teilnehmers zu, da diese gleichermaßen aus einer ehrlichen Antwort wie aus einer durch den Zufallsgenerator erzwungenen Antwort stammen kann. Verweigerer im Sinne des CDMs sind nun Umfrageteilnehmer, die sich ungeachtet der Instruktionen kategorisch für eine „nein“-Antwort entscheiden. Diese Entscheidung kann möglicherweise sowohl anhand von *respondent jeopardy* als auch anhand von *risk of suspicion* erklärt werden; der wahre Merkmalsstatus von Verweigerern bleibt im CDM ausdrücklich ungeklärt. Mehrere Studien konnten bereits zeigen, dass anhand des CDM höhere und damit potentiell validere Ergebnisse erzielt werden können als durch eine direkte Befragung (z.B. Moshagen et al., 2010; Ostapczuk, Moshagen, et al., 2009; Ostapczuk & Musch, 2011; Ostapczuk, Musch, et al., 2009; Ostapczuk et al., 2011; Pitsch, Emrich & Klein, 2007). Andererseits wird die Interpretation der Ergebnisse von CDM-Studien erschwert durch den Befund, dass der geschätzte Anteil von Verweigerern oft substantiell ist und bis zu 50% betragen kann (z.B. Ostapczuk, Moshagen, et al., 2009; Ostapczuk & Musch, 2011; Ostapczuk, Musch, et al., 2009; Ostapczuk et al., 2011). Eine Schätzung der Prävalenz des sensiblen Merkmals ist somit nur noch im mitunter breiten Intervall von π (wenn kein Verweigerer Merkmalsträger ist) bis $\pi+\gamma$ (wenn alle Verweigerer Merkmalsträger sind; Clark & Desharnais, 1998) möglich.

Ein Modell, das potentiell zu unverzerrten Prävalenzschätzungen und zusätzlich zu einer Schätzung von Verweigerern mit *bekanntem* Merkmalsstatus führt, wurde von Moshagen et al. (2012) mit dem *Stochastischen Lügendetektor* (SLD) vorgestellt. Versuchsteilnehmern werden hier wiederum zwei Fragen präsentiert: die sensible Frage A und ihre Negation, Frage B. Basierend auf einem Vorgängermodell von Mangat (1994) werden Merkmalsträger nun jedoch angewiesen, die Randomisierungsprozedur zu ignorieren und in jedem Fall zu Frage A Stellung zu nehmen; Nicht-Merkmalsträger durchlaufen das Zufallsexperiment und antworten mit Wahr-

scheinlichkeit p_i auf Frage A, sowie mit der Gegenwahrscheinlichkeit 1-p_i auf Frage B. Den Modellannahmen von Moshagen et al. (2012) folgend sollten bei einer Befragung unter Anwendung des SLD nur Merkmalsträger den Drang verspüren, ihren wahren Merkmalsstatus zu verschleiern und entgegen den Instruktionen mit „nein" zu antworten; Nicht-Merkmalsträger sollten hingegen ehrlich und instruktionskonform reagieren. Damit werden durch den SLD explizit die Gefahren für die Validität der ermittelten Ergebnisse berücksichtigt, die von *respondent jeopardy* (Antonak & Livneh, 1995) ausgehen. Der Anteil ehrlich antwortender Merkmalsträger wird im SLD repräsentiert durch einen zusätzlichen Parameter t (für „true"); der verbleibende Anteil 1-t beziffert somit den Anteil unehrlich antwortender Merkmalsträger (siehe Abbildung 1).

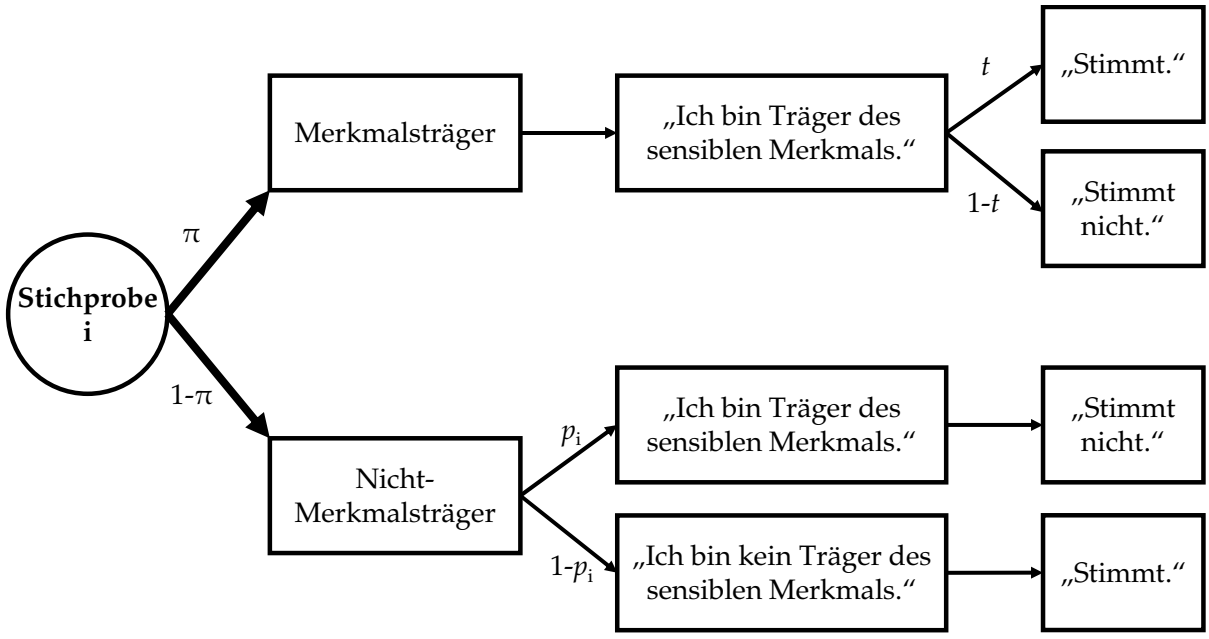

Abbildung 1: Multinomiales Modell des stochastischen Lügendetektors (Moshagen et al., 2012).

Um die Schätzung der beiden Parameter π und t zu ermöglichen, bedarf es der Erhebung zweier unabhängiger Stichproben mit unterschiedlichen Randomisierungswahrscheinlichkeiten p_1 und p_2; eine größere Differenz von p_1 und p_2 geht hierbei mit einer höheren statistischen Effizienz des Modells einher (Clark & Desharnais, 1998;

Moshagen et al., 2012). Eine Maximum-Likelihood-Schätzung für π und t kann Moshagen et al. (2012) zufolge bestimmt werden anhand von

$$\hat{\pi} = \frac{\left(\frac{{n_2}'}{n_2} - \frac{{n_1}'}{n_1}\right) + (p_2 - p_1)}{(p_2 - p_1)} \tag{3}$$

und

$$\hat{t} = \frac{\left(\frac{{n_2}'}{n_2}(1 - p_1)\right) - \left(\frac{{n_1}'}{n_1}(1 - p_2)\right)}{\left(\frac{{n_2}'}{n_2} - \frac{{n_1}'}{n_1}\right) + (p_2 - p_1)} \,. \tag{4}$$

Hierbei beziffern n_1 und n_2 die Stichprobengrößen der beiden getesteten Stichproben und n_1' und n_2' die absoluten Häufigkeiten von „Stimmt"-Antworten in diesen Stichproben. Die Varianz von $\hat{\pi}$ und $\hat{t}$ ist entsprechend gegeben durch

$$Var(\hat{\pi}) = \frac{\frac{{n_1}'}{n_1}\left(1 - \frac{{n_1}'}{n_1}\right)}{n_1(p_1 - p_2)^2} + \frac{\frac{{n_2}'}{n_2}\left(1 - \frac{{n_2}'}{n_2}\right)}{n_2(p_2 - p_1)^2} \tag{5}$$

und

$$\begin{aligned} Var(\hat{t}) = {} & \frac{\frac{{n_1}'}{n_1}\left(1 - \frac{{n_1}'}{n_1}\right)}{n_1}\left(\frac{(p_1 - p_2)\left(p_2 + \frac{{n_2}'}{n_2} - 1\right)}{\left(p_1 - p_2 + \frac{{n_1}'}{n_1} - \frac{{n_2}'}{n_2}\right)^2}\right)^2 \\ & + \frac{\frac{{n_2}'}{n_2}\left(1 - \frac{{n_2}'}{n_2}\right)}{n_2}\left(\frac{(p_1 - p_2)\left(p_1 + \frac{{n_1}'}{n_1} - 1\right)}{\left(p_1 - p_2 + \frac{{n_1}'}{n_1} - \frac{{n_2}'}{n_2}\right)^2}\right)^2 \,. \end{aligned} \tag{6}$$

Ergebnisse aus zwei Pilotstudien mit dem SLD werden in Moshagen et al. (2012) berichtet: In einer Umfrage zu häuslicher Gewalt, in welcher die Befragungstechnik experimentell manipuliert wurde, konnte anhand des SLD zunächst eine Prävalenzschätzung gewonnen werden, die etwa viermal so hoch ausfiel wie eine Schätzung durch eine direkte Frage und etwa doppelt so hoch wie die Schätzung auf Basis des

Vorgängermodells von Mangat (1994). Zusätzlich lag die Schätzung für den Parameter *t* signifikant unter 100%, was vermuten lässt, dass sich ein substantieller Anteil von Merkmalsträgern für die sichere Alternative entschieden und gelogen hatte, um sich vom sensiblen Merkmal zu distanzieren (Moshagen et al., 2012). In einem zweiten Experiment konnte für eine Frage zum Nichtwählen gezeigt werden, dass der SLD wiederum die höchste Prävalenzschätzung produzierte; darüber hinaus korrespondierte diese Schätzung annähernd exakt mit dem Anteil von Nichtwählern in der Population, der aus offiziellen Statistiken für die betreffende Bundestagswahl im Jahr 2009 bekannt war (Moshagen et al., 2012). Noch beeindruckender sind die Ergebnisse einer starken Validierungsstudie des SLD von Moshagen et al. (2014), in welcher die bekannte Prävalenz eines sensiblen Merkmals als Außenkriterium für die Validität der SLD-Prävalenzschätzung herangezogen wurde. Hier wurde Teilnehmern eine Modifikation des *die-under-the-cup*-Paradigmas (vgl. Hilbig & Hessler, 2013) vorgegeben, bei dem der Ausgang eines verdeckt durchgeführten Würfelwurfes berichtet werden sollte und zufällig bestimmte Würfelzahlen mit einer finanziellen Belohnung verbunden waren. Da die Ergebnisse einzelner Würfeldurchgänge den Versuchsleitern unbekannt waren, blieb das Verhalten individueller Teilnehmer vertraulich. Auf Stichprobenebene konnte aus den Antworthäufigkeiten und der bekannten Verteilung von Würfelergebnissen jedoch berechnet werden, dass etwa 53% der Teilnehmer, die einen Gewinn beanspruchten, in Wirklichkeit ein anderes Ergebnis gewürfelt und damit betrogen hatten. Während eine direkte Frage diesen Anteil von Betrügern mit 36% substantiell unterschätzte, konnte auf Basis des SLD eine höhere Schätzung von 48% erreicht werden, die sich darüber hinaus nicht signifikant von der bekannten wahren Prävalenz unterschied. Folglich wurden der SLD und die mit Hilfe dieses Modells gewonnenen Prävalenzschätzungen als hoch valide evaluiert (Moshagen et al., 2014).

In Anbetracht dieser Ergebnisse scheint der SLD geeignet, die Gefahren von *respondent jeopardy* zu kontrollieren, beziehungsweise diese messbar zu machen. Der

SLD verfügt jedoch über eine sichere Antwortalternative: Eine „Nein"-Antwort schließt den Instruktionen folgend eine Merkmalsträgerschaft explizit aus (wenngleich unehrliche „Nein"-Antworten von Merkmalsträgern durch die Einführung des *t*-Parameters berücksichtigt wurden). Sollten sich also Nicht-Merkmalsträger – beispielsweise aufgrund eines hohen wahrgenommenen *risk of suspicion* – für eine unehrliche „Nein"-Antwort entscheiden, so verletzt dies klar die Modellannahmen und könnte in einer Verzerrung der Prävalenzschätzung münden. Das konkurrierende Crosswise-Modell (Yu et al., 2008) bietet hingegen ein Format, das den Einfluss von *respondent jeopardy* und *risk of suspicion* womöglich von vornherein ausschließt; dieses Modell wird im Folgenden genauer dargestellt.

3 Das Crosswise-Modell (CWM)

Eine vielversprechende Weiterentwicklung der RRT stellt die kürzlich etablierte Klasse der sogenannten *Nonrandomized Response Techniques* (NRRT; Tian & Tang, 2013) dar. Diese Verfahren ermöglichen durch ein modifiziertes Frageformat, Teilnehmer ohne die Anwendung eines externen Zufallsgenerators (z.B. eines Würfels) indirekt zu sensiblen Merkmalen zu befragen. Mit dem *Crosswise-Modell* (CWM) haben Yu et al. (2008) eine Befragungstechnik vorgestellt, die verhältnismäßig einfach zu implementieren ist und deren Instruktionen von Teilnehmern womöglich besser verstanden werden als die Instruktionen konkurrierender Modelle. Zusätzlich bietet das Modell den Vorteil symmetrischer Antwortkategorien, die keine sichere Antwortalternative einschließen. Hierzu werden Umfrageteilnehmern zwei Aussagen gleichzeitig präsentiert: eine Aussage, die sich auf das sensible Merkmal mit unbekannter Prävalenz π bezieht, und eine zweite Aussage zu einem nicht-sensiblen Merkmal (z.B. zum Geburtsmonat des Befragten) mit bekannter Prävalenz p. Zur Beantwortung der Frage müssen Teilnehmer nun angeben, ob a) beide Aussagen zutreffen oder keine der beiden Aussagen zutrifft, oder ob b) genau einer der beiden Aussagen (egal welche) zutrifft. Keine der Antwortoptionen a) oder b) lässt für einzelne Befragte Rückschlüsse auf eine Merkmalsträgerschaft zu; ebenso wird von keiner der beiden Optionen sicher ausgeschlossen, dass Befragte das sensible Merkmal tragen. Folglich werden Antwortverzerrungen, die durch *respondent jeopardy* oder *risk of suspicion* zu erklären sind, womöglich gänzlich vermieden. Tatsächlich konnte in Studien zu anderen indirekten Befragungstechniken gezeigt werden, dass Modelle mit symmetrischen Antwortkategorien zu valideren Prävalenzschätzungen als nichtsymmetrische Modelle führen (z.B. Ostapczuk, Moshagen, et al., 2009). Zusammenfassend könnten nach Yu et al. (2008) die klar verständlichen Instruktionen des CWM dazu führen, dass Teilnehmer der Befragungstechnik vertrauen und auf sensible Fragen ehrlich antworten. Die Symmetrie der Antwortoptionen könnte darüber hin-

aus eine Schätzung von Instruktionsverweigerern (wie beispielsweise beim CDM oder SLD) entbehrlich werden lassen.

Da das CWM mathematisch äquivalent mit dem ursprünglichen RRT-Modell von Warner (1965) ist (vgl. Ulrich et al., 2012; Yu et al., 2008), können eine Maximum-Likelihood-Schätzung für die Prävalenz π und die Varianz des Schätzers anhand der bereits dargestellten Gleichungen (1) und (2) bestimmt werden (siehe S. 11); leicht abweichend entspricht hierbei p der bekannten Prävalenz des nicht-sensiblen Merkmals und n' der absoluten Anzahl von Teilnehmern, die die erste Antwortoption gewählt haben („beide Aussagen treffen zu oder keine der beiden Aussagen trifft zu"). Eine Darstellung des CWM als multinomiales Modell kann Abbildung 2 entnommen werden.

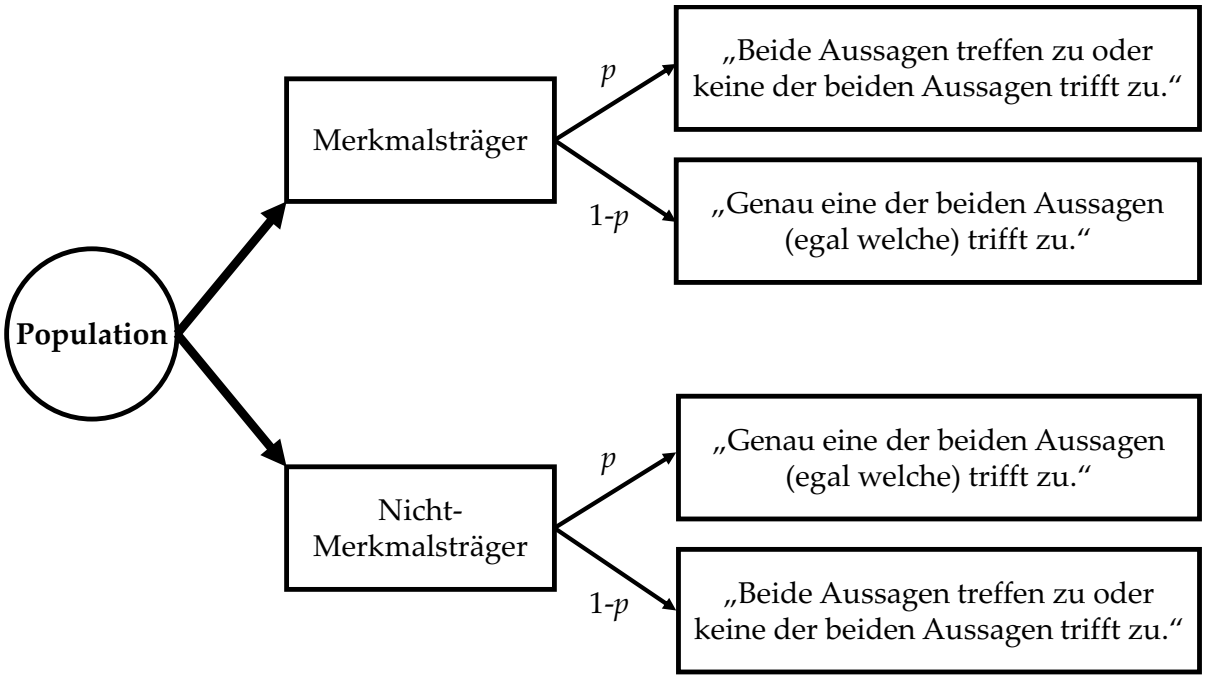

Abbildung 2: Multinomiales Modell des Crosswise-Modells (Yu et al., 2008).

Bisher sind der Literatur nur wenige Veröffentlichungen zu entnehmen, in welchen die Validität des CWM empirisch überprüft wurde. In zwei aktuellen Arbeiten wurde das CWM ohne Vergleich zu einer direkten Frage in Umfragen zu sensiblen Merkmalen mit unbekannter Prävalenz angewandt (Eslami et al., 2013; Vakilian,

Mousavi & Keramat, 2014); diese lassen somit keinerlei Ableitungen zur Validität des Verfahrens zu. Weitere Studien konnten jedoch durch Methodenvergleiche zeigen, dass anhand des CWM höhere und damit potentiell validere Prävalenzschätzungen für sensible Merkmale gewonnen werden können als anhand einer direkten Frage – so beispielsweise für Plagiate in studentischen Abschlussarbeiten (Coutts, Jann, Krumpal & Näher, 2011; Jann, Jerke & Krumpal, 2012) und für die Verwendung illegaler Dopingmittel unter Sportlern (Nakhaee, Pakravan & Nakhaee, 2013). Das CWM wurde somit bisher als eine geeignete Methode zur Reduktion des Einflusses sozialer Erwünschtheit in Umfragen evaluiert (Jann et al., 2012).

Wenngleich die dargestellten Befunde zur Validität des CWM vielversprechend sind, wurde die Vermutung der Autoren, dass das Modell mit einer vergleichsweise besonders hohen Verständlichkeit und daraus resultierenden empfundenen Vertraulichkeit einhergeht (Yu et al., 2008), bisher nicht empirisch überprüft. Darüber hinaus konnte der wissenschaftlichen Literatur zum Zeitpunkt der Anfertigung dieser Dissertation weder ein Methodenvergleich des CWM mit einer indirekten Befragungstechnik mit Verweigererdetektion noch eine starke Validierungsstudie des CWM entnommen werden. Im folgenden Abschnitt werden die Forschungsfragen zusammengefasst, deren Beantwortung das Ziel der vorliegenden Arbeit darstellte.

4 Fragestellungen

Als eine vielversprechende, aktuelle Weiterentwicklung der RRT (Warner, 1965) bietet das CWM (Yu et al., 2008) gegenüber Vorgängermodellen vereinfachte Instruktionen und den spezifischen Vorteil symmetrischer Antwortkategorien. Diese Modelleigenschaften könnten über eine bessere Verständlichkeit, eine höhere subjektiv empfundene Vertraulichkeit und eine damit einhergehende Vermeidung von *response hazards* zu Prävalenzschätzungen führen, deren Validität vergleichbar oder sogar höher ausfällt als bei konkurrierenden indirekten Befragungstechniken. Die vorliegende Dissertation hatte zum Ziel, die Validität des CWM einer ausführlichen empirischen Überprüfung zu unterziehen.

Erstens sollte zunächst die Validität des CWM im Rahmen einer schwachen Validierungsstudie überprüft und mit der Validität eines konkurrierenden RRT-Modells mit Verweigererdetektion verglichen werden. Hierzu wurden den Versuchsteilnehmern in Experiment 1 zu zwei sensiblen Merkmalen Fragen in den Formaten des CWM, des SLD (Moshagen et al., 2012) und einer konventionellen direkten Frage vorgelegt. Dem *more-is-better*-Kriterium folgend sollten hier höhere Prävalenzschätzungen als valider gedeutet werden.

Zweitens sollte in der vorliegenden Arbeit erstmalig quantifiziert werden, wie verständlich die Instruktionen des CWM tatsächlich sind, und wie hoch die subjektiv empfundene Vertraulichkeit bei Anwendung dieses Modells ausfällt. Hierzu wurde das CWM in Experiment 2 hinsichtlich der Verständlichkeit und der subjektiv empfundenen Vertraulichkeit drei konkurrierenden indirekten Befragungstechniken und einer direkten Frage gegenübergestellt.

Drittens sollte im Rahmen einer starken Validierung, die über einen reinen *more-is-better*-Ansatz hinausgeht, überprüft werden, ob Prävalenzschätzungen auf Basis des CWM mit dem wahren Anteil von Merkmalsträgern konvergieren, sofern

dieser innerhalb der Stichprobe bekannt ist. In Experiment 1 wurde zu diesem Zweck zunächst untersucht, ob das Modell die bekannte Prävalenz eines nicht-sensiblen Kontrollmerkmals (Anfangsbuchstabe des Nachnamens) in einer adäquaten Schätzung abbildet. Sofern Befragte die Instruktionen der verwendeten Befragungstechniken tatsächlich verstehen, sollten die Schätzungen mit der bekannten Prävalenz dieses Kontrollmerkmals übereinstimmen. In Experiment 3 wurde schließlich die bekannte Prävalenz eines experimentell erzeugten sensiblen Merkmals als objektives Außenkriterium für die Validität des Modells herangezogen. Sofern das CWM tatsächlich eine geeignete Methode darstellt, den Einfluss sozialer Erwünschtheit in Umfragen zu kontrollieren, sollte die Schätzung auf Basis des CWM höher als die Schätzung auf Basis einer direkten Frage ausfallen und möglichst genau mit der bekannten Prävalenz des sensiblen Merkmals korrespondieren.

5 Zusammenfassung der Einzelarbeiten

Im Folgenden werden die Methodik und die Ergebnisse der durchgeführten Experimente 1 bis 3 zusammengefasst und jeweils kurz diskutiert. Wie unter anderem in Moshagen, Hilbig und Musch (2011), Moshagen et al. (2012) und Ostapczuk et al. (2011) beschrieben, wurden die in den Experimenten 1 und 3 verwendeten Modelle als multinomiale Verarbeitungsbaummodelle (Batchelder, 1998; Batchelder & Riefer, 1999) reformuliert. Auf Basis der Modellgleichungen und der empirisch beobachteten Antworthäufigkeiten sowie mit Hilfe des in der Software multiTree (Moshagen, 2010) implementierten EM-(Expectation-Maximization-)Algorithmus (Dempster, Laird & Rubin, 1977; Hu & Batchelder, 1994) können so innerhalb der jeweiligen experimentellen Bedingungen Maximum-Likelihood-Schätzungen für die Prävalenz der betreffenden Merkmale gewonnen werden. Ein Test der Modellpassung anhand der asymptotisch χ^2-verteilten Log-Likelihood-Statistik G^2 verbleibt für die nicht restringierten Basismodelle grundsätzlich ohne Aussagekraft ($G^2 = 0$), da diese saturiert sind und somit keine Freiheitsgrade aufweisen (die Anzahl der unabhängigen Antwortkategorien entspricht der Anzahl der zu schätzenden Parameter). Unterschiede zwischen Modellparametern können jedoch anhand von Unterschieden in der Modellpassung (ΔG^2) beim Hinzufügen von Parameterrestriktionen auf Signifikanz getestet werden (z.B. $\pi_{CWM} = \pi_{direkte\ Frage}$). Für alle weiteren statistischen Analysen einschließlich der inferenzstatistischen Auswertung der Daten in Experiment 2 wurde das Programm IBM SPSS Statistics 22 verwendet.

5.1 Experiment 1: Methodenvergleich des CWM mit einem konkurrierenden RRT-Modell

In Experiment 1 wurden das CWM und ein konkurrierendes RRT-Modell mit Verweigererdetektion, der SLD (Moshagen et al., 2012), hinsichtlich ihrer Validität mit-

einander verglichen und einer konventionellen direkten Frage gegenübergestellt. Hierzu wurden Versuchsteilnehmer in einem gekreuzten Innersubjektdesign zu zwei sensiblen Merkmalen mit unbekannter Prävalenz und einem nicht-sensiblen Kontrollmerkmal mit bekannter Prävalenz befragt.

Als erstes sensibles Merkmal wurden Vorurteile gegenüber Personen mit Migrationshintergrund (*Xenophobie*) gewählt. Zur Verbreitung von Xenophobie in Deutschland zeigte sich in bisherigen Studien mitunter eine starke Diskrepanz zwischen selbstberichteten Einstellungen und tatsächlichem Verhalten: Während Umfragen auf Basis direkter Selbstauskünfte darauf hindeuten, dass die Verbreitung xenophober Einstellungen in Deutschland moderat ausgeprägt und vergleichbar mit der Verbreitung in anderen westeuropäischen Ländern ist (Zick, Küpper & Hövermann, 2011), stellten Klink und Wagner (1999) in einer Reihe von Feldexperimenten fest, dass Angehörige ethnischer Minoritäten mitunter starker Diskriminierung durch deutsche Versuchsteilnehmer ausgesetzt waren. So erhielten Versuchshelfer, deren Name, Akzent und äußere Erscheinung beispielsweise auf einen türkischen Migrationshintergrund hindeuteten, viel seltener Hilfe in Alltagssituationen als deutsche Versuchshelfer. Hjerm (1998) äußerte bereits den Verdacht, dass die Prävalenz xenophober Einstellungen durch direkte Selbstauskünfte aufgrund des Einflusses sozialer Erwünschtheit systematisch unterschätzt werden könnte. Diese Vermutung wird durch die Ergebnisse mehrerer Studien gestützt, in denen RRT-Fragen eine höhere Prävalenzschätzung für xenophobe Einstellungen ergaben als direkte Fragen (Krumpal, 2012; Ostapczuk, Musch, et al., 2009). In Anlehnung an ein Item aus der *Social Distance Scale* (Bogardus, 1933; in ähnlicher Form eingesetzt beispielsweise von Jimenez, 1999; Ostapczuk, Musch, et al., 2009; Silbermann & Hüsers, 1995) wurde somit folgende Frage zur Erfassung des ersten sensiblen Merkmals, Xenophobie, gewählt: „Würde es Sie stören, wenn Ihre 20 Jahre alte Tochter eine Beziehung mit einem Türken eingehen würde?“

Das zweite sensible Merkmal in Experiment 1 stellten Vorurteile gegenüber religiösen Minoritäten, konkret Mitbürgern islamischen Glaubens, dar (*Islamophobie*). Islamophobie ist in europäischen Ländern weit verbreitet (z.B. EUMC - European Monitoring Center on Racism and Xenophobia, 2006; Savelkoul, Scheepers, van der Veld & Hagendoorn, 2012; Sheridan, 2006; Zick et al., 2011) und wird als eines der wichtigsten politischen Themen diskutiert (Bunzl, 2005). Im europäischen Vergleich schneidet Deutschland als eines der Länder mit der höchsten Verbreitung antimuslimischer Vorurteile ab (Zick et al., 2011), obwohl die deutsche Verfassung Religionsfreiheit als Grundrecht definiert (Grundgesetz für die Bundesrepublik Deutschland, Art. 3, Abs. 3). Imhoff und Recker (2012) konnten außerdem eine starke Assoziation islamophober Einstellungen mit negativen Einstellungen zur Errichtung muslimisch-religiöser Bauwerke (konkret einer Moschee in Köln) nachweisen. Schließlich zeigte sich bei einer Volksabstimmung zum Verbot des Neubaus von Minaretten in der Schweiz im Jahr 2009, dass repräsentative Umfragen im Vorfeld eine Ablehnung des Minarettverbots prognostizierten (gfs.bern, 2009a, 2009b; reformiert, 2009); tatsächlich stimmte in der Volksabstimmung jedoch die Mehrheit für ein solches Minarettverbot, welches im Anschluss in die Schweizer Verfassung aufgenommen wurde. Als mögliche Erklärung für die Diskrepanz zwischen der Prognose und dem tatsächlichen Ergebnis wurde diskutiert, dass eine offene Befürwortung des Minarettverbots durch die mediale Berichterstattung womöglich als Indikator für Islamfeindlichkeit und damit als stigmatisierendes Merkmal wahrgenommen wurde (fög, 2010). Ein Grund für die Unterschätzung durch die vorangehenden Umfragen könnte somit in sozial erwünschtem Antwortverhalten auf eine direkte Frage gefunden werden. Ausgehend von diesen Befunden wurde entschieden, folgende Frage zur Erfassung des zweiten sensiblen Merkmals, Islamophobie, in die Studie aufzunehmen: „Ich bin dafür, in Deutschland den Bau von Minaretten zu verbieten."

Unter anderem von Umesh und Peterson (1991) wurde gefordert, dass Validierungsstudien für indirekte Befragungstechniken über einen *more-is-better*-Ansatz

hinausgehen sollten; hierzu müssten Merkmale untersucht werden, deren Prävalenz in der Population (oder besser noch in der Stichprobe) bekannt ist und als objektives Außenkriterium für die Validität der Prävalenzschätzungen herangezogen werden kann. Da die Erfassung der Prävalenz sensibler Merkmale jedoch häufig sehr aufwändig oder sogar unmöglich ist, wurden bisher nur verhältnismäßig wenige solcher starken Validierungsstudien durchgeführt (Lensvelt-Mulders et al., 2005). Als erster Schritt einer starken Validierung des CWM wurde deshalb in Experiment 1 zunächst ein nicht-sensibles Kontrollmerkmal herangezogen, dessen Prävalenz in der deutschen Bevölkerung verhältnismäßig einfach zu bestimmen war. Nach Angaben des Statistischen Bundesamtes beginnt der Nachname von etwa 43% der in Deutschland lebenden Personen mit einem der Buchstaben K, L, M, R, S oder T (Reinders, 1996). Sofern das CWM und / oder der SLD zu validen Prävalenzschätzungen für sensible Merkmale führen, sollten diese Modelle auch die Prävalenz dieses nicht-sensiblen Kontrollmerkmals punktgenau schätzen können. Für die Erfassung des nicht-sensiblen Merkmals mit bekannter Prävalenz wurde deshalb folgende Frage in die Untersuchung aufgenommen: „Mein Nachname beginnt mit einem der folgenden Buchstaben: K, L, M, R, S, T.“

Insgesamt 1312 Studierende (56% weiblich; M_{Alter} = 21.21 Jahre, SD_{Alter} = 3.14) bekamen unmittelbar vor Lehrveranstaltungen an den Universitäten Düsseldorf (81%), Duisburg (10%) und Bochum (8%) einen einseitigen Papier-Bleistift-Fragebogen ausgehändigt, der sofort ausgefüllt und wieder abgegeben werden sollte. Dieser Fragebogen enthielt aufeinanderfolgend jeweils eine Frage zum sensiblen Merkmal 1 (Xenophobie), zum sensiblen Merkmal 2 (Islamophobie) und zum nicht-sensiblen Kontrollmerkmal (Anfangsbuchstabe des Nachnamens). Die Befragungstechnik wurde als Innersubjektfaktor mit drei Stufen (CWM, SLD, direkte Frage) zufällig den einzelnen Fragen zugewiesen, mit der Vorgabe, dass alle drei Fragen in unterschiedlichen Formaten beantwortet werden sollten.

Die Prävalenzschätzungen für die beiden sensiblen Merkmale und das nicht-sensible Kontrollmerkmal, die in den experimentellen Bedingungen auf Basis der empirisch beobachteten Antworthäufigkeiten bestimmt wurden, können Tabelle 1 entnommen werden.

Tabelle 1

Prävalenzschätzungen für die beiden sensiblen Merkmale und das nicht-sensible Kontrollmerkmal in den unterschiedlichen experimentellen Bedingungen

Merkmal	**Bedingung**	$\hat{\pi}$ (*SE*)	$\hat{t}$ (*SE*)
Sensibel 1 (Xenophobie)	Direkte Frage	26.98% (2.11)	–
	CWM	48.67% (3.48)	–
	SLD	53.38% (6.31)	79.43% (4.59)
Sensibel 2 (Islamophobie)	Direkte Frage	43.33% (2.40)	–
	CWM	51.64% (3.46)	–
	SLD	76.93% (6.62)	67.94% (3.19)
Nicht-sensibel (Anfangsbuchstabe des Nachnamens)	Direkte Frage	40.99% (2.33)	–
	CWM	46.57% (3.54)	–
	SLD	62.72% (6.25)	78.23% (3.92)

Anmerkung: Die Prävalenz des nicht-sensiblen Kontrollmerkmals in der deutschen Bevölkerung liegt nach Angaben des Statistischen Bundesamtes bei etwa 43% (Reinders, 1996).

Paarweise Vergleiche der Prävalenzschätzungen für das sensible Merkmal 1 (Xenophobie) zeigten, dass die Schätzung durch eine direkte Frage signifikant niedriger als die Schätzung durch das CWM ($\Delta G^2[1] = 28.20$, $p < .001$) und die Schätzung durch den SLD ($\Delta G^2[1] = 16.80$, $p < .001$) ausfiel. Die Schätzungen der beiden indirekten Befragungstechniken unterschieden sich nicht signifikant voneinander ($\Delta G^2[1] = 0.43$, $p = .51$). Der Anteil ehrlich antwortender Merkmalsträger (t) wurde durch den SLD signifikant niedriger als ein Referenzwert von 100% geschätzt ($\Delta G^2[1] = 14.56$, $p < .001$). Auch für das sensible Merkmal 2 (Islamophobie) wurde die Prävalenz durch das CWM ($\Delta G^2[1] = 3.89$, $p < .05$) und den SLD ($\Delta G^2[1] = 23.97$, $p < .001$) höher

geschätzt als durch eine direkte Frage. Die Schätzung durch das CWM fiel hier jedoch signifikant niedriger als die Schätzung durch den SLD aus ($\Delta G^2[1] = 11.80$, $p < .001$). Der geschätzte Anteil ehrlich antwortender Merkmalsträger in der SLD-Bedingung (t) unterschied sich auch hier signifikant von 100% ($\Delta G^2[1] = 65.07$, $p < .001$). Schließlich wurden für das nicht-sensible Kontrollmerkmal erwartungsgemäß keine signifikanten Unterschiede zwischen der Schätzung durch eine direkte Frage und der Schätzung durch das CWM ($\Delta G^2[1] = 1.73$, $p = .19$), der Schätzung durch eine direkte Frage und der bekannten Prävalenz von 43% ($\Delta G^2[1] = 0.73$, $p = .39$) sowie der Schätzung durch das CWM und der bekannten Prävalenz ($\Delta G^2[1] = 1.02$, $p = .31$) ermittelt. Entgegen den Erwartungen fiel die Schätzung durch den SLD jedoch signifikant höher aus als die Schätzung durch eine direkte Frage ($\Delta G^2[1] = 11.00$, $p < .001$), die Schätzung durch das CWM ($\Delta G^2[1] = 5.15$, $p < .05$) und die bekannte Prävalenz ($\Delta G^2[1] = 10.42$, $p < .01$). Ebenfalls unerwartet fiel der Test des Anteils ehrlich antwortender Merkmalsträger (t) aus, der wiederum einen signifikanten Unterschied vom Referenzwert (100%) ergab ($\Delta G^2[1] = 23.16$, $p < .001$).

Die dargestellten Ergebnisse legen nahe, dass die Prävalenzen der sensiblen Merkmale 1 (Xenophobie) und 2 (Islamophobie) durch eine direkte Frage unterschätzt wurden – bzw. dass beide indirekte Befragungstechniken (CWM und SLD) entsprechend dem *more-is-better*-Kriterium eine gegenüber einer direkten Frage erhöhte Validität aufwiesen. Das CWM führte auch bei der Erfassung des nicht-sensiblen Kontrollmerkmals zu einer adäquaten Schätzung der in der Population bekannten Prävalenz. Zusammenfassend konnten für dieses Modell also vielversprechende Ergebnisse sowohl im Sinne einer schwachen Validierung als auch in einem ersten Schritt einer starken Validierung auf Basis eines nicht-sensiblen Merkmals ermittelt werden. Die Anwendung des SLD resultierte hingegen in einer substantiellen Überschätzung der Prävalenz des nicht-sensiblen Kontrollmerkmals. Dieser problematische Befund legt nahe, dass sich nicht alle Teilnehmer in der SLD-Bedingung entsprechend den Instruktionen verhalten haben. Das Nichtbefolgen der Instruktio-

nen ist hier jedoch kaum anhand von *response hazards* zu erklären, da das abgefragte Merkmal nicht sensibel war und deshalb kein Grund für die Suche nach einer sicheren Antwortalternative gegeben war. Plausibel scheint hingegen, dass einige Versuchsteilnehmer die verhältnismäßig komplexen Instruktionen des SLD schlicht nicht verstanden haben und folglich Probleme bei der Identifikation der korrekten Antwortalternative hatten. Zur Überprüfung dieser Vermutung und zur weiterführenden Validierung des CWM wurden die beiden in Experiment 1 verwendeten Modelle sowie zwei weitere indirekte Befragungstechniken in Experiment 2 hinsichtlich ihrer Verständlichkeit und der subjektiv empfundenen Vertraulichkeit untereinander und mit einer direkten Frage verglichen.

5.2 Experiment 2: Verständlichkeit und subjektiv empfundene Vertraulichkeit indirekter Befragungstechniken

In indirekten Befragungen wird meist davon ausgegangen, dass Teilnehmer die Instruktionen verstehen und den erhöhten Schutz der Vertraulichkeit ihrer Antwort als solchen erkennen (Abul-Ela et al., 1967; Edgell et al., 1982; Franklin, 1998; Warner, 1965). Diese beiden Variablen, das Instruktionsverständnis und die subjektiv empfundene Vertraulichkeit, stellen zwei zentrale psychologische Aspekte bei der Ermittlung valider Prävalenzschätzungen auf Basis indirekter Fragen dar (Landsheer, van der Heijden & van Gils, 1999; Zdep & Rhodes, 1977). Zur Erfassung dieser Variablen wurden verschiedene Ansätze vorgestellt: Erstens wurde versucht, das Verständnis und die empfundene Vertraulichkeit anhand von Rücklaufquoten und Antwortraten zu quantifizieren (z.B. Boruch, 1972; Chi, Chow & Rider, 1972; Coutts & Jann, 2011; Goodstadt & Gruson, 1975), wobei höhere Rücklaufquoten und Antwortraten meist mit einer höheren Ausprägung auf diesen beiden Variablen in Verbindung gebracht wurden. Zweitens wurde in Studien, in denen der Merkmalsstatus der Befragten bekannt war, ein niedriger Anteil falscher Antworten als Anzeichen für ein hohes Ver-

ständnis und eine hohe empfundene Vertraulichkeit gedeutet (z.B. Edgell et al., 1992; Edgell et al., 1982). Diese beiden Ansätze erlauben jedoch nur eine relativ indirekte Erfassung der interessierenden Größen, da der differentielle Einfluss von Verständnis und empfundener Vertraulichkeit auf das Antwortverhalten nicht ermittelt werden kann. Eine alternative Strategie kann in der direkten Abfrage von Verständnis und empfundener Vertraulichkeit durch Selbst- oder Fremdeinschätzungen von Befragten oder Befragenden gefunden werden (z.B. Coutts & Jann, 2011; Locander et al., 1976; Miller, 1984; van der Heijden, van Gils, Bouts & Hox, 1998). Besonders bezüglich des Instruktionsverständnisses stellen diese Einschätzungen jedoch subjektive Maße dar, die nicht gezwungenermaßen mit der tatsächlichen Verständlichkeit der jeweiligen Methode korrespondieren müssen. Außerdem existieren zum Einfluss des Bildungsniveaus der Teilnehmer auf das Instruktionsverständnis widersprüchliche Befunde: Einigen Studien sind Hinweise auf eine mögliche positive Assoziation zu entnehmen (z.B. Abul-Ela et al., 1967; Chi et al., 1972), andere Studien fanden keinen (Landsheer et al., 1999) oder sogar einen unerwartet negativen Zusammenhang (Holbrook & Krosnick, 2010). Schließlich lagen für viele aktuelle indirekte Befragungstechniken bisher keine Daten zur Verständlichkeit, zur subjektiv empfundenen Vertraulichkeit oder zu einem möglichen Einfluss des Bildungsniveaus vor; vier dieser Befragungstechniken wurden in Experiment 2 implementiert: das CWM, der SLD, das CDM (alle bereits dargestellt) und ein verwandtes Verfahren, die *Unmatched-Count-Technik* (UCT; Miller, 1984). In Befragungen mit der UCT werden zwei unabhängige Gruppen untersucht: Einer ersten (Experimental-) Gruppe wird eine Liste mit mehreren Fragen zu nicht-sensiblen Merkmalen, sowie einer Frage zu dem interessierenden sensiblen Merkmal vorgelegt. Einer zweiten (Kontroll-) Gruppe werden nur die nicht-sensiblen Fragen präsentiert. Versuchsteilnehmer sollen nun lediglich angeben, wie viele der Fragen sie mit „ja“ beantworten, unabhängig davon, welche der Aussagen bejaht werden. Sofern durch eine angemessene Auswahl der Art und Anzahl nicht-sensibler Fragen vermieden wird, dass Teilnehmer keine oder

alle der präsentierten Aussagen bejahen müssen, bleibt die Vertraulichkeit individueller Antworten gewahrt (Erdfelder & Musch, 2006; Fox, J. A. & Tracy, 1986). Anhand der Differenz in der mittleren Anzahl berichteter „ja"-Antworten zwischen Experimental- und Kontrollgruppe kann schließlich eine Schätzung für die Prävalenz des sensiblen Merkmals gewonnen werden. Auch anhand der UCT konnten bereits mehrfach höhere und damit potentiell validere Prävalenzschätzungen erreicht werden als anhand einer direkten Frage (z.B. Ahart & Sackett, 2004; Coutts & Jann, 2011; LaBrie & Earleywine, 2000; Wimbush & Dalton, 1997). Darüber hinaus existieren Hinweise auf eine mögliche Überlegenheit der UCT gegenüber einer Forced-Response-Variante der RRT bezüglich der Verständlichkeit und subjektiv empfundenen Vertraulichkeit, die jedoch aus möglicherweise verzerrten Selbsteinschätzungen der Befragten gewonnen wurden (Coutts & Jann, 2011). Studien, die eine objektive Erfassung der Verständlichkeit und einen Vergleich mit den zuvor genannten aktuellen RRT-Modellen zum Ziel hatten, waren der Literatur zum Zeitpunkt der Anfertigung dieser Arbeit nicht zu entnehmen. Das Ziel von Experiment 2 bestand zusammenfassend darin, vier aktuelle indirekte Befragungstechniken hinsichtlich der Verständlichkeit der Instruktionen und der subjektiv empfundenen Vertraulichkeit untereinander zu vergleichen und einer direkten Frage gegenüberzustellen, sowie einen potentiell moderierenden Einfluss des Bildungsniveaus zu erfassen.

Zu diesem Zweck wurde 401 Teilnehmern im Alter von 25 bis 35 Jahren (53% weiblich; M_{Alter} = 30.72 Jahre, SD_{Alter} = 3.34) ein szenariobasierter Online-Fragebogen dargeboten. In einem 5x2 quasi-experimentell-gemischten Design wurde zum einen der Innersubjektfaktor Befragungstechnik in den Stufen direkte Frage (Referenz), CWM, SLD, CDM und UCT in randomisierter Reihenfolge variiert. Als zweiter, quasi-experimenteller Zwischensubjektfaktor wurde das Bildungsniveau der Teilnehmer in den Extremgruppen niedrige Bildung (höchstens Hauptschulabschluss) und hohe Bildung (mindestens Fachabitur) berücksichtigt. Den Teilnehmern wurden zunächst kurze Personenbeschreibungen der vier fiktiven Personen Wil-

helm, Hans, Ernst und Ludwig dargeboten. Über diese vier Personen wurden alle möglichen Merkmalskombinationen in Bezug auf ein sensibles Merkmal (Betrug in Prüfungen; bereits mehrfach in Studien mit indirekten Befragungstechniken implementiert, z.B. Hejri et al., 2013; Ostapczuk, Moshagen, et al., 2009; Scheers & Dayton, 1987) und ein nicht-sensibles Merkmal (Geburtsmonat im November oder Dezember, Zufallsgenerator für CWM, SLD und CDM) realisiert. Zwei weitere, nicht-sensible Merkmale wurden konstant gehalten (Geschlecht männlich, noch nie die Stadt London besucht) und ermöglichten die Anwendung der UCT. Innerhalb jedes Fragetechnik-Blocks wurden den Teilnehmern alle vier fiktiven Charaktere in randomisierter Reihenfolge dargeboten. Um eine objektive Messung der ersten abhängigen Variablen Verständlichkeit zu erreichen, wurde diese als relativer Anteil korrekter Antworten über die vier Charaktere auf die Frage „Was müsste [Wilhelm; Hans; Ernst; Ludwig] hier antworten?“ operationalisiert. Versuchsteilnehmer konnten somit pro Befragungstechnik einen Wert zwischen 0% (wenn für keinen der vier Charaktere die jeweils korrekte Antwort identifiziert wurde) und 100% (wenn für alle vier Charaktere die jeweils korrekte Antwort identifiziert wurde) erzielen. Die Messung der zweiten abhängigen Variablen (subjektiv empfundene Vertraulichkeit) erfolgte über die Frage „Wie gut geschützt ist [Wilhelm; Hans; Ernst; Ludwig] Ihrer Meinung nach davor, dass wir herausfinden, ob er tatsächlich schon einmal in einer Prüfung betrogen hat?“ mit einem siebenstufig likert-skalierten Antwortformat von -3 („gar nicht geschützt“) bis +3 (sehr gut geschützt). Hierbei wurde pro Versuchsteilnehmer das arithmetische Mittel der Bewertungen über alle vier fiktiven Charaktere gebildet. Einen Screenshot aus dem Online-Fragebogen mit einem Beispiel für die Abfrage der beiden abhängigen Variablen im Format des CWM für den fiktiven Charakter Wilhelm zeigt Abbildung 3.

Abbildung 3: Screenshot aus dem Online-Fragebogen von Experiment 2, der die Abfrage der beiden abhängigen Variablen beispielhaft für das CWM und den fiktiven Charakter Wilhelm zeigt.

Eine 5x2-gemischt-faktorielle ANOVA für die Variable Verständlichkeit zeigte einen signifikanten Haupteffekt des Faktors Befragungstechnik (F[4,1596] = 75.46, p < .001, η^2 = .16), einen signifikanten Haupteffekt des Faktors Bildungsniveau (F[1,399] = 17.07, p < .001, η^2 = .04) und eine signifikante Interaktion dieser Faktoren (F[4,1596] = 4.13, p < .001, η^2 = .01). Eine Darstellung der mittleren Verständlichkeit in Abhängigkeit von der Befragungstechnik für die Gesamtstichprobe und getrennt nach Bildungsgruppen kann den Abbildungen 4 und 5 entnommen werden.

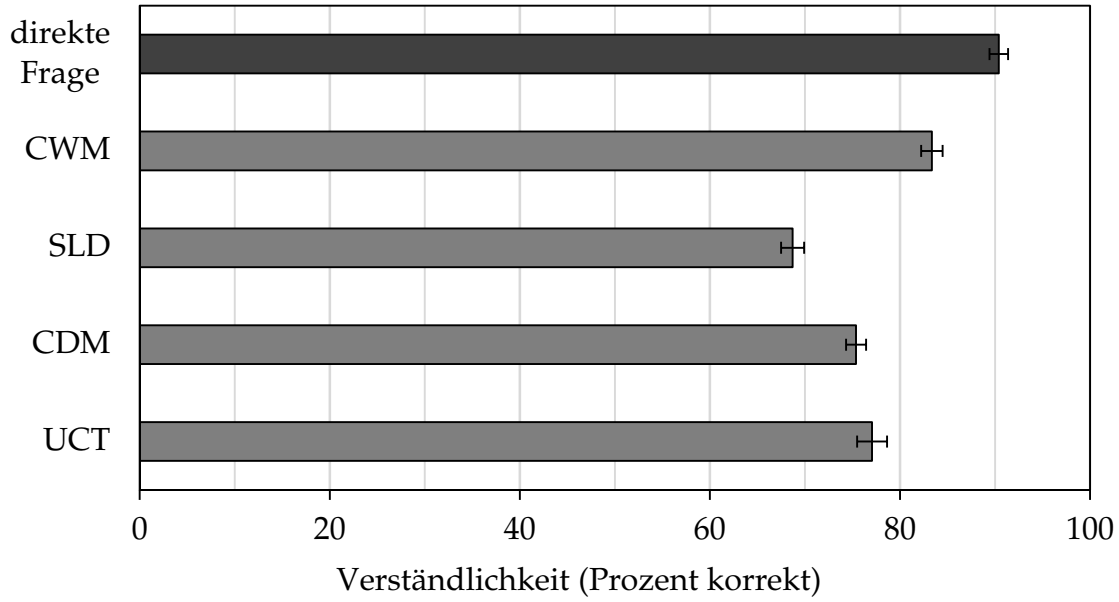

Abbildung 4: Mittlere Verständlichkeit in Abhängigkeit von der Befragungstechnik für die Gesamtstichprobe (die Fehlerbalken repräsentieren den Standardfehler).

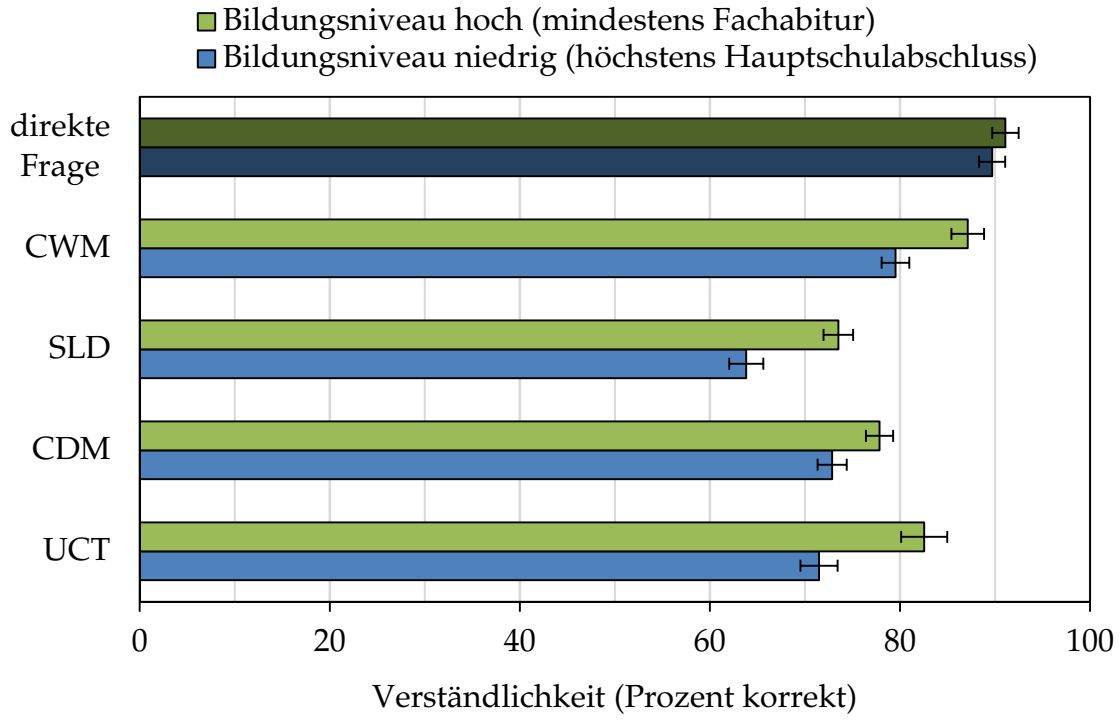

Abbildung 5: Mittlere Verständlichkeit in Abhängigkeit von der Befragungstechnik, getrennt nach dem Bildungsniveau der Teilnehmer (die Fehlerbalken repräsentieren den Standardfehler).

Paarvergleiche zwischen den Befragungstechnik-Stufen anhand von Bonferroni-post-hoc-Tests ergaben, dass die mittlere Verständlichkeit einer direkten Frage signifikant höher ausfiel als die mittlere Verständlichkeit des CWM ($p < .001$; $r = .49$, $dz = 0.33$ nach Cohen, 1988), des SLD ($p < .001$; $r = .23$, $dz = 0.79$), des CDM ($p < .001$; $r = .44$, $dz = 0.70$) und der UCT ($p < .001$; $r = .52$, $dz = 0.49$). Alle Vergleiche zwischen den indirekten Befragungstechniken fielen signifikant aus (alle $p < .001$; CDM vs. SLD: $r = .38$, $dz = 0.26$; CDM vs. CWM: $r = .39$, $dz = 0.33$; SLD vs. CWM: $r = .29$, $dz = 0.52$; SLD vs. UCT: $r = .25$, $dz = 0.24$; CWM vs. UCT: $r = .52$, $dz = 0.23$), mit Ausnahme des Vergleichs des CDM und der UCT ($p > .99$; $r = .42$, $dz = 0.06$). Erwartungsgemäß zeigten höher gebildete Teilnehmer insgesamt ein höheres Verständnis als niedriger gebildete Teilnehmer. Anhand fünf paariger t-Tests innerhalb der Befragungstechnik-Stufen auf einem Bonferroni-korrigierten Alphafehler-Niveau (korrigiertes $\alpha = .05 / 5 = .01$) konnte darüber hinaus gezeigt werden, dass kein Einfluss des Bildungsniveaus auf die Verständlichkeit einer direkten Frage bestand ($t[399] = -0.71$, $p = .48$; $d = 0.07$). Bei allen indirekten Befragungstechniken zeigte sich für niedriger gebildete Teilnehmer jedoch eine geringere mittlere Verständlichkeit (CWM: $t[399] = -3.39$, $p < .001$, $d = 0.34$; SLD: $t[399] = -4.07$, $p < .001$, $d = 0.41$; UCT: $t[399] = -3.56$, $p < .001$, $d = 0.36$), wobei dieser Test für das CDM auf dem korrigierten Alphafehler-Niveau knapp nicht signifikant ausfiel ($t[399] = -2.37$, $p = .02$; $d = 0.24$).

Anhand einer weiteren 5x2-gemischt-faktoriellen ANOVA für die Variable subjektiv empfundene Vertraulichkeit konnte ein signifikanter Haupteffekt des Faktors Befragungstechnik ($F[4,1596] = 18.76$, $p < .001$, $\eta^2 = .05$) und eine signifikante Interaktion zwischen der Befragungstechnik und dem Bildungsniveau der Teilnehmer ermittelt werden ($F[4,1596] = 9.21$, $p < .001$, $\eta^2 = .02$); der Faktor Bildungsniveau zeigte keinen signifikanten Haupteffekt ($F[1,399] = 0.96$, $p = .33$, $\eta^2 < .01$). Eine Darstellung der mittleren empfundenen Vertraulichkeit in Abhängigkeit von der Befragungstechnik für die Gesamtstichprobe und getrennt nach Bildungsniveau kann den Abbildungen 6 und 7 entnommen werden.

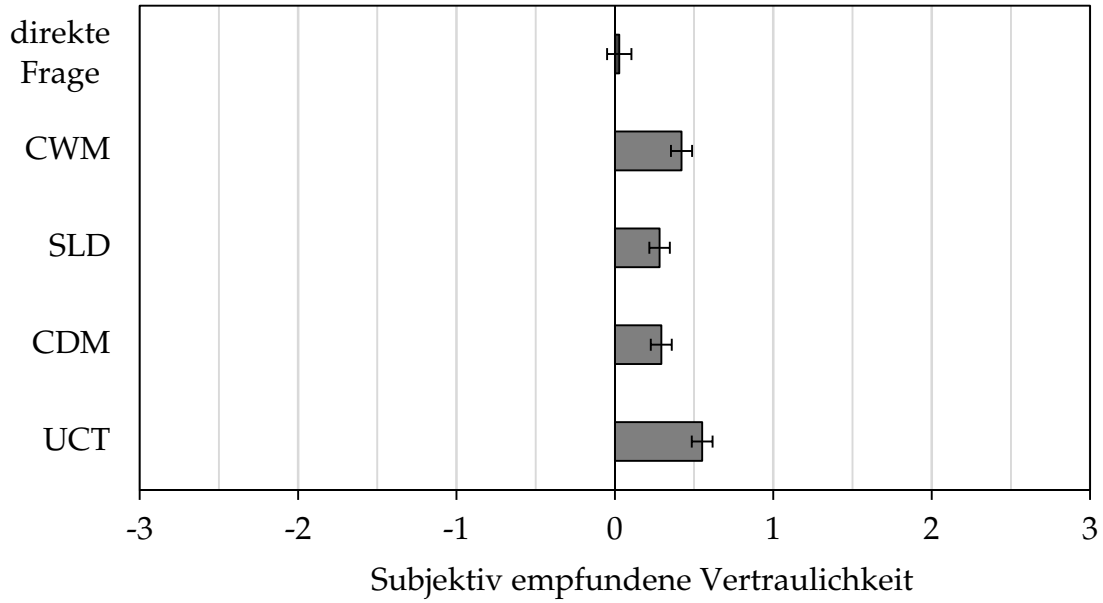

Abbildung 6: Mittlere subjektiv empfundene Vertraulichkeit in Abhängigkeit von der Befragungstechnik für die Gesamtstichprobe (die Fehlerbalken repräsentieren den Standardfehler).

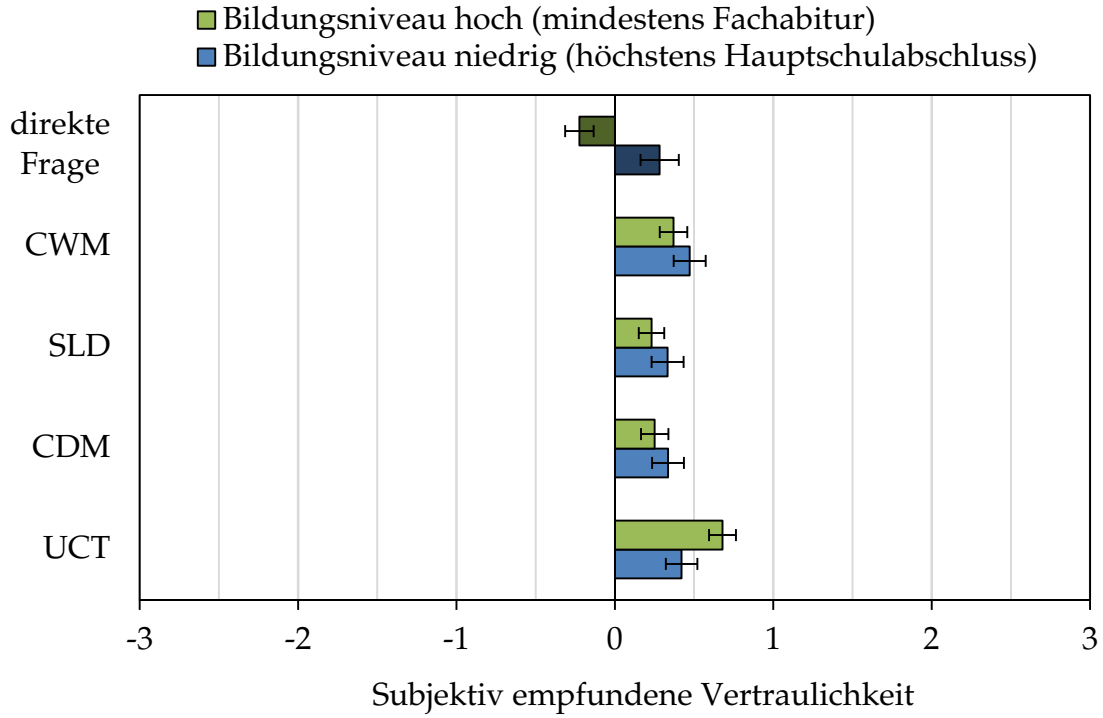

Abbildung 7: Mittlere subjektiv empfundene Vertraulichkeit in Abhängigkeit von der Befragungstechnik, getrennt nach dem Bildungsniveau der Teilnehmer (die Fehlerbalken repräsentieren den Standardfehler).

Bonferroni-post-hoc-Tests zwischen den Befragungstechnik-Stufen zeigten zunächst, dass die mittlere subjektiv empfundene Vertraulichkeit für eine direkte Frage konsistent niedriger ausfiel als für alle indirekten Befragungstechniken (CWM: $p < .001$, $r = .39$, $dz = 0.25$; SLD: $p < .01$, $r = .53$, $dz = 0.18$; CDM: $p < .001$, $r = .57$, $dz = 0.19$; UCT: $p < .001$, $r = .40$, $dz = 0.33$). Die höchste empfundene Vertraulichkeit zeigte sich für die UCT; diese war gegenüber dem CWM nur geringfügig und nicht signifikant ($p = .21$; $r = .64$, $dz = 0.12$), gegenüber dem SLD ($p < .001$; $r = .64$, $dz = 0.24$) und dem CDM ($p < .001$; $r = .61$, $dz = 0.22$) jedoch signifikant erhöht. Die übrigen Paarvergleiche verblieben auf nicht-signifikantem Niveau (CWM vs. SLD: $p = .10$; $r = .67$, $dz = 0.13$; CWM vs. CDM: $p = .31$; $r = .61$, $dz = 0.11$; SLD vs. CDM: $p > .99$; $r = .65$, $dz = 0.01$). Fünf paarige t-Tests innerhalb der einzelnen Befragungstechnik-Stufen auf einem Bonferroni-korrigierten Signifikanzniveau (korrigiertes $\alpha = .05 / 5 = .01$) zeigten nur für die direkte Frage, dass höher gebildete Teilnehmer im Mittel eine niedrigere empfundene Vertraulichkeit berichteten ($t[399] = 3.35$, $p < .001$; $d = 0.33$). Bei keiner der indirekten Befragungstechniken konnte ein Bildungseffekt festgestellt werden (CWM: $t[399] = 0.77$, $p = .44$, $d = 0.07$; SLD: $t[399] = 0.78$, $p = .43$, $d = 0.08$; CDM: $t[399] = 0.64$, $p = .53$, $d = 0.07$; UCT: $t[399] = 1.98$, $p = .05$, $d = 0.20$).

Schließlich wurde ein möglicher Zusammenhang zwischen der Verständlichkeit und der subjektiv empfundenen Vertraulichkeit für die Gesamtstichprobe und getrennt für die Bildungsgruppen anhand von Produkt-Moment-Korrelationen (*Pearsons r*) innerhalb der einzelnen Befragungstechnik-Stufen untersucht. Mit einer Ausnahme (eine geringe Korrelation in der SLD-Bedingung bei niedrig gebildeten Teilnehmern) fanden sich hierbei keine signifikanten Zusammenhänge (siehe Tabelle 2).

Tabelle 2

Produkt-Moment-Korrelationen (*Pearsons r*) zur Messung des Zusammenhangs zwischen Verständlichkeit und subjektiv empfundener Vertraulichkeit

Gruppe	**Befragungstechnik**				
	direkte Frage	CWM	SLD	CDM	UCT
Gesamtstichprobe	- .08	.04	.07	- .01	.09
Bildungsniveau hoch	- .12	.02	.03	- .09	.03
Bildungsniveau niedrig	- .01	.07	.16 *	.09	.13

Anmerkung: * $p < .05$.

Zusammenfassend zeigte das CWM gegenüber den konkurrierenden indirekten Befragungstechniken die vergleichsweise höchste Verständlichkeit. Alle indirekten Verfahren wiesen im Vergleich mit einer direkten Frage eine erhöhte Vertraulichkeit auf; das CWM schnitt hierbei mit einem der höchsten Werte ab. Die Ergebnisse zum Einfluss des Bildungsniveaus legen nahe, dass bei der indirekten Befragung weniger gebildeter Stichproben eine gezielte Überprüfung des Instruktionsverständnisses ratsam ist. Weiter scheinen weniger gebildete Teilnehmer eine verminderte Einsicht in die mit einer direkten Befragung einhergehende reduzierte Vertraulichkeit zu haben. Die fehlende Assoziation zwischen Verständnis und empfundener Vertraulichkeit unterstreicht schließlich, dass eine getrennte Erfassung dieser Konstrukte sinnvoll und erstrebenswert ist. Indirekte Befragungstechniken, die zur Abfrage sensibler Merkmale eingesetzt werden, sollten gleichermaßen verständlich sein und zu einer möglichst hohen empfundenen Vertraulichkeit führen. Da das CWM bezüglich dieser Größen aus Experiment 2 als das insgesamt vielversprechendste Verfahren hervorging, wurde dieses Modell in Experiment 3 mit Hilfe einer starken Validierungsuntersuchung einer besonders strengen Prüfung unterzogen. Diese sollte abschließend klären, ob die Anwendung des CWM tatsächlich zu validen Prävalenzschätzungen für sensible Merkmale führt.

5.3 Experiment 3: „Starke“ Validierung des CWM

In der Mehrzahl von Validierungsstudien für indirekte Befragungstechniken werden Prävalenzschätzungen für sensible Merkmale auf Basis direkter und indirekter Befragungen miteinander verglichen. Hierbei werden höhere Prävalenzschätzungen für sozial *un*erwünschte Merkmale als valider angesehen, weil diese potentiell weniger durch sozial erwünschtes Antwortverhalten verzerrt sind, was zu einer *Unter*schätzung der Prävalenz (und damit vergleichsweise niedrigeren Prävalenzschätzungen) führen sollte (*more-is-better*-Kriterium; Umesh & Peterson, 1991). Da der wahre Anteil von Trägern des sensiblen Merkmals jedoch nicht bekannt ist, liefern Befunde dieser Art nur schwache Evidenz für die Validität einer Befragungsmethode. Selbst höhere Schätzer könnten nämlich immer noch Unterschätzungen (oder schlimmstenfalls sogar Überschätzungen) der Prävalenz darstellen. Starke Validierungsstudien hingegen nutzen die bekannte Prävalenz eines sensiblen Merkmals als objektives Außenkriterium. Durch den Vergleich von Prävalenzschätzungen mit dem wahren Wert können somit eindeutige Rückschlüsse auf die Validität einer Befragungsmethode gezogen werden. Da die Durchführung starker Validierungsstudien mit hohen Kosten verbunden ist, weil die Prävalenz sensibler Merkmale in der Regel unbekannt und ihre Erhebung meist sehr aufwändig oder sogar unmöglich ist, sind starke Validierungsstudien verhältnismäßig selten (z.B. nur 6 der insgesamt 38 berücksichtigten Studien in der Metaanalyse von Lensvelt-Mulders et al., 2005). Für das CWM wurden bereits vielversprechende Ergebnisse in Bezug auf das *more-is-better*-Kriterium (z.B. Coutts et al., 2011; Jann et al., 2012; Nakhaee et al., 2013, sowie Experiment 1 in dieser Arbeit), die punktgenaue Schätzung der bekannten Prävalenz eines nichtsensiblen Merkmals (Experiment 1) sowie die Verständlichkeit der Instruktionen und die subjektiv empfundene Vertraulichkeit (Experiment 2) ermittelt; folglich wurde in Experiment 3 der Versuch einer starken Validierung dieses Modells unternommen.

Ein vielversprechender Ansatz, den üblicherweise mit starken Validierungsstudien verbundenen Problemen zu begegnen, wurde kürzlich vorgeschlagen: Moshagen et al. (2014) gaben Versuchsteilnehmern die Möglichkeit, beim Bericht der Ergebnisse von verdeckt durchgeführten Würfeldurchgängen zu lügen, um ihren finanziellen Gewinn zu maximieren. Während das Verhalten einzelner Teilnehmer den Versuchsleitern unbekannt blieb, konnte auf Basis der bekannten Verteilung von Würfelergebnissen die Prävalenz von Betrügern in der Stichprobe ermittelt werden. Anschließend wurden Teilnehmer befragt, ob sie betrogen hatten. Die bekannte Prävalenz des sensiblen Merkmals (Betrug im Würfelspiel) wurde hierbei erfolgreich als objektives Außenkriterium bei der Validierung einer indirekten Befragungstechnik verwendet. So elegant dieser Ansatz die Schwierigkeiten starker Validierungsstudien zu umgehen vermag, so schwierig erscheint seine Implementation in Online-Studien, da Würfelaufgaben dieser Art in der Regel mit einer direkten Interaktion mit einzelnen Versuchsteilnehmern verbunden sind. Online-Studien bieten jedoch optimale Rahmenbedingungen für die Erhebung großer Stichproben, die zur Kompensation der verringerten Effizienz indirekter Befragungstechniken nötig sind (Musch, Bröder & Klauer, 2001). Anknüpfend an Moshagen et al. (2014) wurde in Experiment 3 deshalb zunächst ein Paradigma entwickelt, das einer modifizierten Version der Anagrammrätsel-Aufgabe von Wiltermuth (2011) entsprach und auch in Online-Studien anwendbar sein sollte. In der ursprünglichen Version dieser Anagrammrätsel-Aufgabe wurde Versuchsteilnehmern eine Liste mit neun Anagrammen (Wörter, bei denen die Buchstabenreihenfolge durcheinandergebracht wurde) vorgegeben, für die jeweils durch Umstellung der Buchstaben in Gedanken das Zielwort identifiziert werden sollte. Diese Anagramme sollten der Reihe nach bearbeitet werden; anschließend wurde die Anzahl der aufeinanderfolgend gelösten Anagramme abgefragt, nicht jedoch die Anagrammlösungen selbst. Höhere Ergebnisse waren hier mit einer höheren finanziellen Entlohnung verknüpft. Ohne Kenntnis der Teilnehmer hatte sich das dritte Anagramm in dieser Reihe als empirisch unlösbar erwiesen, da keiner

der 30 Versuchsteilnehmer in einer Vorstudie das Zielwort hatte identifizieren können. Folglich wurden Teilnehmer, die drei oder mehr aufeinanderfolgend gelöste Anagramme angaben, als Betrüger klassifiziert.

Dieses Paradigma wurde in Experiment 3 adaptiert und nochmals vereinfacht: Versuchsteilnehmer sollten nur drei Anagramme bearbeiten, von denen zwei sicher lösbar und eines sicher unlösbar war. Die Teilnehmer sollten dann lediglich die Anzahl insgesamt gelöster Anagramme angeben, ohne Berücksichtigung der Reihenfolge. Ein finanzieller Anreiz sollte hierbei einen möglichst substantiellen Anteil von Teilnehmern dazu animieren, die theoretisch zwar mögliche, praktisch jedoch unmöglich zu erreichende Anzahl von drei gelösten Anagrammen zu berichten. Bei der Konstruktion geeigneter Anagramme wurden drei Variablen berücksichtigt, die in Vorgängerstudien mit der Anagrammschwierigkeit assoziiert waren: Erstens sind Anagramme, deren Zielwörter in der betreffenden Sprache häufiger vorkommen, leichter zu lösen als Anagramme mit selteneren Zielwörtern (z.B. Dominowski, 1967; Lemay, 1972; Mayzner & Tresselt, 1958, 1959; Mendelsohn, 1976). Zweitens sind Anagramme umso leichter zu lösen, je weniger Buchstabenbewegungen nötig sind, um das Zielwort zu identifizieren (z.B. Dominowski, 1966; Mendelsohn & Obrien, 1974). Drittens wird die Lösung von Anagrammen leichter gefunden, wenn sich das Anagramm und das Zielwort ähnlicher sind; diese Ähnlichkeit wurde unter anderem operationalisiert über die absolute Anzahl an Buchstaben im Anagramm, die in der korrekten Reihenfolge verblieben sind (Gilhooly & Johnson, 1978) sowie über die Rangkorrelation der Buchstabenpositionen von Anagramm und Zielwort (z.B. Johnson, 1966; Terakoa, 1959). Als Basis für die Identifikation von Zielwörtern diente ein Datensatz des Projekts *Deutscher Wortschatz* (Quasthoff, Richter & Biemann, 2006), der eine Million deutscher Wörter enthält, die zufällig aus einer Million Sätze aus der deutschsprachigen Wikipedia (http://de.wikipedia.org) gezogen wurden. Dieser Datensatz enthält für jedes Wort eine Angabe zur Häufigkeitsklasse, die die relative Häufigkeit des betreffenden Wortes im Vergleich zum häufigsten Wort im

Korpus („der“) quantifiziert (zur Berechnung von Häufigkeitsklassen siehe z.B. Zipf, 1935); niedrigere Häufigkeitsklassen stehen hierbei für eine höhere Häufigkeit. Um zu verhindern, dass Teilnehmer Hypothesen über einen Zusammenhang von Wortlänge und Anagrammschwierigkeit generieren würden, um einfache Instruktionen vorgeben zu können und um ein einheitliches Erscheinungsbild der Aufgabe zu erreichen, wurde dieser Datensatz zunächst auf sechsbuchstabige Nomen in ihrer Grundform reduziert. Eigennamen und Wörter mit doppelten Buchstaben wurden ausgeschlossen. Schließlich wurden nur Wörter in die Auswahl eingeschlossen, die im Sinne einer Anagrammaufgabe einlösig waren – für die sich also durch weiteres Umstellen der Buchstaben kein zweites Anagramm bilden ließ. Aus dem resultierenden Datensatz wurden 32 Wörter ausgewählt, die in der deutschen Sprache verhältnismäßig häufig vorkommen (Häufigkeitsklassen 7 bis 11) und als potentielle Zielwörter für leichte Anagramme dienen sollten, sowie 13 Wörter mit einer niedrigen Worthäufigkeit (Häufigkeitsklassen 18 bis 22), die potentielle Zielwörter für schwere Anagramme darstellten. Leichte Anagramme wurden anschließend so gebildet, dass nur eine Buchstabenverschiebung zum Finden der Lösung nötig war, dass fünf der sechs Buchstaben in der korrekten Reihenfolge verblieben und dass die Rangkorrelation zwischen den Buchstabenpositionen von Anagramm und Zielwort so hoch wie möglich ausfiel ($\tau > .50$). Schwere Anagramme wurden so konstruiert, dass die maximal mögliche Anzahl von vier Buchstabenverschiebungen zur Lösung nötig war, dass keine Buchstaben in der korrekten Reihenfolge verblieben und dass die Rangkorrelation zwischen den Buchstabenpositionen von Anagramm und Zielwort möglichst niedrig ausfiel ($-.10 < \tau < .10$). In einer Online-Vorstudie wurden diese 45 Anagramme in randomisierter Reihenfolge $N = 136$ Versuchsteilnehmern vorgegeben mit einer Präsentationszeit von jeweils 20 Sekunden und zusätzlichen 20 Sekunden für die Eingabe der Lösung. Auf diesem Wege konnten zwei extrem leichte Anagramme identifiziert werden, die von allen Teilnehmern gelöst wurden und mit verhältnismäßig kurzen mittleren Präsentations- und Antwortzeiten einhergingen, sowie ein

extrem schweres Anagramm, das von keinem Teilnehmer gelöst wurde und lange mittlere Präsentations- und Antwortzeiten aufwies (siehe Tabelle 3). Diese drei Anagramme wurden in die finale Version der Anagrammrätselaufgabe aufgenommen.

Tabelle 3

Itemkennwerte aus der Vorstudie mit $N = 136$ Teilnehmern für die drei Anagramme, die in die finale Version der Anagrammaufgabe aufgenommen wurden

Zielwort	**Anagramm**	**% gelöst**	M_{pt} **(*SD*)**	M_{rt} **(*SD*)**
PRAXIS	APRXIS	100.00	1.99 (1.36)	3.51 (1.43)
UMWELT	UMELTW	100.00	1.90 (0.93)	3.09 (1.36)
TRIOLE	IERTLO	0.00	18.73 (3.60)	5.30 (5.90)

Anmerkung: M_{pt} = mittlere Präsentationszeit, M_{rt} = mittlere Antwortzeit.

In der Hauptstudie wurde $N = 664$ Teilnehmern (57% weiblich; $M_{Alter} = 36.72$ Jahre, $SD_{Alter} = 13.73$) im Rahmen eines Online-Fragebogens zunächst die konstruierte Anagrammaufgabe vorgegeben. Die Präsentationszeit der einzelnen Anagramme betrug nach wie vor 20 Sekunden; abweichend von der Vorstudie wurden nun jedoch nur noch die drei ausgewählten Anagramme in festgelegter Reihenfolge (APRXIS, UMELTW, IERTLO) präsentiert. Bei der Abfrage der Anzahl gelöster Anagramme wurden zusätzlich die gesuchten Zielwörter dargeboten, um sicherzustellen, dass Teilnehmer offensichtlich lügen mussten, wenn Sie angaben, drei Anagramme gelöst zu haben; gleichzeitig wurde die Angabe von drei gelösten Anagrammen durch einen finanziellen Anreiz (die Teilnahme an einem Gewinnspiel für Geschenkgutscheine im Wert von 20€, 30€ und 100€) provoziert. Tatsächlich entschieden sich 103 der 664 Teilnehmer (15.51%) dazu, die maximale Anzahl von drei gelösten Anagrammen zu berichten. Diese Teilnehmer wurden folglich als Betrüger klassifiziert, da drei gelöste Anagramme im Lichte der in der Vorstudie ermittelten Anagrammschwierigkeiten als praktisch unmöglich zu erreichen angesehen wurden. Anschließend sollten Teilnehmer zu der Aussage „Ich habe mehr gelöste Anagramme angegeben, als ich tatsächlich gelöst habe." Stellung beziehen, wobei die Befra-

gungstechnik als Zwischensubjektfaktor in den Stufen direkte Frage und CWM manipuliert wurde (Zuweisungsquote direkte Frage / CWM = 1 / 4, um die verringerte Effizienz des CWM zu kompensieren). Die Ergebnisse der Prävalenzschätzungen in Abhängigkeit von der Befragungstechnik, sowie die Ergebnisse der inferenzstatistischen Vergleiche der Schätzer miteinander und mit der bekannten Prävalenz von 15.51% können Tabelle 4 entnommen werden.

Tabelle 4

Prävalenzschätzungen für das sensible Merkmal Betrug in der Anagrammaufgabe in den unterschiedlichen experimentellen Bedingungen

Bedingung	$\hat{\pi}$ (*SE*)	**Test gegen wahren Wert (15.51%)**		**Test gegen CWM-Schätzer**	
		ΔG^2 (*df*=1)	*p*	ΔG^2 (*df*=1)	*p*
Direkte Frage	5.07% (1.87)	14.88	< .001	5.29	< .05
CWM	13.03% (2.75)	0.79	.37	–	

Die bekannte Prävalenz von Betrug in der Anagrammaufgabe wurde durch eine direkte Frage substantiell unterschätzt. Eine Schätzung auf Basis des CWM fiel signifikant höher aus als die Schätzung durch eine direkte Frage und wich darüber hinaus nur geringfügig und nicht signifikant von der bekannten Prävalenz des sensiblen Merkmals ab. In Anbetracht der Ergebnisse dieser starken Validierung wurde das CWM somit als geeignete Methode evaluiert, den Einfluss sozial erwünschten Antwortverhaltens in Umfragen zu sensiblen Merkmalen zu kontrollieren und valide Prävalenzschätzungen hervorzubringen. Zusätzlich wurde in Experiment 3 ein Paradigma zur experimentellen Induktion eines sensiblen Merkmals entwickelt, das einfach anzuwenden ist und auch in Online-Studien eingesetzt werden kann. Dieses Paradigma bietet vielversprechende Möglichkeiten zur Durchführung weiterer starker Validierungsstudien für Methoden, die auf die Kontrolle sozialer Erwünschtheit abzielen.

6 Diskussion und Ausblick

In Studien, die direkte Selbstauskünfte als Basis für die Schätzung der Prävalenz sensibler Merkmale nutzen, wird die Validität der Ergebnisse durch den Einfluss sozial erwünschten Antwortverhaltens bedroht. Indirekte Befragungstechniken wie die RRT (Warner, 1965) wurden entwickelt, um den Einfluss sozialer Erwünschtheit in Umfragen zu kontrollieren. Techniken dieser Art garantieren durch eine Zufallsverschlüsselung die Vertraulichkeit individueller Antworten, sollen damit ehrliches Antwortverhalten fördern und schließlich in valideren Prävalenzschätzungen münden. Neuere RRT-Modelle wie der SLD (Moshagen et al., 2012) berücksichtigen zusätzlich, dass manche Teilnehmer auch in RRT-Studien versuchen, ihren wahren Merkmalsstatus zu verschleiern und ermöglichen neben einer Prävalenzschätzung für das sensible Merkmal auch eine Schätzung des Anteils unehrlich antwortender Merkmalsträger. Eine vielversprechende aktuelle Weiterentwicklung der RRT stellt das CWM (Yu et al., 2008) dar. Dieses Modell verfügt über besonders einfache Instruktionen und ein symmetrisches Antwortformat, das es Teilnehmern unmöglich macht, durch die Wahl einer Antwortalternative eine Merkmalsträgerschaft sicher auszuschließen. Falls diese Eigenschaften tatsächlich zu vollständig ehrlichem Antwortverhalten führen, könnte das CWM konkurrierenden Modellen wie dem SLD trotz fehlender Verweigererdetektion überlegen sein und zu validen Prävalenzschätzungen führen. In der vorliegenden Dissertation wurde die Validität des CWM in drei Experimenten einer empirischen Überprüfung unterzogen.

Bezugnehmend auf die in Kapitel 4 formulierten Forschungsfragen konnte anhand der ermittelten Ergebnisse erstens gezeigt werden, dass das CWM für insgesamt drei sensible Merkmale (Experimente 1 und 3) höhere Prävalenzschätzungen hervorbrachte als eine konventionelle direkte Frage. Dem *more-is-better*-Kriterium (Umesh & Peterson, 1991) folgend werden diese höheren Schätzer als valider angesehen, weil sie potentiell weniger durch den Einfluss sozialer Erwünschtheit verzerrt

sind. Die Gegenüberstellung mit dem SLD in Experiment 1 ergab außerdem, dass die CWM- und SLD-Prävalenzschätzungen für eines der sensiblen Merkmale (Xenophobie) vergleichbar ausfielen; für ein zweites sensibles Merkmal (Islamophobie) hingegen fiel die Schätzung des SLD höher aus als die Schätzung durch das CWM. Die Interpretation dieses uneindeutigen Ergebnismusters wird zusätzlich erschwert durch den Befund, dass der SLD die bekannte Prävalenz eines nicht-sensiblen Kontrollmerkmals signifikant überschätzte. Da das CWM die Prävalenz dieses Kontrollmerkmals hingegen adäquat abbildete und somit für alle drei Merkmale in Experiment 1 plausible Ergebnisse lieferte, kann den Schätzungen auf Basis dieses Modells vermutlich eher vertraut werden als Schätzungen auf Basis des SLD. Aus der schwachen Validierung und dem Methodenvergleich mit einem RRT-Modell mit Verweigererdetektion ging das CWM deshalb als Favorit hervor. Zweitens zeigte das CWM verglichen mit drei konkurrierenden indirekten Befragungstechniken, dem SLD, dem CDM (Clark & Desharnais, 1998) und der UCT (Miller, 1984), die vergleichsweise höchste Verständlichkeit (Experiment 2). Auch in der subjektiv empfundenen Vertraulichkeit erreichte das Modell einen der höchsten Werte, der außerdem gegenüber einer direkten Frage substantiell erhöht war. Da das Instruktionsverständnis und das Vertrauen der Teilnehmer, das der verwendeten Befragungstechnik entgegengebracht wird, zwei zentrale psychologische Aspekte bei der validen Erfassung der Prävalenz sensibler Merkmale darstellen, wurde das CWM auch in Experiment 2 als die vielversprechendste indirekte Befragungstechnik bewertet. Drittens konnte im Rahmen einer starken Validierungsstudie gezeigt werden, dass das CWM für die bekannte Prävalenz eines experimentell erzeugten sensiblen Merkmals eine genaue Schätzung lieferte, während diese Prävalenz durch eine direkte Frage drastisch unterschätzt wurde. Dieser Befund lieferte abschließend starke Evidenz für die Validität von Prävalenzschätzungen, die mit Hilfe des CWM ermittelt werden.

Da durch die vorliegende Arbeit weitere schwache, nun jedoch vor allem auch starke Evidenz für die Validität des CWM vorliegt, bietet sich die praktische An-

wendung dieses Modells bei der Erfassung von potentiell sozial unerwünschten Merkmalen in der Umfrageforschung an. Beispielsweise hat schon die Befragung zum Minarettverbot (als Index für Islamophobie) in Experiment 1 das Potential des CWM als mögliche Alternative zu direkten Befragungen für Wahlprognosen aufgezeigt: Während eine direkte Frage eine Ablehnung des Minarettverbots prognostizierte, sagte das CWM vorher, dass die Mehrheit der Befragten für ein solches Verbot stimmen würde. Vielleicht hätte sich auch das Ergebnis der Volksabstimmung in der Schweiz im Jahr 2009 bei der Kontrolle sozial erwünschten Antwortverhaltens, z.B. durch das CWM, genauer vorhersagen lassen. Weiter könnte das CWM in der empirischen Dunkelfeldforschung zu kriminell relevantem Verhalten zum Einsatz kommen: Besonders für die Prävalenz schwerer Straftaten, die gesellschaftlich eine hohe Relevanz besitzen (z.B. Vergewaltigung und sexuelle Nötigung, sexueller Missbrauch von Kindern, etc.), für die jedoch nur wenig verlässliche Daten existieren (z.B. die polizeiliche Kriminalstatistik, die nur das Hellfeld, also die tatsächlich zur Anzeige gebrachten Fälle, widerspiegelt), könnten repräsentative Opfer- und Täterbefragungen mit dem CWM eine attraktive alternative Datenquelle darstellen. Darüber hinaus ließe sich das CWM einsetzen, um die Stärke des Einflusses sozialer Erwünschtheit für spezifische Merkmale zu quantifizieren: Da das Modell offensichtlich von sozialer Erwünschtheit unverzerrte Prävalenzschätzungen liefert, sollten Differenzen zwischen den Schätzungen auf Basis einer direkten Frage und Schätzungen auf Basis des CWM die Sensibilität der betreffenden Merkmale abbilden. So wäre beispielsweise denkbar, dass der Selbstbericht xenophober Einstellungen von politisch linksorientierten Befragten als sensibler empfunden wird als von politisch rechtsorientierten Befragten. Dieser Sensibilitätsunterschied sollte sich in einer größeren Differenz zwischen direkt und indirekt gewonnenen Schätzern innerhalb der Stichprobe der linksorientierten Befragten ausdrücken.

Das CWM könnte darüber hinaus zur weiteren Erforschung konkurrierender Befragungstechniken verwendet werden. Beispielsweise ist für das CDM nicht be-

kannt und bisher nicht erforscht, wie sich die Gruppe der Verweigerer (γ) bezüglich einer Merkmalsträgerschaft zusammensetzt. So könnte eine mit Hilfe des CWM gewonnene, unverzerrte Prävalenzschätzung für ein sensibles Merkmal zur Beantwortung der Frage beitragen, ob im Sinne des CDM in Wirklichkeit alle Verweigerer Merkmalsträger sind (in diesem Fall müsste die CWM-Schätzung etwa bei $\pi + \gamma$ in der CDM-Bedingung liegen), kein Verweigerer Merkmalsträger ist (in diesem Fall läge die CWM-Schätzung etwa bei π in der CDM-Bedingung) oder sowohl Merkmalsträger als auch Nicht-Merkmalsträger zu den Verweigerern gehören (in diesem Fall läge die CWM-Schätzung im Intervall zwischen π und $\pi + \gamma$ in der CDM-Bedingung). Für Methodenvergleiche dieser Art, oder auch Vergleiche mit anderen indirekten Befragungstechniken, böte sich die Verwendung des in Experiment 3 entwickelten Paradigmas zur experimentellen Induktion eines sensiblen Merkmals an, um gleichzeitig ein objektives Außenkriterium für die Validität der verwendeten Verfahren zur Verfügung zu haben.

Schließlich könnte das CWM als ein Modell, das eine verhältnismäßig hohe Verständlichkeit mit sich bringt (Experiment 2), in Kontrast zu einem schwerer verständlichen Modell (z.B. dem SLD) eingesetzt werden, um die Determinanten des Instruktionsverständnisses für indirekte Befragungstechniken weiter zu ergründen. Qualitative Interviewstudien (wie beispielsweise in Boeije & Lensvelt-Mulders, 2002) und Experimente, die verschiedene Aspekte der Instruktionen systematisch variieren (z.B. die Länge der Instruktionen, den Einsatz von Verständnistests, etc.) könnten zur Entwicklung optimierter, standardisierter Instruktionen beitragen, die grundsätzlich zu hohen Verständnisraten führen. Das Ziel sollte hierbei sein, die Verständlichkeit für indirekte Befragungstechniken mindestens auf das Niveau einer direkten Frage zu heben, um Versuchsteilnehmern eine Einsicht in die immensen Vorteile solcher Befragungen zu geben und die subjektiv empfundene Vertraulichkeit möglichst zu maximieren.

Werden die Ergebnisse der durchgeführten Studien in Zusammenhang gebracht, so kann das CWM als geeignetes Verfahren für die Kontrolle des Einflusses sozialer Erwünschtheit in Umfragen evaluiert werden. Das CWM ist leichter zu verstehen als konkurrierende Verfahren und geht gegenüber einer direkten Befragung mit einem substantiell erhöhten Vertrauen auf Seiten der Befragten einher. Mit dem CWM liegt somit eine indirekte Befragungstechnik vor, deren Einsatz in Umfragen zu sensiblen Merkmalen die Ermittlung unverzerrter und damit valider Prävalenzschätzungen verspricht.

Literaturverzeichnis

Abernathy, J. R., Greenberg, B. G. & Horvitz, D. G. (1970). Estimates of Induced Abortion in Urban North-Carolina. *Demography, 7*(1), 19-29.

Abul-Ela, A. L. A., Greenberg, B. G. & Horvitz, D. G. (1967). A Multi-Proportions Randomized Response Model. *Journal of the American Statistical Association, 62*, 990-1008.

Ahart, A. M. & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods, 7*, 101-114.

Akers, R. L., Massey, J., Clarke, W. & Lauer, R. M. (1983). Are Self-Reports of Adolescent Deviance Valid? Biochemical Measures, Randomized-Response, and the Bogus Pipeline in Smoking-Behavior. *Social Forces, 62*, 234-251.

Antonak, R. F. & Livneh, H. (1995). Randomized-Response Technique - a Review and Proposed Extension to Disability Attitude Research. *Genetic, Social, and General Psychology Monographs, 121*, 97-145.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10*, 331-344.

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

Boeije, H. & 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.

Boruch, R. F. (1972). Relations among statistical methods for assuring confidentiality of social research data. *Social Science Research, 1*, 403-414.

Bunzl, M. (2005). Between anti-Semitism and Islamophobia: Some thoughts on the new Europe. *American Ethnologist, 32*, 499-508.

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

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

Chaudhuri, A. & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

Chi, I., Chow, L. P. & Rider, R. V. (1972). Randomized Response Technique as Used in Taiwan Outcome of Pregnancy Study. *Studies in Family Planning, 3*, 265-269.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). Hillsdale: Erlbaum.

Coutts, E. & Jann, B. (2011). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research, 40*, 169-193.

Coutts, E., Jann, B., Krumpal, I. & Näher, A. F. (2011). Plagiarism in Student Papers: Prevalence Estimates Using Special Techniques for Sensitive Questions. *Jahrbücher für Nationalökonomie Und Statistik, 231*, 749-760.

Dawes, R. M. & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen. In F. Petermann (Hrsg.), *Einstellungsmessung, Einstellungsforschung*. 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.

Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P. & Ulrich, R. (2013). Randomized Response Estimates for the 12-Month Prevalence of Cognitive-Enhancing Drug Use in University Students. *Pharmacotherapy, 33*, 44-50.

Dominowski, R. L. (1966). Anagram Solving as a Function of Letter Moves. *Journal of Verbal Learning and Verbal Behavior, 5*, 107-111.

Dominowski, R. L. (1967). Anagram Solving as a Function of Bigram Rank and Word Frequency. *Journal of Experimental Psychology, 75*, 299-306.

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.

Erdfelder, E. & Musch, J. (2006). Experimental methods of psychological assessment. In M. Eid & E. Diener (Hrsg.), *Handbook of Multimethod Measurement in Psychology* (S. 205-220). Washington, D.C.: American Psychological Association.

Eriksson, S. A. (1973). A New Model for Randomized Response. *International Statistical Review, 41*, 101-113.

Eslami, M., Yazdanpanah, M., Taheripanah, R., Andalib, P., Rahimi, A. & Nakhaee, N. (2013). Importance of Pre-pregnancy Counseling in Iran: Results from the High Risk Pregnancy Survey 2012. *International Journal of Health Policy and Management, 1*, 213-218.

EUMC - European Monitoring Center on Racism and Xenophobia. (2006). *Muslims in the European Union: Discrimination and Islamophobia*. Wien: FRA.

Fidler, D. S. & Kleinknecht, R. E. (1977). Randomized Response Versus Direct Questioning - 2 Data-Collection Methods for Sensitive Information. *Psychological Bulletin, 84*, 1045-1049.

fög. (2010). *Berichterstattung zur Volksinitiative 'Gegen den Bau von Minaretten'*. Zugriff am 8. Juni 2010 unter http://www.foeg.unizh.ch.

Fox, J. A. & Tracy, P. E. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills: Sage.

Fox, J. P. & Meijer, R. R. (2008). Using Item Response Theory to Obtain Individual Information From Randomized Response Data: An Application Using Cheating Data. *Applied Psychological Measurement, 32*, 595-610.

Franklin, L. (1998). Randomized Response Techniques. In P. Armitage & T. Colton (Hrsg.), *Encyclopedia of Biostatistics* (Bd. 5, S. 3696-3703). New York: Wiley.

gfs.bern. (2009a). *'Minarett-Initiative': Das Nein überwiegt – SVP-Wählerschaft dafür*. Zugriff am 6. Juni 2010 unter www.gfsbern.ch.

gfs.bern. (2009b). *'Minarett-Initiative': Ja nimmt zu - Nein unverändert stärker*. Zugriff am 8. Juni 2010 unter www.gfsbern.ch.

Gilhooly, K. J. & Johnson, C. E. (1978). Effects of solution word attributes on anagram difficulty - a regression analysis. *Quarterly Journal of Experimental Psychology, 30*, 57-70.

Goodstadt, M. S. & Gruson, V. (1975). Randomized Response Technique - Test on Drug-Use. *Journal of the American Statistical Association, 70*, 814-818.

Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R. & Horvitz, D. G. (1969). Unrelated Question Randomized Response Model - Theoretical Framework. *Journal of the American Statistical Association, 64*, 520-539.

Greenberg, B. G., Horvitz, D. G. & Abernathy, J. R. (1974). A comparison of randomized response designs. In F. Proschan & R. J. Serfling (Hrsg.), *Reliability and biometry, statistical analysis of life length* (S. 787-815). Philadelphia: SIAM.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R. & Horvitz, D. G. (1971). Application of Randomized Response Technique in Obtaining Quantitative Data. *Journal of the American Statistical Association, 66*, 243-250.

Hejri, S. M., Zendehdel, K., Asghari, F., Fotouhi, A. & Rashidian, A. (2013). Academic disintegrity among medical students: a randomised response technique study. *Medical Education, 47*, 144-153.

Hilbig, B. E. & Hessler, C. M. (2013). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology, 49*, 263-266.

Himmelfarb, S. & Edgell, S. E. (1980). Additive Constants Model - a Randomized-Response Technique for Eliminating Evasiveness to Quantitative Response Questions. *Psychological Bulletin, 87*, 525-530.

Hjerm, M. (1998). National identities, national pride and xenophobia: A comparison of four Western countries. *Acta Sociologica, 41*, 335-347.

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.

Horvitz, D. G., Greenberg, B. G. & Abernathy, J. R. (1976). Randomized Response - Data-Gathering Device for Sensitive Questions. *International Statistical Review, 44*, 181-196.

Horvitz, D. G., Shah, S. & Simmons, W. R. (1967). The Unrelated Question Randomized Response Model. *Proceedings of the Social Statistics Section, American Statistical Association*.

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

Iacono, W. G. (2000). The detection of deception. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Hrsg.), *Handbook of Psychophysiology* (2. Auflage, S. 772-793). New York: Cambridge University Press.

IIT Research Institute and the Chicago Crime Commission. (1971). *A study of organized crime in Chicago*. Chicago: IITRI Project No. H-6031, Bericht an die Illinois Enforcement Commission.

Imhoff, R. & Recker, J. (2012). Differentiating Islamophobia: Introducing a New Scale to Measure Islamoprejudice and Secular Islam Critique. *Political Psychology, 33*, 811-824.

James, R. A., Nepusz, T., Naughton, D. P. & Petroczi, A. (2013). A potential inflating effect in estimation models: Cautionary evidence from comparing performance enhancing drug and herbal hormonal supplement use estimates. *Psychology of Sport and Exercise, 14*, 84-96.

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

Jimenez, P. (1999). Weder Opfer noch Täter - die alltäglichen Einstellungen 'unbeteiligter' Personen gegenüber Ausländern. In R. Dollase, T. Kliche & H. Moser (Hrsg.), *Politische Psychologie der Fremdenfeindlichkeit. Opfer - Täter - Mittäter* (S. 293–306). Weinheim: Juventa.

Johnson, D. M. (1966). Solution of Anagrams. *Psychological Bulletin, 66*, 371-384.

Jones, E. E. & Sigall, H. (1971). The Bogus Pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76*, 349-364.

Klink, A. & Wagner, U. (1999). Discrimination against ethnic minorities in Germany: Going back to the field. *Journal of Applied Social Psychology, 29*, 402-423.

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.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity, 47*, 2025-2047.

Kuk, A. Y. C. (1990). Asking Sensitive Questions Indirectly. *Biometrika, 77*, 436-438.

Kulka, R. A., Weeks, M. F. & Folsom, R. E. (1981). *A comparison of the randomized response approach and direct questioning approach to asking sensitive survey questions* (working paper). North Carolina: Research Triangle Institute.

LaBrie, J. W. & Earleywine, M. (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. *Journal of Sex Research, 37*, 321-326.

Landsheer, J. A., van der Heijden, P. & 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.

Lemay, E. H. (1972). Anagram Solutions as a Function of Task Variables and Solution Word Models. *Journal of Experimental Psychology, 92*, 65-68.

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.

Liu, P. T. & Chow, L. P. (1976). A New Discrete Quantitative Randomized Response Model. *Journal of the American Statistical Association, 71*, 72-73.

Liu, P. T., Chow, L. P. & Mosley, W. H. (1975). Use of Randomized Response Technique with a New Randomizing Device. *Journal of the American Statistical Association, 70*, 329-332.

Locander, W., Sudman, S. & Bradburn, N. (1976). An Investigation of Interview Method, Threat and Response Distortion. *Journal of the American Statistical Association, 71*, 269-275.

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

Mangat, N. S. & Singh, R. (1990). An Alternative Randomized-Response Procedure. *Biometrika, 77*, 439-442.

Mayzner, M. S. & Tresselt, M. E. (1958). Anagram Solution Times: A Function of Letter Order and Word-Frequency. *Journal of Experimental Psychology, 56*, 376-379.

Mayzner, M. S. & Tresselt, M. E. (1959). Anagram Solution Times: A Function of Transition-Probabilities. *Journal of Psychology, 47*, 117-125.

Mendelsohn, G. A. (1976). An Hypothesis Approach to the Solution of Anagrams. *Memory & Cognition, 4*, 637-642.

Mendelsohn, G. A. & Obrien, A. T. (1974). The Solution of Anagrams - A Reexamination of the Effects of Transition Letter Probabilities, Letter Moves, and Word Frequency on Anagram Difficulty. *Memory & Cognition, 2*, 566-574.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior* (unveröffentlichte Doktorarbeit). Washington, D.C.: George Washington University, Department of Sociology.

Moors, J. J. A. (1971). Optimization of Unrelated Question Randomized Response Model. *Journal of the American Statistical Association, 66*, 627-629.

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

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.

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.

Moshagen, M., Musch, J. & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44*, 222-231.

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

Musch, J., Brockhaus, R. & Bröder, A. (2002). Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit. *Diagnostica, 48*, 121-129.

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 (Hrsg.), *Dimensions of Internet science* (S. 179-192). Lengerich: 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*(3-4), 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.

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.

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.

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.

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

Paulhus, D. L. (1994). Balanced Inventory of Desirable Responding: Reference manual for BIDR Version 6. *Unpublished manuscript, University of British Columbia, Vancouver, Canada.*

Paulhus, D. L. (1998). *The Balanced Inventory of Desirable Responding*. Toronto, Canada: Multi-Health Systems.

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

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.

Pollock, K. H. & Bek, Y. (1976). A Comparison of 3 Randomized Response Models for Quantitative Data. *Journal of the American Statistical Association, 71*, 884-886.

Quasthoff, U., Richter, M. & Biemann, C. (2006). *Corpus Portal for Search in Monolingual Corpora.* Paper presented at the fifth international conference on Language Resources and Evaluation, LREC, Genoa.

reformiert. (2009). *Mehrheit ist gegen ein Minarettverbot*. Zugriff am 8. Juni 2010 unter www.ref.ch.

Reinders, M. (1996). Häufigkeit von Namensanfängen. *Statistische Rundschau Nordrhein-Westfalen, 11*, 651-660.

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.

Scheers, N. J. (1992). A Review of Randomized-Response Techniques. *Measurement and Evaluation in Counseling and Development, 25*, 27-41.

Scheers, N. J. & Dayton, C. M. (1987). Improved Estimation of Academic Cheating Behavior Using the Randomized-Response Technique. *Research in Higher Education, 26*(1), 61-69.

Sheridan, L. P. (2006). Islamophobia pre- and post-September 11th, 2001. *Journal of Interpersonal Violence, 21*, 317-336.

Silbermann, A. & Hüsers, F. (1995). *Der 'normale' Haß auf die Fremden. Eine sozialwissenschaftliche Studie zu Ausmaß und Hintergründen von Fremdenfeindlichkeit in Deutschland*. München: Quintessenz.

Simon, P., Striegel, H., Aust, F., Dietz, K. & Ulrich, R. (2006). Doping in fitness sports: estimated number of unreported cases and individual probability of doping. *Addiction, 101*, 1640-1644.

Soeken, K. L. & Damrosch, S. P. (1986). Randomized-Response Technique - Applications to Research on Rape. *Psychology of Women Quarterly, 10*, 119-125.

Stöber, J. (1999). Die Soziale-Erwünschtheits-Skala-17 (SES-17). *Diagnostica, 45*, 173-177.

Striegel, H., Ulrich, R. & Simon, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence, 106*, 230-232.

Sudman, S. & Bradburn, N. (1974). *Response effects in surveys*. Chicago: Aldine.

Terakoa, T. (1959). Effects of letter-orders and material words on the anagram solution. *Japanese Journal of Psychology, 30*, 253-263.

Tian, G. L. & Tang, M. L. (2013). *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys*. Boca Raton: CRC Press, Taylor & Francis Group.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883.

Tracy, D. S. & Mangat, N. S. (1996). Some development in randomized response sampling during the last decade - a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science, 4*, 147-158.

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.

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.

Vakilian, K., Mousavi, S. A. & Keramat, A. (2014). Estimation of sexual behavior in the 18-to-24-years-old Iranian youth based on a crosswise model study. *BMC Research Notes, 7*(28), 1-4.

van der Heijden, P. G. M., van Gils, G., Bouts, J. & Hox, J. J. (1998). A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of social security fraud. *Kwantitatieve Methoden, 19*, 15-34.

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.

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

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes, 115*, 157-168.

Wimbush, J. C. & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology, 82*, 756-763.

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.

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.

Zdep, S. M. & Rhodes, I. N. (1977). Making the Randomized Response Technique Work. *Public Opinion Quarterly, 40*, 531-537.

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

Zipf, G. K. (1935). *The psycho-biology of language*. Oxford: Houghton-Mifflin.

Anhang: Einzelarbeiten

Experiment 1:

Hoffmann, A., & Musch, J. (2014). Assessing the validity of two indirect questioning techniques: a Stochastic Lie Detector versus the Crosswise Model. *Manuscript submitted for publication*.

Experiment 2:

Hoffmann, A., Schmidt, A. F., Waubert de Puiseau, B., & Musch, J. (2014). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Manuscript submitted for publication*.

Experiment 3:

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2014). A strong validation of the Crosswise Model using experimentally induced cheating behavior. *Manuscript submitted for publication*.

Assessing the Validity of Two Indirect Questioning Techniques: a Stochastic Lie Detector Versus the Crosswise Model

Adrian Hoffmann and Jochen Musch

University of Duesseldorf

Author Note

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

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

Estimates of the prevalence of sensitive attributes obtained through direct questions are prone to being distorted by untruthful responding. Indirect questioning procedures such as the Randomized Response Technique (RRT; Warner, 1965) aim to control for the influence of social desirability bias. However, even on RRT surveys, some participants may disobey the instructions in an attempt to conceal their true status. In the present study, we experimentally compared the validity of two competing indirect questioning techniques that presumably offer a solution to the problem of nonadherent respondents: the Stochastic Lie Detector (Moshagen, Musch, & Erdfelder, 2012) and the Crosswise Model (Yu, Tian, & Tang, 2008). For two sensitive attributes, both techniques met the "more is better" criterion. Their application resulted in higher, and thus presumably more valid, prevalence estimates than a direct question. Only the Crosswise Model, however, adequately estimated the known prevalence of a nonsensitive control attribute.

Keywords: Randomized Response Technique, Stochastic Lie Detector, Crosswise Model, social desirability bias

Assessing the Validity of Two Indirect Questioning Techniques: a Stochastic Lie Detector Versus the Crosswise Model

When assessing the prevalence of sensitive personal attributes, the validity of prevalence estimates obtained via direct questioning procedures (*DQ*) is threatened by response bias. Respondents frequently choose to align their answers to sensitive questions with social norms in order to make or uphold a socially desirable impression (Krumpal, 2013; Marquis, Marquis, & Polich, 1986; Paulhus, 1991; Paulhus & Reid, 1991; Phillips & Clancy, 1972; Rasinski, Willis, Baldwin, Yeh, & Lee, 1999; Stocké, 2007; Sudman & Bradburn, 1974; Tourangeau & Yan, 2007). Consequently, prevalence estimates of sensitive attributes may be distorted by the underreporting of socially undesirable and the overreporting of socially desirable attitudes and behaviors.

Warner (1965) proposed the Randomized Response Technique (RRT) to increase respondents' willingness to cooperate on sensitive surveys. This technique improves the confidentiality of individual answers by employing a randomization procedure that removes the direct association between a respondent's answer and his or her standing on the sensitive attribute. However, even on RRT surveys, respondents may fail to adhere to the instructions in order to conceal their true status. After providing a brief introduction to the Randomized Response Technique, we will therefore describe and evaluate two recently proposed advanced models that were designed to address the problem of nonadherence to the instructions: The *Stochastic Lie Detector* (SLD; Moshagen et al., 2012) and the competing *Crosswise Model* (CWM; Yu et al., 2008). The Stochastic Lie Detector (SLD; Moshagen et al., 2012) implements an additional parameter to estimate the proportion of sensitive-attribute carriers who cheat on the survey. Arguably, this should result in a more accurate prevalence estimate than traditional RRT procedures. The competing Crosswise Model (CWM; Yu et al., 2008) does not model cheating

but is instead characterized by rather simple instructions that make it particularly easy to understand how the confidentiality of answers is protected. Like the original Warner (1965) model, the CWM is symmetrical in the sense that it does not provide a "safe" answer option that offers the opportunity to explicitly deny being a carrier of the sensitive attribute. There are, however, no studies that have compared the validity of the two approaches. Therefore, we conducted a large-scale experimental survey that aimed to evaluate and compare the two models with regard to their convergent validity and their ability to estimate the known prevalence of a control attribute. We also tested the two models against a direct questioning control condition.

The Randomized Response Technique

The general idea behind the RRT is to ensure the confidentiality of individual answers to sensitive questions by adding random noise to the responses. In the original *Related Questions Model* (RQM; Warner, 1965), respondents are simultaneously presented with two Questions A ("Are you a carrier of the sensitive attribute?") and B ("Are you *not* a carrier of the sensitive attribute?"). Depending on the outcome of a randomization procedure, the respondents are asked to answer either of these questions. If, for example, a die is used, subjects are instructed to respond to Question A if the die shows one of the numbers 1 to 4 (randomization probability p = 4/6 = .67) and to respond to Question B if the die shows either of the numbers 5 or 6 (1 - p = 2/6 = .33). Because the outcome of the randomization procedure remains unknown to the questioner, the true status of an individual respondent with respect to the sensitive attribute cannot be derived from his or her answer: A "Yes" response could possibly have been given by a carrier of the sensitive attribute who was instructed to respond to Statement A or from a noncarrier who was instructed to respond to Statement B. In view of the confidentiality thus afforded, respondents are expected to answer more truthfully than when questioned directly. In spite of the

confidentiality guaranteed to the individual respondent, an estimate of the prevalence π of the sensitive attribute can be obtained at the sample level. Warner (1965) showed the maximum likelihood estimate of π in the RQM to be

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

where *p* is the known probability that the randomization device would select Statement A, *n'* represents the total number of "Yes" responses, and *n* reflects the sample size. Compared with a conventional direct questioning procedure, the RRT has lower statistical efficiency because the randomization procedure adds unsystematic variance to the answers. The reduced efficiency, however, is supposed to be overcompensated for by an increase in the validity of the prevalence estimates resulting from the presumably higher proportion of honest respondents.

Within the last almost 50 years, a large number of RRT models have been developed with various objectives such as improving efficiency (e.g., Boruch, 1971; Dawes & Moore, 1980; Eriksson, 1973; Mangat, 1994; Mangat & Singh, 1990; Moors, 1971), including questions with multicategorical or quantitative answers (e.g., Abul-Ela, Greenberg, & Horvitz, 1967; Himmelfarb & Edgell, 1980; Liu & Chow, 1976; Pollock & Bek, 1976), increasing respondents' cooperation (e.g., Greenberg, Abul-Ela, Simmons, & Horvitz, 1969; Horvitz, Shah, & Simmons, 1967; Kuk, 1990; Ostapczuk, Moshagen, Zhao, & Musch, 2009), and accounting for cheating or noncompliance with the instructions (e.g., Clark & Desharnais, 1998; Moshagen et al., 2012). The RRT has been applied in surveys covering a variety of sensitive topics such as drug use (Dietz et al., 2013; Goodstadt & Gruson, 1975), doping (James, Nepusz, Naughton, & Petroczi, 2013; Simon, Striegel, Aust, Dietz, & Ulrich, 2006; Striegel, Ulrich, & Simon, 2010), crime (IIT Research Institute and the Chicago Crime Commission, 1971; Wolter & Preisendörfer, 2013),

unwed motherhood (Abul-Ela et al., 1967), promiscuity (Liu, Chow, & Mosley, 1975), abortion (Abernathy, Greenberg, & Horvitz, 1970; Greenberg, Kuebler, Abernathy, & Horvitz, 1971), rape (Fidler & Kleinknecht, 1977; Soeken & Damrosch, 1986), homosexuality (Clark & Desharnais, 1998), tax evasion (Edgell, Himmelfarb, & Duchan, 1982), fraud (van der Heijden, van Gils, Bouts, & Hox, 2000), academic cheating (J. P. Fox & Meijer, 2008; Hejri, Zendehdel, Asghari, Fotouhi, & Rashidian, 2013; Ostapczuk, Moshagen, et al., 2009), xenophobia (Ostapczuk, Musch, & Moshagen, 2009), negative attitudes toward people with disabilities (Ostapczuk & Musch, 2011), dental hygiene (Moshagen, Musch, Ostapczuk, & Zhao, 2010), and domestic violence (Moshagen et al., 2012). Overviews of RRT models and their applications have been given by Greenberg, Horvitz, and Abernathy (1974), Horvitz, Greenberg, and Abernathy (1976), J. A. Fox and Tracy (1986), Chaudhuri and Mukerjee (1988), Umesh and Peterson (1991), Scheers (1992), Antonak and Livneh (1995), Tracy and Mangat (1996), Franklin (1998), Chaudhuri (2011), and Chaudhuri and Christofides (2013).

In two meta-analyses, Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) reported an overall positive effect of the RRT on the validity of self-reports. The 32 comparative studies they found generally arrived at higher prevalence estimates of sensitive attributes in the RRT condition than in the direct questioning (DQ) control condition. Applying a “more is better” criterion, these higher estimates were usually considered to be more valid. However, this validation approach can be criticized as providing only relatively weak evidence because it is possible that both direct and indirect questioning techniques will provide inaccurate prevalence estimates (e.g., Umesh & Peterson, 1991). It is therefore important that in an additional meta-analysis of six methodologically stronger validation studies in which the respondents’ true status with respect to the sensitive attribute was known to the questioner, Lensvelt-Mulders et al.

(2005) also found RRT estimates to be more valid than estimates obtained via direct questioning because the RRT estimates deviated less from the known values in the population.

Despite the apparent advantages of RRT questioning, however, not all studies have supported its alleged superiority over conventional questioning methods. In some studies, estimates obtained via the RRT did not differ from those obtained via direct questioning (e.g., Akers, Massey, Clarke, & Lauer, 1983; Locander, Sudman, & Bradburn, 1976; Wolter & Preisendörfer, 2013). In other studies, they were even lower (e.g., Holbrook & Krosnick, 2010; Kulka, Weeks, & Folsom, 1981). Furthermore, Edgell et al. (1982) showed that a substantial proportion of participants failed to follow the RRT instructions, especially on surveys addressing highly sensitive issues. In view of these diverging patterns of results, Holbrook and Krosnick (2010) called the validity of RRT surveys into question.

Respondent jeopardy and *risk of suspicion* provide potential explanations for the divergent findings because either of these response hazards may lead to a violation of the assumptions underlying RRT models (Antonak & Livneh, 1995). The influence of these response hazards can primarily be observed in—and best be described with—forced-choice RRT designs (Boruch, 1971; Dawes & Moore, 1980). In this design variant, all participants are confronted with a single sensitive question, and a randomly chosen subsample is instructed to respond "Yes" regardless of their true status. Hence, a "Yes" response can stem from either a carrier or a noncarrier of the sensitive attribute who is either responding truthfully (carrier) or has simply been told to answer in the affirmative (carriers *and* noncarriers). It is important to note, however, that participants can still explicitly decline being carriers of the sensitive attribute by ignoring the instructions and simply responding "No." In this situation, *respondent jeopardy* refers to the problem that guilty respondents make themselves more vulnerable by answering a sensitive

question in the affirmative because they can be identified as carriers with a higher probability after a "Yes" than after a "No" response. If carriers perceive the risk of being identified as carriers as too high, they may choose to disobey the instructions by dishonestly responding "No." Innocent respondents, on the other hand, suffer from a *risk of suspicion* because noncarriers have a higher risk of being falsely identified as carriers if they are forced to respond "Yes" by the randomization device. For this reason, they may also be inclined to disregard the instructions and to respond "No" in spite of being told otherwise (Antonak & Livneh, 1995). Lying carriers and suspicion-avoiding noncarriers were explicitly accounted for by the introduction of the cheating detection model.

Detection of Cheating on RRT Surveys

Clark and Desharnais (1998) argued that even on RRT surveys, participants may refuse to adhere to the instructions if there is an answer option that allows them to avoid being identified as a carrier. They therefore proposed the *Cheating Detection Model* (*CDM*) as an improvement over the forced-response procedure. In addition to considering carriers of the sensitive attribute who answer honestly (π) and noncarriers who answer honestly (β), it considers a third class of respondents, namely cheaters (γ) who respond "No" regardless of the outcome of the randomization procedure. Various studies have shown that the CDM provides higher and thus potentially more valid prevalence estimates of sensitive attributes than a direct question (e.g., Moshagen et al., 2010; Ostapczuk, Moshagen, et al., 2009; Ostapczuk & Musch, 2011; Ostapczuk, Musch, et al., 2009; Ostapczuk, Musch, & Moshagen, 2011; Pitsch, Emrich, & Klein, 2007). However, the CDM does not make any assumptions about the real status of cheaters; they may be either lying carriers or noncarriers who wish to avoid suspicion. Consequently, a precise estimate of the total prevalence of a sensitive attribute can be obtained only if the proportion of

cheaters is zero. Whenever cheating occurs ($\gamma > 0$), the prevalence of carriers of the sensitive attribute can be located anywhere within the range of π (if no cheater is a carrier) and $\pi + \gamma$ (if all cheaters carried the sensitive attribute; Clark & Desharnais, 1998). Thus, whenever $\gamma > 0$, the CDM provides only a lower (π) and an upper bound ($\pi + \gamma$) for the proportion of carriers. Several studies using the CDM have suggested that the proportion of cheaters on surveys covering sensitive topics may often be substantial and amount to up to 50% of the sample (e.g., Ostapczuk, Moshagen, et al., 2009; Ostapczuk & Musch, 2011; Ostapczuk, Musch, et al., 2009; Ostapczuk et al., 2011). On the one hand, this underlines the importance of a cheating detection approach to RRT surveys; on the other hand, this means that if the rate at which people cheat is substantial, the CDM allows for only a very rough estimate of the proportion of carriers in a given population. Moreover, when the CDM is applied, nothing is or can be said about the true status of respondents who have to be classified as cheaters according to the model.

Moshagen et al. (2012) recently introduced a new RRT model that is presumed to be capable of providing an estimate of the prevalence of carriers (π) and of the proportion of cheaters with a *known* status: the *Stochastic Lie Detector* (*SLD*). Based on a modification of the original RQM (Warner, 1965) by Mangat (1994), the randomization process in the SLD is restricted to the group of noncarriers. All respondents are presented with two Statements A (the sensitive statement) and B (the negation of statement A), but only carriers are asked to respond to Statement A unconditionally. Noncarriers are instructed to respond to Statement A with randomization probability p_i and to statement B with complementary probability $1 - p_i$. Moshagen et al. (2012) argued that when confronted with these instructions, carriers perceiving respondent jeopardy may have an incentive to disobey the instructions by responding “No” to the sensitive statement, whereas noncarriers should have no reason to lie about their status. To model

potential cheating among carriers, Moshagen et al. (2012) introduced a parameter *t* representing the proportion of carriers responding truthfully; the remaining carriers (1 - *t*) are assumed to be lying to conceal their true status. Figure 1 illustrates the tree diagram of the resulting SLD.

[Insert Figure 1 about here]

To allow for the estimation of the two parameters π and *t* in the SLD, two independent randomly drawn subsamples have to be assessed with different randomization probabilities $p_1 \neq p_2$ (Clark & Desharnais, 1998), with a larger difference of p_1 and p_2 resulting in a higher statistical efficiency of the model (Moshagen et al., 2012). Moshagen et al. (2012) showed the maximum likelihood estimates of π and *t* to be

$$\hat{\pi} = \frac{\left(\frac{n_2'}{n_2} - \frac{n_1'}{n_1}\right) + (p_2 - p_1)}{(p_2 - p_1)} \qquad (2)$$

and

$$\hat{t} = \frac{\left[\frac{n_2'}{n_2}(1 - p_1)\right] - \left[\frac{n_1'}{n_1}(1 - p_2)\right]}{\left(\frac{n_2'}{n_2} - \frac{n_1'}{n_1}\right) + (p_2 - p_1)} \qquad (3)$$

where n_1 and n_2 denote the sample sizes of the two samples tested with different randomization probabilities p_1 and p_2, and n_1' and n_2' represent the absolute frequencies of "Yes" responses in these groups. Equations deriving the variances of π and *t* were also provided by Moshagen et al. (2012).

The SLD was first applied in two pilot studies by Moshagen et al. (2012): In an experimental survey assessing the prevalence of domestic violence, the SLD yielded a prevalence estimate that was about four times higher than with direct questioning and more than

two times higher than with Mangat (1994) model. In addition, the estimated proportion of carriers responding truthfully (t) differed significantly from 100%, which indicated that a substantial number of carriers had decided to "play it safe" by choosing an answer option that would not make them look suspiciously like carriers (Moshagen et al., 2012). In a second experiment, estimates of the prevalence of nonvoting in the 2009 German federal elections obtained via DQ, the SLD, and Mangat (1994) model were compared with the known true proportion of nonvoters in the general population obtained by official statistics. Again, the SLD provided an estimate of the proportion of nonvoters that was higher than the ones provided by direct questioning and by applying Mangat (1994) model. Moreover, only the SLD estimate concurred almost exactly with the known true proportion of nonvoters (Moshagen et al., 2012).

The most compelling evidence supporting the validity of the SLD was provided in a recent validation study by Moshagen, Hilbig, Erdfelder, and Moritz (2014). In an adaption of the "die-under-the-cup" paradigm (cf. Hilbig & Hessler, 2013), participants were instructed to secretly roll a die and to report the outcome to the experimenter. Some of the outcomes were associated with a monetary reward. As the outcome of the individual die rolls was unknown to the questioner, the participants' actual behavior remained confidential. On the sample level, however, Moshagen et al. (2014) determined that cheaters comprised 53% of the alleged "winners." This known prevalence could then be used as an external criterion for the validation of the prevalence estimate obtained with the SLD and a DQ procedure. Moshagen et al. (2014) showed that a conventional DQ procedure substantially underestimated the known prevalence of cheaters (36%), whereas the application of the SLD resulted in an estimate of 48%, which did not differ significantly from the ground truth. In light of these results, Moshagen et al. (2014) considered the Stochastic Lie Detector to be a promising candidate within the class of advanced

RRT models. It is important to note, however, that the SLD offers a "safe" answer category because a "No" response can stem only from a noncarrier. If noncarriers are attracted to this answer to avoid the risk of suspicion, the model assumptions are violated, and distorted prevalence estimates are to be expected. We therefore felt it necessary to conduct a further validation of the SLD and to compare it with the competing Crosswise Model (Yu et al., 2008), which claims to counteract both respondent jeopardy *and* risk of suspicion.

The Crosswise Model

Within the last couple of years, a new class of so-called *nonrandomized response models* has been proposed (for an overview, see Tian & Tang, 2013). The goal of these models is to question the respondents indirectly without having to employ a randomization procedure such as the rolling of a die. With the *Crosswise Model* (*CWM*) as a member of this class, Yu et al. (2008) introduced a questioning technique that is arguably easier for the respondents to understand than other models. Moreover, the CWM holds the particular advantage of response symmetry because none of the answer options provides a "safe" alternative that clearly dispels the possibility of the respondent being a carrier of the sensitive attribute. In the CWM, participants are simultaneously presented with two statements: one statement referring to a sensitive attribute with unknown prevalence π and another statement referring to a nonsensitive attribute with known prevalence p (e.g., a statement about the month of the respondent's birth). Respondents are then asked to indicate whether "both statements are true or both statements are false" or whether "exactly one and only one of the two statements is true." Neither of these answer options directly indicates whether the respondent is a carrier of the sensitive attribute, and neither of them clearly marks the respondent as a noncarrier. Respondent jeopardy and risk of suspicion are thus thoroughly circumvented. Yu et al. (2008) argued that the clear and easy-to-understand rationale and the

convincing protection offered to the respondents by the symmetric CWM "would presumably not only make [them] willing to participate in the survey, but also persuade them to provide truthful responses" (p. 254). Response symmetry has, in fact, been shown to increase compliance with the instructions in other RRT models (e.g., Ostapczuk, Moshagen, et al., 2009). If response symmetry makes cheating-detection mechanisms such as the ones implemented in the CDM and the SLD dispensable, the CWM may be the model of choice for the assessment of sensitive attributes.

Because the CWM is mathematically equivalent to the model by Warner (1965), the maximum likelihood estimator for π is given by

$$\hat{\pi} = \frac{p - 1 + \frac{n'}{n}}{2p - 1} \quad , \qquad p \neq 1/2 \tag{4}$$

where p is the known prevalence of the nonsensitive statement, n' represents the total number of "both true or both false" responses, and n reflects the sample size. Equations deriving the variance of π are provided in Yu et al. (2008). Figure 2 illustrates the CWM as a tree diagram.

[Insert Figure 2 about here]

So far, a small number of publications have presented data from applications of the CWM. In two recently published studies, the CWM was applied without a direct questioning control group (Eslami et al., 2013; Vakilian, Mousavi, & Keramat, 2014). More relevant to the present research are studies comparing the CWM and a direct questioning procedure. In two such studies, the CWM yielded a higher and therefore arguably more valid prevalence estimate for plagiarism in student papers than direct questions (Coutts, Jann, Krumpal, & Näher, 2011; Jann, Jerke, & Krumpal, 2012). When assessing the incidence of tax evasion in small and medium Serbian firms,

Kundt, Misch, and Nerré (2013) also obtained significantly higher prevalence estimates when using the CWM than direct self-reports. The lifetime prevalence of anabolic steroid use in athletes was estimated as being more than two times higher when using the CWM rather than a direct question in a study by Nakhaee, Pakravan, and Nakhaee (2013). Jann et al. (2012) therefore evaluated the existing body of research as showing that "the [CWM] is successful in decreasing the social-desirability bias" (p. 13). It is important to note, however, that none of the existing studies provided a strong validation and direct evidence for the validity of the CWM because estimates obtained with the model were never compared with a known prevalence of carriers or noncarriers. If the CWM or the SLD does not provide correct estimates for the known prevalence of control attributes as well, the validity of the respective model will be called into question. We therefore decided to investigate whether the CWM and the SLD can correctly recover the known prevalence of a nonsensitive control attribute.

In contrast to models implementing a cheating detection device, the application of the CWM does not allow the user to test whether participants adhered to the instructions. Hence, it seemed worthwhile to compare its performance with the SLD as an alternative model that is not symmetrical but is rather based on a cheating detection procedure. To investigate the extent to which either model would succeed in motivating respondents to provide truthful answers to questions addressing a sensitive topic, we also included a direct questioning (DQ) control condition.

Xenophobia, Islamophobia, and the Influence of the Social Desirability Bias

We used a repeated measures design to compare the three questioning procedures (SLD, CWM, and DQ). To assess the ability of the different questioning techniques to control for social desirability, we included two questions pertaining to sensitive issues and a control question

pertaining to an issue that was nonsensitive in nature but for which the true prevalence was known from official statistics. The two sensitive questions referred to xenophobia and islamophobia, respectively.

Xenophobia (i.e., a negative attitude toward people with an immigration background) has been shown to be an attitude that is rather widespread but that is usually met with social disapproval in Germany. Whereas self-reports using direct questioning procedures have shown only a moderate level of xenophobia in Germany that was comparable to the level observed in other Western European countries (Zick, Küpper, & Hövermann, 2011), Klink and Wagner (1999) demonstrated a heavy discrimination against ethnic minorities in a series of field experiments in which they manipulated the names, accents, and appearances of confederates presumably seeking help in everyday situations. Confederates mimicking a Turkish immigration background were far less likely to obtain a viewing appointment for a vacant house or to receive help in several situations in which they needed support. Regarding the frequently observed deviance of self-reports and actual behavior, Hjerm (1998) commented:

> "The problem of social desirability is obviously important in studies that deal with such issues as xenophobia. It is possible that although they respond anonymously, people give socially desirable answers so as not to appear xenophobic. This might lead to an underestimation of the actual prevalence of xenophobia in a society." (p. 338)

Krumpal's (2012) results support this conjecture. Using a forced-response variant of the RRT (Boruch, 1971; Dawes & Moore, 1980) in a German telephone survey, he found that the RRT produced higher estimates of xenophobia than conventional DQ methods. Similarly, Ostapczuk, Musch, et al. (2009) showed that in a German sample, the proportion of xenophobes was

substantially higher under the truth-eliciting CDM questioning procedure (Clark & Desharnais, 1998) than under a direct questioning procedure. They therefore concluded that the participants seemed to be "less unprejudiced than their answers to a direct question had suggested" (Ostapczuk, Musch, et al., 2009, p. 928). The question we used to assess the prevalence of xenophobia read: "I would mind if my daughter had a relationship with a Turkish man." This question was modeled after Bogardus (1933) and had been used before with other ethnic minorities by Silbermann and Hüsers (1995), Jimenez (1999), and Ostapczuk, Musch, et al. (2009).

As a second sensitive attribute, we assessed islamophobia, that is, a negative attitude toward, or even a fear of, people of the Muslim religion. Islamophobia is widespread in European countries (e.g., EUMC - European Monitoring Center on Racism and Xenophobia, 2006; Savelkoul, Scheepers, van der Veld, & Hagendoorn, 2012; Sheridan, 2006; Zick et al., 2011) and has been argued to be one of the most important political issues in modern Europe, possibly even "[m]uch more pressing" than anti-Semitism (Bunzl, 2005, p. 506). Even though the German constitution guarantees religious freedom, Germany is one of the highest ranked European countries in anti-Muslim attitudes (Zick et al., 2011). A strong connection between islamophobia and negative attitudes toward the construction of Muslim religious buildings was recently reported by Imhoff and Recker (2012). Individual scores of German participants on an Islamoprejudice subscale proved highly predictive of negative attitudes toward the construction of a great new mosque in the city of Cologne. We therefore decided to use an item that asked for negative attitudes toward the construction of minarets in Germany. This item was chosen because, in a recent popular vote, the citizens of Switzerland had voted in favor of a constitutional addendum that prohibited any further construction of minarets. The result of this popular vote

had not been predicted by representative polls (gfs.bern, 2009a, 2009b; reformiert, 2009), arguably because voters refrained from revealing what had been stigmatized as an attitude that tends to be met with social disapproval in the debate preceding the poll (fög, 2010).

Umesh and Peterson (1991) have argued that "[s]tudies that compared the RR[T] with other forms of questioning […] are not validation studies" and that a "true validation study must compare the randomized response estimate and the actual value" (p. 127). Two such "strong" validation studies have been conducted for the SLD (Moshagen et al., 2014; Moshagen et al., 2012), but such studies have yet to be reported for the CWM. Therefore, one goal of the present study was to investigate whether the SLD and the CWM would be capable of recovering the known prevalence of an attribute. Unfortunately, however, the ground truth for sensitive attributes is usually unknown and difficult to obtain, as reflected in the relatively small number of only six "strong" validation studies that compared Randomized Response estimates with a known prevalence as reported in Lensvelt-Mulders et al.'s (2005) meta-analysis. In one of these studies that reported on social security fraud, the assessment of a sample that had a true prevalence of carriers of 100% was possible only because of the public availability of databases containing the addresses of people who had previously been convicted of such crimes in the Netherlands (van der Heijden et al., 2000). No such databases are available in Germany, however. Because there was no way to know the true value of a sensitive attribute in our student sample, we included a nonsensitive control question that pertained to the first letter of the respondents' surname, for which the incidence in the general population could be determined. This allowed us to go beyond the usual "more is better" validation approach and to detect method-specific biases in the assessment of the prevalence of an attribute. Official statistics from the German *Statistisches Bundesamt* (*Federal Office of Statistics*) show that the proportion of citizens in

Germany with a surname that begins with one of the relatively frequent letters K, L, M, R, S, or T is about 43% (Reinders, 1996). If the SLD and the CWM are capable of obtaining valid prevalence estimates of sensitive attributes, they should also perform well when applied to a nonsensitive control attribute.

To summarize, the present experiment addressed the following two questions: (a) Are the SLD and the CWM capable of controlling for social desirability? To the extent to which they are, the two indirect questioning techniques were expected to provide higher prevalence estimates of the two sensitive attributes than a direct question. (b) Are the SLD and the CWM prone to a method-specific bias that results in systematic over- or underestimates? If so, the two indirect questioning techniques should provide estimates that are at odds with official statistics with regard to the prevalence of surnames that begin with certain letters.

Method

A total of 1,312 subjects volunteered to participate in our survey. The sample (56% female, *mean age* = 21.21 years, *SD* = 3.14) consisted of students from three German universities (Duesseldorf 81%, Duisburg 10%, and Bochum 9%) who were recruited and assessed in groups in lecture halls before classes began.

Survey Design

Respondents filled out a one-page questionnaire consisting of a short introduction, the three (sensitive and nonsensitive) experimental questions, and two demographic questions asking for the respondents' age and gender. The questioning technique was varied as an independent within-subjects variable and consisted of the *SLD* (randomization device: mother's month of birth; subdivided into two groups with low vs. high randomization probabilities of $p_1 = .158$ vs. $p_2 = .842$, respectively), the *CWM* (randomization device: father's month of birth; randomization

probability $p = .158$), and the *DQ* format. The question format was determined randomly for each question with the constraint that each of the three questions was answered in a different format. Two questions referred to sensitive attributes (xenophobia/negative attitudes toward Turkish immigrants; islamophobia/negative attitudes toward the construction of minarets in Germany) with unknown prevalences π_{s1} and π_{s2}, respectively. The third question referred to the first letter of the respondents' surname as a nonsensitive control attribute. The prevalence of this nonsensitive attribute was known ($\pi_{ns} = .43$) because it could be obtained from official statistics for the set of letters that was used for this question (first letter K, L, M, R, S, or T; Reinders, 1996). Examples of all three questioning formats are given below.

SLD format. For the question referring to xenophobia, the SLD format (with a low randomization probability of $p1 = .158$) read as follows:

Assume that you have a 20-year-old daughter: Would you mind if she had a relationship with a Turkish man? If yes, please respond to Statement A. If not, please respond to…

- Statement A if your mother was born in November or December,
- Statement B if your mother was born in any other month.

Statement A: I would mind if my daughter had a relationship with a Turkish man.

Statement B: I would *not* mind if my daughter had a relationship with a Turkish man.

Answer options: *True* versus *False*.

For the two other topics, the SLD questioning format was adapted accordingly.

CWM format. For the question referring to islamophobia, the CWM format (with a randomization probability of $p = .158$) read as follows:

Statement A: The construction of minarets should be prohibited in Germany.

Statement B: My father was born in November or December.

Answer options: *Both true or both false* versus *Exactly one true (regardless of which one)*.

For the two other topics, the CWM format was adapted accordingly.

DQ format. For the nonsensitive control question with known prevalence (π_{ns} = .43), the DQ format read as follows:

Statement: My surname begins with one of the following letters: K, L, M, R, S, or T.

Answer options: *True* versus *False*.

For the two sensitive questions, the DQ format was adapted accordingly.

Statistical Analysis

Following the procedure detailed in Moshagen, Hilbig, and Musch (2011), Moshagen and Musch (2012), Moshagen et al. (2012), Moshagen et al. (2010), Ostapczuk, Moshagen, et al. (2009), Ostapczuk and Musch (2011), Ostapczuk, Musch, et al. (2009), and Ostapczuk et al. (2011), multinomial processing tree models (MPT; Batchelder, 1998; Batchelder & Riefer, 1999) were formulated for all three questioning techniques. Within the multinomial modeling framework and using the procedures detailed in Hu and Batchelder (1994), it was possible to estimate the prevalence parameters for each questioning technique and to conduct the necessary statistical tests of our hypotheses. On the basis of the empirically observed answer frequencies in the different experimental conditions, we computed maximum likelihood estimates for all parameters using the expectation-maximization algorithm (EM; Dempster, Laird, & Rubin, 1977; Hu & Batchelder, 1994) implemented in the software multiTree (Moshagen, 2010). The model fit was tested via the asymptotically χ^2-distributed log-likelihood statistic G^2. The MPT models for all three questioning techniques were saturated with $df = 0$ and $G^2 = 0$ as the number of independent answer categories just sufficient to estimate all parameters in the three questioning technique conditions: The two proportions of "Yes" responses in the conditions with a low versus

high randomization probability allowed us to estimate the two parameters π and *t* in the *SLD* condition; the proportion of "Both true or both false" responses allowed us to estimate π in the *CWM* condition; and the proportion of "Yes" responses allowed us to estimate π in the DQ condition. Comparisons between parameter estimates and comparisons between the parameters and a constant were conducted by assessing the significance of the difference in model fit (ΔG^2) between an unrestricted baseline model and an alternative model in which either the two parameters under question were restricted to be equal or one parameter was set to a constant value (e.g., $\pi_{ns} = .43$). The multinomial model equations and the observed answering frequencies for all conditions are given in Appendices A and B.

Results

Xenophobia (Sensitive Attribute 1)

Table 1 shows the prevalence estimates for the xenophobia item obtained via DQ, the SLD, and the CWM.

[Insert Table 1 about here]

Pairwise comparisons between questioning techniques revealed that in comparison with the DQ condition (26.98%), respondents were more likely to answer truthfully in both the SLD (53.38%) and the CWM conditions (48.67%), ΔG^2 ($df = 1$) = 16.80, $p < .001$ and ΔG^2 ($df = 1$) = 28.20, $p < .001$, respectively. This pattern suggests that the prevalence of xenophobia was presumably underestimated in the DQ condition. A comparison of the two indirect questioning techniques revealed no significant difference in the prevalence estimates between the SLD and the CWM conditions, ΔG^2 ($df = 1$) = 0.43, $p = .51$. The SLD estimated the proportion of carriers of the sensitive attribute answering honestly at $t = .79$, a value significantly below 1.0, ΔG^2 ($df =$

1) = 14.56, $p < .001$. This finding suggests that according to the SLD, a substantial proportion of 21% of the carriers seems to have cheated in an attempt to conceal their true status.

Islamophobia (Sensitive Attribute 2)

Table 2 contains the prevalence estimates for the islamophobia item obtained via the three questioning methods.

[Insert Table 2 about here]

As for the xenophobia item, the pattern of results suggests an underestimation of the prevalence in the DQ condition. In both the SLD (76.93%) and CWM (51.64%) conditions, the proportion of respondents with negative attitudes was estimated as higher than in the DQ condition (43.33%), ΔG^2 ($df = 1$) = 23.97, $p < .001$ and ΔG^2 ($df = 1$) = 3.89, $p < .05$, respectively. However, unlike for the xenophobia item, the two indirect questioning techniques showed diverging results: In the SLD condition, the estimated proportion of carriers was significantly higher than in the CWM condition, ΔG^2 ($df = 1$) = 11.80, $p < .001$. The SLD estimated the proportion of carriers of the sensitive attribute answering honestly at $t = .68$, a value significantly below 1.0, ΔG^2 ($df = 1$) = 65.07, $p < .001$. This finding suggests that a substantial proportion of 32% of the carriers seems to have cheated on the survey to conceal their true status.

Nonsensitive Control Attribute with Known Prevalence: First Letter of Surname

The parameter estimates for the nonsensitive attribute are shown in Table 3.

[Insert Table 3 about here]

As expected for a nonsensitive attribute, there was no significant difference between the prevalence estimates obtained via DQ (40.99%) and the CWM (46.57%), ΔG^2 ($df = 1$) = 1.73, $p = .19$. Unexpectedly, however, the SLD estimate (62.72%) was significantly higher than both the DQ and CWM estimates, ΔG^2 ($df = 1$) = 11.00, $p < .001$ and ΔG^2 ($df = 1$) = 5.15, $p < .05$,

respectively. The DQ and CWM estimates deviated only slightly and nonsignificantly from the known prevalence of $\pi_{ns} = .43$, ΔG^2 ($df = 1$) = 0.73, $p = .39$ and ΔG^2 ($df = 1$) = 1.02, $p = .31$, respectively. By contrast, the SLD significantly overestimated the known prevalence, ΔG^2 ($df = 1$) = 10.42, $p < .01$. The SLD estimated the proportion of carriers of the sensitive attribute answering honestly at $t = .78$, a value significantly below 1.0, ΔG^2 ($df = 1$) = 23.16, $p < .001$, indicating that approximately 22% of the carriers of the nonsensitive attribute seemed to have disobeyed the instructions.

Discussion

Social desirability bias may lead to the underreporting of socially undesirable attributes. The present study investigated the validity of two competing indirect questioning techniques, the Stochastic Lie Detector (SLD; Moshagen et al., 2012) and the Crosswise Model (CWM; Yu et al., 2008), both of which aim to experimentally address the problem of social desirability bias. According to the "more is better" criterion, higher estimates of socially undesirable attributes can be considered more valid as they presumably suffer less from distortion. Using a large-scale survey, we therefore assessed whether the application of the two indirect questioning techniques would result in higher prevalence estimates than a conventional direct questioning (DQ) approach for two sensitive statements. Because the "more is better" criterion fails if a questioning technique provides estimates that surpass the known prevalence of a criterion, we also tested whether the application of the SLD or the CWM would result in undistorted estimates of a third nonsensitive control attribute. To the extent to which estimates provided by an indirect questioning technique are higher than the actual known prevalence of a control attribute, the validity of this indirect questioning technique is called into question.

With regard to the xenophobia item, which referred to being willing to let one's daughter marry a Turkish man, both indirect questioning techniques outperformed the direct question. When questioned directly, only about 27% of the respondents admitted to negative attitudes toward Turkish men. By contrast, both the SLD (53%) and the CWM (49%) yielded prevalence estimates that were approximately twice as high and thus presumably more valid than the estimate from the direct question. The SLD estimated the prevalence of carriers responding truthfully to be substantially lower than 100%. This finding suggests that the topic of xenophobia seems to be rather sensitive in nature, and this can also explain why ignoring the social desirability bias by relying on a direct question would have resulted in an underestimate. Apparently, respondents were considerably more willing to respond truthfully when they were granted full confidentiality for their answers.

A quite similar pattern of results was observed for the islamophobia item, which referred to the construction of minarets in Germany. Whereas only 43% of the directly questioned respondents provided an islamophobic answer to this question, the CWM estimated the true prevalence of islamophobia to be significantly higher at 52%. The SLD estimate, however, even surpassed the CWM estimate and located the prevalence of xenophobic participants at a surprising 77%. The SLD model estimated the proportion of carriers answering truthfully at 68%. These results add to the evidence that suggests that the self-report of both xenophobic and islamophobic attitudes may be distorted by a social desirability bias and that indirect questioning techniques may be capable of yielding more valid prevalence estimates.

It is interesting to note that our results would predict diverging outcomes for a hypothetical popular vote on the issue under investigation. On the basis of the results of the direct question condition, one would have to predict that a majority would vote against the

introduction of a law prohibiting the construction of minarets; according to the results obtained in the SLD and CWM conditions, however, one would have to predict that the proposal of such a law would pass a referendum. The latter result was in fact the outcome of a popular vote conducted in Switzerland in 2009, a result that was generally considered surprising because a poll had predicted the opposite outcome just prior to the vote. This poll, however, had been based on a direct question.

In summary, the results we obtained for the two sensitive questions attest to the validity of the indirect questioning techniques with regard to the "more is better" criterion. The application of both indirect questioning techniques resulted in higher and therefore presumably more valid prevalence estimates for the two sensitive topics under investigation.

To determine whether a method bias that would result in a general tendency to over- or underestimate the prevalence of any attribute is inherent to either the SLD or the CWM, we included a control question that referred to a nonsensitive attribute with a known prevalence. In accordance with the assumption of no bias, the CWM yielded a prevalence estimate (47%) that was fairly close to and not significantly different from the known true prevalence of 43%. Supporting the validity of this estimate, the CWM estimate did not differ significantly from the estimate yielded by the direct question (41%). This result was to be expected considering that the first letter of a person's surname is not a sensitive attribute, and corresponding self-reports should therefore not be distorted by social desirability bias. Thus, the validity of the CWM was confirmed with regard to both better control over social desirability bias as compared with a direct question and the lack of a method bias resulting in a systematic tendency to over- or underestimate.

Unlike the CWM, however, the SLD substantially overestimated the known prevalence of the control attribute (SLD: 63% vs. true: 43%). The SLD estimate also differed significantly from the estimate yielded by the direct question, which closely mirrored the known true prevalence of the nonsensitive control attribute (DQ: 41% vs. true: 43%). The proportion of carriers responding truthfully on the SLD was estimated at 78%, which is significantly lower than the 100% that would have to be expected if all participants had completely complied with the instructions.

Several alternative explanations for this unexpected outcome seem possible. First, the SLD may have a harmful tendency to overestimate the prevalence of any given attribute. Holbrook and Krosnick (2010) called the validity of the RRT method into question when obtaining an estimate for the prevalence of a socially desirable attribute that was unexpectedly higher than the corresponding estimate obtained with a direct question and even reached "impossible levels" (Holbrook & Krosnick, 2010, p. 336) of over 100%. Wolter and Preisendörfer (2013), however, argued that to draw general conclusions regarding the validity of the RRT might be premature. When assessing the validity of the SLD, it has to be taken into account that the technique performed well in two studies by Moshagen et al. (2012) and Moshagen et al. (2014), both of which found that the SLD provided estimates that converged with the known prevalence of a sensitive attribute. Moreover, the SLD performed well for the xenophobia item in the current study, providing an estimate that converged with the estimate obtained using the CWM, which in turn provided presumably valid estimates for all questions in the present investigation.

An alternative explanation for why the SLD did not yield valid results for all questions in the present study may be found in its specific implementation. Although the SLD was designed

to address one particular type of nonadherence to instructions, namely, untruthful responding by carriers of a sensitive attribute, its assumptions are clearly violated if (a) noncarriers falsely claim to carry the attribute, (b) carriers strategically use the randomization procedure to appear as though they are noncarriers, (c) response behavior varies for different randomization probabilities, or (d) respondents generally fail to understand and follow the instructions (cf. Moshagen et al., 2012). Any of the above problems can lead to distorted prevalence estimates. However, given that the surname control item was nonsensitive in nature, the three potential violations described in (a), (b), and (c) would be unlikely causes of the observed distortion. Moreover, as pointed out by Moshagen et al. (2012), violations of the assumptions according to (b) and (c) should have led to an under- rather than an overestimation of the attribute's prevalence. A general failure to understand and follow the instructions, however, might offer a potential explanation for the present findings. Various researchers have pointed out that RRT questions may be difficult for some participants to understand (e.g., Landsheer, van der Heijden, & van Gils, 1999; Locander et al., 1976; van der Heijden, van Gils, Bouts, & Hox, 1998). The validity of an RRT estimate, however, strongly depends on the participants' comprehension of the instructions (e.g., Abul-Ela et al., 1967; Holbrook & Krosnick, 2010; Soeken & Macready, 1982). In a recent survey using the unrelated question variant of the RRT, James et al. (2013) surmised that a misunderstanding of the instructions might have led to the inflated estimates they obtained for the use of performance-enhancing drugs.

Another potential reason for the performance problems we observed for the SLD might be that, unlike the CWM, the SLD does not offer response symmetry to the respondents. Using the SLD, it is possible to respond in a way that allows the respondent to avoid looking suspicious of being a carrier of the sensitive attribute. This possibility to "play it safe" may lead to distorted

response behavior. However, the distortion we observed occurred for a question that was nonsensitive in nature. Thus, a tendency to "play it safe" can hardly explain why we obtained an overestimate for a nonsensitive control item using the SLD. However, it is conceivable that the application of the SLD to a nonthreatening control question may have seemed odd to some of the respondents. This may have led to some confusion or even to a rejection of the method, but unfortunately, no post hoc test of this explanation was possible with the data we collected. It should however be noted that any response behavior that deviated from the instructions—including random responses—that equally extends to both the low and high randomization probability conditions can be shown to necessarily lead to an overestimation when using the SLD whenever $\pi < 1.00$. Therefore, it seems necessary to provide a more systematic investigation of the comprehensibility of RRT questions and compliance with the instructions in future research. Judging from the present results, it would appear that both comprehensibility and compliance with the instructions might be better for the CWM than for the SLD.

In conclusion, our results suggest that the CWM offers a valid and useful means for achieving the experimental control of social desirability. Our results also tentatively suggest that the CWM might be superior to the SLD with regard to applicability and validity. Even though both models met the "more is better" criterion in the assessment of two sensitive attributes, only the CWM succeeded in estimating the known prevalence of a nonsensitive control attribute. On the basis of this pattern of results, it seems justifiable to recommend the use of the CWM in future studies investigating sensitive issues.

References

Abernathy, J. R., Greenberg, B. G., & Horvitz, D. G. (1970). Estimates of Induced Abortion in Urban North-Carolina. *Demography, 7*(1), 19-29.

Abul-Ela, A. L. A., Greenberg, B. G., & Horvitz, D. G. (1967). A Multi-Proportions Randomized Response Model. *Journal of the American Statistical Association, 62*, 900-1008.

Akers, R. L., Massey, J., Clarke, W., & Lauer, R. M. (1983). Are Self-Reports of Adolescent Deviance Valid? Biochemical Measures, Randomized-Response, and the Bogus Pipeline in Smoking-Behavior. *Social Forces, 62*, 234-251.

Antonak, R. F., & Livneh, H. (1995). Randomized-Response Technique - a Review and Proposed Extension to Disability Attitude Research. *Genetic, Social, and General Psychology Monographs, 121*, 97-145.

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

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.

Bunzl, M. (2005). Between anti-Semitism and Islamophobia: Some thoughts on the new Europe. *American Ethnologist, 32*, 499-508. doi: DOI 10.1525/ae.2005.32.4.499

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.

Chaudhuri, A., & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

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

Coutts, E., Jann, B., Krumpal, I., & Näher, A. F. (2011). Plagiarism in Student Papers: Prevalence Estimates Using Special Techniques for Sensitive Questions. *Jahrbücher für Nationalökonomie Und Statistik, 231*, 749-760.

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]*. 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.

Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., & Ulrich, R. (2013). Randomized Response Estimates for the 12-Month Prevalence of Cognitive-Enhancing Drug Use in University Students. *Pharmacotherapy, 33*, 44-50.

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

Eriksson, S. A. (1973). A New Model for Randomized Response. *International Statistical Review, 41*, 101-113.

Eslami, M., Yazdanpanah, M., Taheripanah, R., Andalib, P., Rahimi, A., & Nakhaee, N. (2013). Importance of Pre-pregnancy Counseling in Iran: Results from the High Risk Pregnancy Survey 2012. *International Journal of Health Policy and Management, 1*, 213-218.

EUMC - European Monitoring Center on Racism and Xenophobia. (2006). Muslims in the European Union: Discrimination and Islamophobia. Vienna, Austria: FRA.

Fidler, D. S., & Kleinknecht, R. E. (1977). Randomized Response Versus Direct Questioning - 2 Data-Collection Methods for Sensitive Information. *Psychological Bulletin, 84*, 1045-1049.

fög. (2010). Berichterstattung zur Volksinitiative 'Gegen den Bau von Minaretten'. Retrieved June 8th, 2010, from http://www.foeg.unizh.ch

Fox, J. A., & Tracy, P. E. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.

Fox, J. P., & Meijer, R. R. (2008). Using Item Response Theory to Obtain Individual Information From Randomized Response Data: An Application Using Cheating Data. *Applied Psychological Measurement, 32*, 595-610. doi: 10.1177/0146621607312277

Franklin, L. (1998). Randomized Response Techniques. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (Vol. 5, pp. 3696-3703). New York: Wiley.

gfs.bern. (2009a). 'Minarett-Initiative': Das Nein überwiegt – SVP-Wählerschaft dafür. Retrieved June 6th, 2010, from www.gfsbern.ch

gfs.bern. (2009b). 'Minarett-Initiative': Ja nimmt zu - Nein unverändert stärker. Retrieved June 8th, 2010, from www.gfsbern.ch

Goodstadt, M. S., & Gruson, V. (1975). Randomized Response Technique - Test on Drug-Use. *Journal of the American Statistical Association, 70*, 814-818.

Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R., & Horvitz, D. G. (1969). Unrelated Question Randomized Response Model - Theoretical Framework. *Journal of the American Statistical Association, 64*, 520-539.

Greenberg, B. G., Horvitz, D. G., & Abernathy, J. R. (1974). A comparison of randomized response designs. *Reliability and biometry, statistical analysis of life length* (pp. 787-815). Philadelphia: SIAM.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., & Horvitz, D. G. (1971). Application of Randomized Response Technique in Obtaining Quantitative Data. *Journal of the American Statistical Association, 66*, 243-250.

Hejri, S. M., Zendehdel, K., Asghari, F., Fotouhi, A., & Rashidian, A. (2013). Academic disintegrity among medical students: a randomised response technique study. *Medical Education, 47*, 144-153. doi: 10.1111/Medu.12085

Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology, 49*, 263-266. doi: 10.1016/j.jesp.2012.11.010

Himmelfarb, S., & Edgell, S. E. (1980). Additive Constants Model - a Randomized-Response Technique for Eliminating Evasiveness to Quantitative Response Questions. *Psychological Bulletin, 87*, 525-530. doi: 10.1037//0033-2909.87.3.525

Hjerm, M. (1998). National identities, national pride and xenophobia: A comparison of four Western countries. *Acta Sociologica, 41*, 335-347.

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

Horvitz, D. G., Greenberg, B. G., & Abernathy, J. R. (1976). Randomized Response - Data-Gathering Device for Sensitive Questions. *International Statistical Review, 44*, 181-196.

Horvitz, D. G., Shah, S., & Simmons, W. R. (1967). The Unrelated Question Randomized Response Model. *Proceedings of the Social Statistics Section, American Statistical Association*.

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

IIT Research Institute and the Chicago Crime Commission. (1971). A study of organized crime in Chicago. Chicago: IITRI Project No. H-6031, Report prepared for the Illinois Enforcement Commission.

Imhoff, R., & Recker, J. (2012). Differentiating Islamophobia: Introducing a New Scale to Measure Islamoprejudice and Secular Islam Critique. *Political Psychology, 33*, 811-824. doi: 10.1111/j.1467-9221.2012.00911.x

James, R. A., Nepusz, T., Naughton, D. P., & Petroczi, A. (2013). A potential inflating effect in estimation models: Cautionary evidence from comparing performance enhancing drug and herbal hormonal supplement use estimates. *Psychology of Sport and Exercise, 14*, 84-96. doi: 10.1016/j.psychsport.2012.08.003

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

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.

Klink, A., & Wagner, U. (1999). Discrimination against ethnic minorities in Germany: Going back to the field. *Journal of Applied Social Psychology, 29*, 402-423. doi: 10.1111/j.1559-1816.1999.tb01394.x

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

Kuk, A. Y. C. (1990). Asking Sensitive Questions Indirectly. *Biometrika, 77*, 436-438. doi: 10.1093/biomet/77.2.436

Kulka, R. A., Weeks, M. F., & Folsom, R. E. (1981). *A comparison of the randomized response approach and direct questioning approach to asking sensitive survey questions*. Working paper. NC: Research Triangle Institute.

Kundt, T. C., Misch, F., & Nerré, B. (2013). Re-assessing the merits of measuring tax evasions through surveys: Evidence from Serbian firms. ZEW Discussion Papers, No. 13-047. Retrieved Dec 12th, 2013, from http://hdl.handle.net/10419/78625

Landsheer, J. A., van der Heijden, P., & 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

Liu, P. T., & Chow, L. P. (1976). A New Discrete Quantitative Randomized Response Model. *Journal of the American Statistical Association, 71*, 72-73. doi: 10.2307/2285733

Liu, P. T., Chow, L. P., & Mosley, W. H. (1975). Use of Randomized Response Technique with a New Randomizing Device. *Journal of the American Statistical Association, 70*, 329-332.

Locander, W., Sudman, S., & Bradburn, N. (1976). An Investigation of Interview Method, Threat and Response Distortion. *Journal of the American Statistical Association, 71*, 269-275. doi: 10.2307/2285297

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

Mangat, N. S., & Singh, R. (1990). An Alternative Randomized-Response Procedure. *Biometrika, 77*, 439-442. doi: 10.1093/biomet/77.2.439

Marquis, K. H., Marquis, M. S., & Polich, J. M. (1986). Response Bias and Reliability in Sensitive Topic Surveys. *Journal of the American Statistical Association, 81*, 381-389. doi: 10.2307/2289227

Moors, J. J. A. (1971). Optimization of Unrelated Question Randomized Response Model. *Journal of the American Statistical Association, 66*, 627-629.

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

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

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*(3-4), 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.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and Denial in Socially Desirable Responding. *Journal of Personality and Social Psychology, 60*, 307-317. doi: 10.1037/0022-3514.60.2.307

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

Pollock, K. H., & Bek, Y. (1976). A Comparison of 3 Randomized Response Models for Quantitative Data. *Journal of the American Statistical Association, 71*, 884-886. doi: 10.2307/2286855

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W. C., & Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology, 13*, 465-484. doi: 10.1002/(Sici)1099-0720(199910)13:5<465::Aid-Acp609>3.0.Co;2-Y

reformiert. (2009). Mehrheit ist gegen ein Minarettverbot. Retrieved June 8th, 2010, from www.ref.ch

Reinders, M. (1996). Häufigkeit von Namensanfängen. *Statistische Rundschau Nordrhein-Westfalen, 11*, 651-660.

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

Scheers, N. J. (1992). A Review of Randomized-Response Techniques. *Measurement and Evaluation in Counseling and Development, 25*, 27-41.

Sheridan, L. P. (2006). Islamophobia pre- and post-September 11th, 2001. *Journal of Interpersonal Violence, 21*, 317-336. doi: 10.1177/0886260505282885

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.

Simon, P., Striegel, H., Aust, F., Dietz, K., & Ulrich, R. (2006). Doping in fitness sports: estimated number of unreported cases and individual probability of doping. *Addiction, 101*, 1640-1644. doi: 10.1111/j.1360-0443.2006.01568.x

Soeken, K. L., & Damrosch, S. P. (1986). Randomized-Response Technique - Applications to Research on Rape. *Psychology of Women Quarterly, 10*, 119-125. doi: 10.1111/j.1471-6402.1986.tb00740.x

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

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.

Striegel, H., Ulrich, R., & Simon, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence, 106*, 230-232. doi: 10.1016/j.drugalcdep.2009.07.026

Sudman, S., & Bradburn, N. (1974). *Response effects in surveys*. Chicago: Aldine.

Tian, G. L., & Tang, M. L. (2013). *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

Tracy, D. S., & Mangat, N. S. (1996). Some development in randomized response sampling during the last decade - a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science, 4*, 147-158.

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.

Vakilian, K., Mousavi, S. A., & Keramat, A. (2014). Estimation of sexual behavior in the 18-to-24-years-old Iranian youth based on a crosswise model study. *BMC Research Notes, 7*(28), 1-4.

van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (1998). A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of social security fraud. *Kwantitatieve Methoden, 19*, 15-34.

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.

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

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. In N. Langenbacher (Ed.). Berlin: Friedrich-Ebert-Stiftung.

Table 1

Prevalence Estimates (Standard Errors in Parentheses) for Sensitive Attribute 1: "I would mind if my daughter had a relationship with a Turkish man"

Parameter	Direct questioning	Stochastic Lie Detector	Crosswise Model
π_{s1}	26.98% (2.11)	53.38% (6.31)	48.67% (3.48)
t	–	79.43% (4.59)	–

Table 2

Prevalence Estimates (Standard Errors in Parentheses) for Sensitive Attribute 2: "The construction of minarets should be prohibited in Germany"

Parameter	Direct questioning	Stochastic Lie Detector	Crosswise Model
π_{s2}	43.33% (2.40)	76.93% (6.62)	51.64% (3.46)
t	–	67.94% (3.19)	–

Table 3

Prevalence Estimates (Standard Errors in Parentheses) for the Nonsensitive Control Attribute: "My surname begins with one of the following letters: K, L, M, R, S, or T" (known prevalence $\pi_{ns} = .43$)

Parameter	Direct questioning	Stochastic Lie Detector	Crosswise Model
π_{ns}	40.99% (2.33)	62.72% (6.25)	46.57% (3.54)
t	–	78.23% (3.92)	–

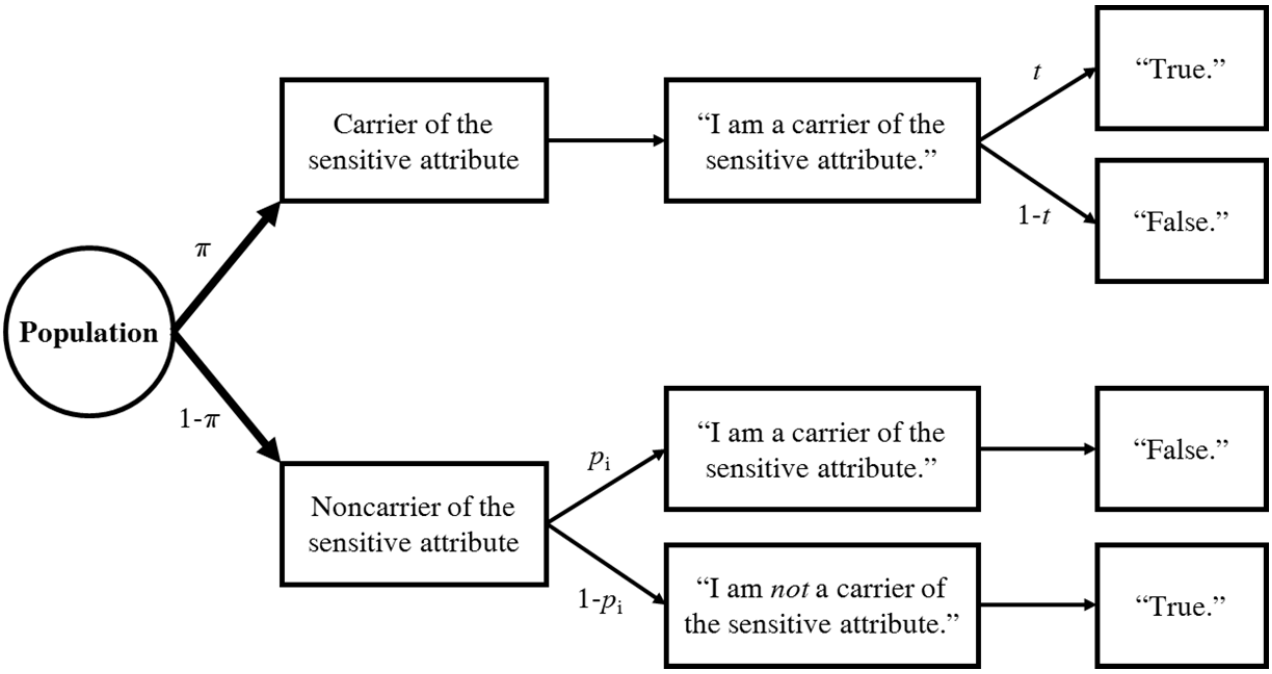

Figure 1. Tree diagram of the Stochastic Lie Detector (Moshagen et al., 2012).

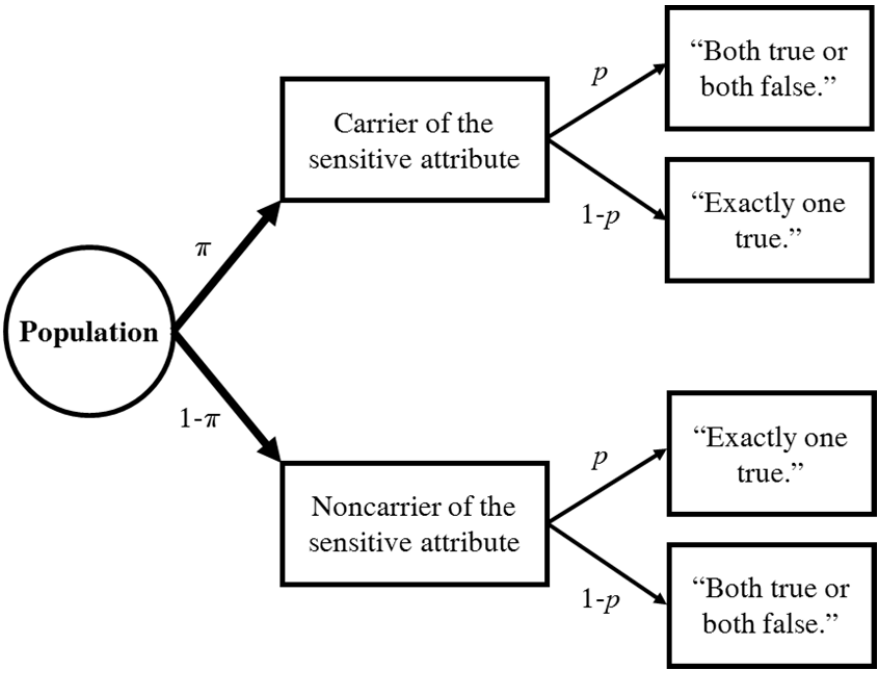

Figure 2. Tree diagram of the Crosswise Model (Yu et al., 2008).

Appendix A

Multinomial model equations established for parameter estimation in multiTree (Moshagen, 2010):

xenoph_SLDp1	xenoph_SLDp1_true	xenoph_PiSLD*xenoph_t
xenoph_SLDp1	xenoph_SLDp1_false	xenoph_PiSLD*(1-xenoph_t)
xenoph_SLDp1	xenoph_SLDp1_false	(1-xenoph_PiSLD)*p1
xenoph_SLDp1	xenoph_SLDp1_true	(1-xenoph_PiSLD)*(1-p1)
xenoph_SLDp2	xenoph_SLDp2_true	xenoph_PiSLD*xenoph_t
xenoph_SLDp2	xenoph_SLDp2_false	xenoph_PiSLD*(1-xenoph_t)
xenoph_SLDp2	xenoph_SLDp2_false	(1-xenoph_PiSLD)*p2
xenoph_SLDp2	xenoph_SLDp2_true	(1-xenoph_PiSLD)*(1-p2)
xenoph_CW	xenoph_CW_bothtrueorfalse	xenoph_PiCW*p1
xenoph_CW	xenoph_CW_onetrue	xenoph_PiCW*(1-p1)
xenoph_CW	xenoph_CW_onetrue	(1-xenoph_PiCW)*p1
xenoph_CW	xenoph_CW_bothtrueorfalse	(1-xenoph_PiCW)*(1-p1)
xenoph_DQ	xenoph_DQ_true	xenoph_PiDQ
xenoph_DQ	xenoph_DQ_false	(1-xenoph_PiDQ)
islamoph_SLDp1	islamoph_SLDp1_true	islamoph_PiSLD*islamoph_t
islamoph_SLDp1	islamoph_SLDp1_false	islamoph_PiSLD*(1-islamoph_t)
islamoph_SLDp1	islamoph_SLDp1_false	(1-islamoph_PiSLD)*p1
islamoph_SLDp1	islamoph_SLDp1_true	(1-islamoph_PiSLD)*(1-p1)
islamoph_SLDp2	islamoph_SLDp2_true	islamoph_PiSLD*islamoph_t
islamoph_SLDp2	islamoph_SLDp2_false	islamoph_PiSLD*(1-islamoph_t)

islamoph_SLDp2	islamoph_SLDp2_false	(1-islamoph_PiSLD)*p2
islamoph_SLDp2	islamoph_SLDp2_true	(1-islamoph_PiSLD)*(1-p2)
islamoph_CW	islamoph_CW_bothtrueorfalse	islamoph_PiCW*p1
islamoph_CW	islamoph_CW_onetrue	islamoph_PiCW*(1-p1)
islamoph_CW	islamoph_CW_onetrue	(1-islamoph_PiCW)*p1
islamoph_CW	islamoph_CW_bothtrueorfalse	(1-islamoph_PiCW)*(1-p1)
islamoph_DQ	islamoph_DQ_true	islamoph_PiDQ
islamoph_DQ	islamoph_DQ_false	(1-islamoph_PiDQ)
surname_SLDp1	surname_SLDp1_true	surname_PiSLD*surname_t
surname_SLDp1	surname_SLDp1_false	surname_PiSLD*(1-surname_t)
surname_SLDp1	surname_SLDp1_false	(1-surname_PiSLD)*p1
surname_SLDp1	surname_SLDp1_true	(1-surname_PiSLD)*(1-p1)
surname_SLDp2	surname_SLDp2_true	surname_PiSLD*surname_t
surname_SLDp2	surname_SLDp2_false	surname_PiSLD*(1-surname_t)
surname_SLDp2	surname_SLDp2_false	(1-surname_PiSLD)*p2
surname_SLDp2	surname_SLDp2_true	(1-surname_PiSLD)*(1-p2)
surname_CW	surname_CW_bothtrueorfalse	surname_PiCW*p1
surname_CW	surname_CW_onetrue	surname_PiCW*(1-p1)
surname_CW	surname_CW_onetrue	(1-surname_PiCW)*p1
surname_CW	surname_CW_bothtrueorfalse	(1-surname_PiCW)*(1-p1)
surname_DQ	surname_DQ_true	surname_PiDQ
surname_DQ	surname_DQ_false	(1-surname_PiDQ)

Appendix B

Empirically observed frequencies of answers used in parameter estimation in multiTree (Moshagen, 2010):

xenoph_SLDp1_true	178
xenoph_SLDp1_false	40
xenoph_SLDp2_true	106
xenoph_SLDp2_false	107
xenoph_CW_bothtrueorfalse	224
xenoph_CW_onetrue	216
xenoph_DQ_true	119
xenoph_DQ_false	322
islamoph_SLDp1_true	157
islamoph_SLDp1_false	62
islamoph_SLDp2_true	123
islamoph_SLDp2_false	97
islamoph_CW_bothtrueorfalse	218
islamoph_CW_onetrue	228
islamoph_DQ_true	185
islamoph_DQ_false	242
surname_SLDp1_true	177
surname_SLDp1_false	43
surname_SLDp2_true	122
surname_SLDp2_false	100

surname_CW_bothtrueorfalse	223
surname_CW_onetrue	203
surname_DQ_true	182
surname_DQ_false	262

On the comprehensibility and perceived privacy protection of indirect questioning techniques

Adrian Hoffmann*[1], Alexander F. Schmidt*[2], Berenike Waubert de Puiseau*[1], Jochen Musch*[1]

[1]University of Duesseldorf, [2]University of Luxembourg

«fn»*A. Hoffmann, A. F. Schmidt, B. Waubert de Puiseau and J. Musch contributed equally to this work.

Author Note

Adrian Hoffmann, Berenike Waubert de Puiseau & Jochen Musch, Department of Experimental Psychology, University of Duesseldorf.

Alexander F. Schmidt, Institute for Health and Behavior, Integrative Research Unit on Social and Individual Development, University of Luxembourg.

Correspondence concerning this article should be addressed to Adrian Hoffmann, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, 40225 Duesseldorf, Germany. E-mail: adrian.hoffmann@hhu.de

Abstract

On surveys that assess sensitive personal attributes, indirect questioning increases respondents' willingness to answer truthfully by protecting confidentiality. However, the assumption that subjects understand questioning procedures fully and trust them to protect their privacy is tested rarely. We compared four indirect questioning procedures in terms of comprehensibility and perceived privacy protection. All indirect questioning techniques were found less comprehensible for respondents than a conventional direct question used for comparison. Less-educated respondents experienced more difficulties when confronted with any indirect questioning technique. Regardless of education, the Crosswise Model was found most comprehensible among the four indirect methods. Indirect questioning was perceived to increase privacy protection in comparison to a direct question. Unexpectedly, comprehension and perceived privacy protection did not correlate. We recommend assessing these factors separately in future evaluations of indirect questioning.

Keywords: confidentiality, comprehension, randomized response technique, stochastic lie detector, crosswise model

On the comprehensibility and perceived privacy protection of indirect questioning techniques

When queried about sensitive personal attributes, some respondents conceal their true statuses by responding untruthfully to present themselves in a socially desirable manner (Krumpal, 2013; Marquis, Marquis, & Polich, 1986; Paulhus, 1991; Paulhus & Reid, 1991; Phillips & Clancy, 1972; Rasinski, Willis, Baldwin, Yeh, & Lee, 1999; Sudman & Bradburn, 1974; Tourangeau & Yan, 2007). To increase respondents' willingness to respond honestly, indirect questioning procedures such as the randomized response technique (Warner, 1965) enhance the confidentiality of individual answers to sensitive questions. Consequently, prevalence estimates for sensitive personal attributes obtained through indirect questioning are considered more valid than prevalence estimates based on conventional direct questioning. However, use of indirect questioning relies on the assumption that participants understand all instructions, and understand how the procedures increase privacy protection (Landsheer, van der Heijden, & van Gils, 1999). Violation of this assumption is potentially at odds with a method's acceptance and validity of results. Employing a quasi-experimental design, this study investigates the influence of questioning techniques and education on comprehension and perceived privacy protection. Four indirect questioning techniques were investigated, and a conventional direct question served as a control condition.

Indirect Questioning Techniques

To minimize bias due to respondents not answering truthfully to a sensitive question, Warner (1965) introduced the randomized response technique (RRT). With the original RRT procedure, respondents are confronted simultaneously with two related questions: a sensitive question A ("Do you carry the sensitive attribute?") and its negation question B ("Do you not carry the sensitive attribute?"). Participants answer one of these two questions, depending on the

outcome of a randomization procedure, which is known only to the respondent and not the experimenter. When using a die as a randomization device, for example, respondents might be asked to answer question A if the die shows a number between 1 and 4 (randomization probability $p = 4/6$), and to answer question B if the die shows either 5 or 6 ($p = 2/6$). Using this procedure, a "Yes" response does not allow conclusions regarding a respondent's true status. He or she might be a carrier of the sensitive attribute who was instructed to respond to statement A, or a non-carrier instructed to respond to B. Since the randomization probability p is known, the proportion of carriers of the sensitive attribute π can be estimated at the sample level (Warner, 1965). Since the collection of individual data related directly to the sensitive attribute is avoided, respondents queried about sensitive topics are expected to answer more truthfully when asked indirectly, rather than through direct questioning (DQ). Prevalence estimates obtained via RRT are supposed to exceed DQ estimates, and this has been found repeatedly (e.g., Chi, Chow, & Rider, 1972; Goodstadt & Gruson, 1975; Scheers & Dayton, 1987; Simon, Striegel, Aust, Dietz, & Ulrich, 2006). However, non-significantly different estimates in RRT and DQ conditions, and estimates higher in the DQ than in the RRT condition, have been reported (e.g., Akers, Massey, Clarke, & Lauer, 1983; Holbrook & Krosnick, 2010a; Kulka, Weeks, & Folsom, 1981; Wolter & Preisendörfer, 2013). Many validation studies considered higher estimates to be evidence of higher validity due to better control of social desirability (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005; Umesh & Peterson, 1991), whereas lower estimates might be the result of sampling error occurring when sample sizes are insufficient to compensate for the low efficiency of RRT designs. Randomized response estimates are always accompanied by a higher standard error than direct questions since employing randomization adds unsystematic variance to the estimator (Ulrich, Schröter, Striegel, & Simon, 2012).

Following the original model from Warner (1965), various, more advanced RRT models have been proposed that focus on optimizing the statistical efficiency, validity, and applicability of the method (e.g., Dawes & Moore, 1980; Horvitz, Shah, & Simmons, 1967; Mangat & Singh, 1990; Moors, 1971). Several reviews and monographs provide detailed descriptions of RRT models and their applications (e.g., Antonak & Livneh, 1995; Chaudhuri & Christofides, 2013; Chaudhuri & Mukerjee, 1988; Fox & Tracy, 1980; Horvitz, Greenberg, & Abernathy, 1976; Tracy & Mangat, 1996; Umesh & Peterson, 1991). We present four indirect questioning procedures used in studies that investigate the prevalence of sensitive, personal attributes, and compare them in terms of comprehensibility and perceived privacy protection.

The Cheating Detection Model

With the cheating detection model (CDM; Clark & Desharnais, 1998), participants are confronted with a forced-response paradigm. After presentation of a single, sensitive question, the outcome of a randomization procedure determines whether respondents answer truthfully to this question with probability p or ignore the question and answer "Yes" with probability 1-p. Since the outcome of the randomization procedure remains confidential, a "Yes" response does not allow for conclusion concerning an individual's status with respect to a sensitive attribute. Clark and Desharnais (1998) suspect some participants disobey instructions by responding "No" regardless of the outcome of randomization, to avoid risk of being marked as a carrier of a sensitive attribute. Consequently, three disjoint and exhaustive classes are considered with CDM: carriers of the sensitive attribute responding truthfully (π), honest non-carriers (β), and respondents concealing their true statuses by answering "No" without regard for instructions. Clark and Desharnais refer to the latter class as cheaters (γ). An example of a CDM question using a respondent's month of birth as a randomization device is shown in Figure 1.

TAKE IN FIGURE 1

The CDM has been shown repeatedly to produce higher, and thus presumably more valid, prevalence estimates than direct questions or other indirect questioning techniques that do not consider instruction disobedience. Validation studies arrive frequently at estimates of γ that exceed zero substantially, demonstrating the usefulness of a cheating-detection approach (e.g., Moshagen, Musch, Ostapczuk, & Zhao, 2010; Ostapczuk & Musch, 2011; Ostapczuk, Musch, & Moshagen, 2009, 2011). However, in the case of $\gamma > 0$, the CDM provides only a lower bound for the proportion of carriers since the true statuses of respondents classified as cheaters are unknown. Hence, the rate of carriers could be located within the range of π (were no cheater a carrier) and $\pi + \gamma$ (were all cheaters carriers).

The Stochastic Lie Detector

Similar to the original RRT procedure (Warner, 1965), the recently proposed stochastic lie detector (SLD; Moshagen, Musch, & Erdfelder, 2012) confronts respondents with sensitive question A and its negation B. Similar to the modified RRT model that Mangat (1994) proposes, only part of the participants is instructed to engage in randomization. Carriers of the sensitive attribute respond to question A unconditionally, and if they respond truthfully, their answer should always be "Yes". Non-carriers respond to question A with randomization probability *p*, and to question B with probability 1-*p*. Consequently, neither a "Yes" nor "No" response unequivocally reveals a respondent's true status. However, Moshagen et al. (2012) argues that some carriers of the sensitive attribute might feel a desire to lie and respond "No", even if instructed otherwise. In contrast, non-carriers should not have any reason to lie. These assumptions were represented by a new parameter *t*, which accounts for the proportion of carriers of the sensitive attribute answering truthfully. The remaining proportion, 1-*t*, of the

carriers are assumed to lie about their true statuses. An example of an SLD question with the respondent's month of birth as a randomization device is shown in Figure 2.

TAKE IN FIGURE 2

During a pilot study, application of the SLD resulted in a prevalence estimate for domestic violence that exceeded an estimate obtained using a direct question. During a second validation study, an SLD estimate for the proportion of nonvoters in the German federal elections in 2009 nearly concurred with the known true prevalence of this attribute (Moshagen et al., 2012). Even more compelling were results of a third study from Moshagen, Hilbig, Erdfelder, and Moritz (2014), in which cheating behaviors were induced experimentally among respondents to allow direct determination of the proportion of cheaters as an external criterion against which the SLD could be validated. Again, SLD nearly reproduced the now known proportion of carriers of the sensitive attribute, while the DQ condition produced an underestimate. In contrast to these results, a recent experimental comparison of SLD with the competing crosswise model (Yu, Tian, & Tang, 2008) and DQ conditions found SLD to overestimate the known prevalence of a non-sensitive control question (Hoffmann & Musch, 2014). Although this mixed pattern of results might be explained in terms of sampling error, difficulties regarding understanding SLD instructions offer an explanation.

The Crosswise Model

A new class of non-randomized response techniques was proposed recently (Tian & Tang, 2013), offering simplified assessment of the prevalence of sensitive attributes since no external randomization device is required. One of the most promising candidates among these is the crosswise model (CWM; Yu et al., 2008) because it offers symmetric answer categories (i.e., none of the answer options is a safe alternative that eliminates identification as a carrier). With

CWM, participants are presented with two statements simultaneously: one statement refers to the sensitive attribute with unknown prevalence π, and a second to a non-sensitive control attribute with known prevalence *p* (e.g., a respondent's month of birth). Participants indicate whether "both statements are true or both statements are false", or whether "exactly one of the two statements is true (irrespective of which one)". If a respondent's month of birth is unknown to the questioner, CWM presumably provides an undistorted estimate of a sensitive attribute's prevalence, without forcing respondents to reveal anything about their true statuses. Figure 3 shows an example of a CWM question, with a respondent's month of birth as a randomization device.

TAKE IN FIGURE 3

In various studies, application of CWM resulted in higher prevalence estimates for sensitive attributes than DQ approaches did (e.g., Coutts, Jann, Krumpal, & Näher, 2011; Jann, Jerke, & Krumpal, 2012; Kundt, Misch, & Nerré, 2013; Nakhaee, Pakravan, & Nakhaee, 2013). An experimental comparison of CWM, SLD, and a DQ condition showed that CWM and SLD prevalence estimates of xenophobia and Islamophobia exceeded those obtained in a DQ condition. The CWM estimated the known prevalence of a non-sensitive control attribute accurately (Hoffmann & Musch, 2014). Yu et al. (2008) argue that non-randomized models are "easy to operate for both interviewer and interviewee" (p. 261), which offers an explanation for promising results observed to date using the CWM.

The Unmatched Count Technique

Introduced by Miller (1984), the unmatched count technique (UCT; sometimes called the item count technique or the randomized list technique) also offers comparably simple instructions to respondents, who are assigned randomly to one of two groups both of which are

confronted with a list of statements. In the first (experimental) group, the list contains a number of non-sensitive statements and one statement containing a sensitive attribute. In the second (control) group, only the non-sensitive statements are presented. In both groups, respondents indicate how many, but not which, of the statements apply to them. Since the only disparity between the two groups is the addition of a question referring to the sensitive attribute in the experimental group, a difference in mean reported total counts estimates the proportion π of carriers of the sensitive attribute (Erdfelder & Musch, 2006; Miller, 1984). The individual statuses of respondents in the experimental group confronted with a list containing the sensitive attribute remain confidential as long as the total reported count is different from zero (in which case all statements, including the sensitive statement, could be deduced to have been answered negatively), and different from the maximum count possible (in which case all statements, including the sensitive statement, could be deduced to have been answered affirmatively). Thus, experimenters should prevent such extreme counts cautiously by including a sufficient number of non-sensitive statements (Erdfelder & Musch, 2006; Fox & Tracy, 1986). An example of a UCT question with one sensitive and three non-sensitive items is shown in Figure 4.

TAKE IN FIGURE 4

UCT has repeatedly provided higher prevalence estimates for sensitive attributes than DQ approaches did (e.g., Ahart & Sackett, 2004; Coutts & Jann, 2011; Dalton, Wimbush, & Daily, 1994; Holbrook & Krosnick, 2010b; LaBrie & Earleywine, 2000; Wimbush & Dalton, 1997). Comprehensibility of the instructions and trust in the method were found to exceed that of the RRT and a conventional DQ approach (Coutts & Jann, 2011). These results however were limited to a comparison of UCT and a forced-response RRT design, and comprehension was evaluated only by means of potentially forgeable self-ratings.

A promising new approach to detect noncompliance has recently been proposed by Nepusz, Petroczi, Naughton, Epton, and Norman (2014). Their SSC-MLE model is based on a maximum-likelihood extension of the single sample count method and allows to detect both, guilty noncompliance and innocent noncompliers. Moreover, the authors of the model derived and tested several hypotheses regarding the nature of different kinds of noncompliance. However, the SSC-MLE model has more degrees of freedom than competing item count models, and has to be based on four innocuous questions with a probability of .5 each. For this reason, the SSC-MLE model was not included in the present analysis.

A meta-analytic evaluation of indirect questioning studies (Lensvelt-Mulders et al., 2005) reveals that prevalence estimates obtained through RRT largely meet the more-is-better criterion (i.e., RRT prevalence estimates for socially undesirable attributes exceed estimates based on direct questions). Higher estimates indicate increased validity since social desirability biases them less. Another, more selective meta-analytic accumulation of strong validation studies in which the true prevalence of a sensitive attribute was known and could be used as an objective standard found that RRT yields prevalence estimates that are substantially less biased than direct questions (Lensvelt-Mulders et al., 2005). Some studies present RRT estimates that are indifferent from (e.g., Kulka et al., 1981) or even lower than (e.g., Holbrook & Krosnick, 2010a) DQ estimates. Regarding thorough examination of the validity of indirect questioning, in some strong validation studies, RRT estimates deviated substantially from known population values (e.g., Kulka et al., 1981; van der Heijden, van Gils, Bouts, & Hox, 2000). These results might be explained in terms of participants' noncompliance with instructions even under RRT conditions, especially concerning surveys that cover highly sensitive personal attributes (e.g., Clark & Desharnais, 1998; Edgell, Himmelfarb, & Duchan, 1982; Moshagen et al., 2012). Two

psychological aspects that are also likely to play a role in respondents' willingness to cooperate are a) the ability to understand instructions and b) whether respondents trust the promise of confidentiality associated with use of indirect questioning.

Comprehensibility and perceived privacy protection from indirect questioning

Most indirect questioning relies on the assumption that participants comply with instructions; they are capable and willing to cooperate (e.g., Abul-Ela, Greenberg, & Horvitz, 1967; Edgell et al., 1982; Franklin, 1998; Warner, 1965). Many researchers raise concerns that some participants might not understand instructions for indirect questions fully since they are generally more complex in comparison to direct questioning formats (e.g., Coutts & Jann, 2011; Landsheer et al., 1999; O'Brien, 1977). Participants might also not trust indirect questioning to protect their privacy, and might therefore disregard instructions (e.g., Clark & Desharnais, 1998; Landsheer et al., 1999; Moshagen et al., 2012). Response bias resulting from lack of understanding or trust toward a method threatens the validity of prevalence estimates determined through indirect questions (Holbrook & Krosnick, 2010a; James, Nepusz, Naughton, & Petroczi, 2013; Landsheer et al., 1999; Umesh & Peterson, 1991). Hence, trust and understanding are two psychological factors that determine the validity of indirect questioning (Fox & Tracy, 1980; Landsheer et al., 1999; Zdep & Rhodes, 1977).

One strategy used to evaluate the comprehensibility and perceived privacy protection is assessment of response rates in surveys that use indirect questioning. Following the logic of these studies, higher response rates indicate higher trust and understanding. In an early pilot study, Boruch (1972) reports that two RRT conditions showed a "slight (insignificant) decrement in the likelihood of response" (p. 411) when compared to a conventional DQ condition. A similar study observed full cooperation in a DQ condition, but slightly reduced response rates for

respondents confronted with an RRT question (Coutts & Jann, 2011). Other studies report comparable response rates for indirect and direct questioning (e.g., Chi et al., 1972; Locander, Sudman, & Bradburn, 1976), or higher response rates during indirect questioning (e.g., Fidler & Kleinknecht, 1977; Goodstadt & Gruson, 1975; Zdep & Rhodes, 1977). These results only allow indirect conclusions regarding the comprehensibility and perceived privacy protection of the questioning techniques used since there exist numerous explanations for disparities in response rates (e.g., motivational factors and the content of sensitive questions). Differential influences of trust and understanding cannot be disentangled based on analysis of response rates.

Using more controlled approaches, some validation studies use known individual statuses of respondents regarding sensitive attributes to determine whether they responded in accordance with instructions. The rate of demonstrably untrue responses was used to estimate the rate of participants who did not understand or trust the questioning procedure. Edgell et al. (1982) argue that a low rate of only 4% incorrect responses to a moderately sensitive question indicates a high level of comprehension. However, the rate of false answers rose to 26% for a highly sensitive question. In a similar study, Edgell, Duchan, and Himmelfarb (1992) show that between 2% and 10% of respondents give incorrect answers to increasingly sensitive questions. It is plausible that this stronger bias might in part be caused by respondents distorting answers to increasingly distance themselves from more sensitive attributes (Edgell et al., 1982). A meta-analytic investigation of strong validation studies in which participants' true statuses concerning a sensitive attribute was known identified a mean rate of 38% incorrect responses for RRT questions, while other questioning formats produced up to 49% false answers (Lensvelt-Mulders et al., 2005). Disparities between RRT and DQ estimates increased for questions with higher sensitivity. This pattern could be interpreted as evidence that respondents trust the confidentiality

offered by indirect questioning but require enhanced privacy protection, and use it only if a sensitive issue is at stake. However, designs used in these studies did not separate the influences of comprehension and perceived privacy protection.

A more direct strategy to determine trust and understanding for varying questioning procedures is to assess these two constructs directly on a survey. In a study of the comprehensibility of indirect questioning, 94% of respondents claimed they were able to understand instructions of the RRT question used (Chi et al., 1972). In another study, trained interviewers estimated the rate of interviewee understanding of the RRT procedure at 78% to 90% (Locander et al., 1976). A similar study reports that 79% to 83% of respondents experience no difficulties when responding to a RRT questionnaire (van der Heijden, van Gils, Bouts, & Hox, 1998). Coutts and Jann (2011) report rates of 80% to 93% of respondents who understand fully RRT questions employing randomization devices. For a question using the unmatched count technique (Miller, 1984), the rate is 92%. In another study, the comprehensibility of an RRT question was rated as normal or easy by 89% of respondents, and 10% indicated it was difficult (Hejri, Zendehdel, Asghari, Fotouhi, & Rashidian, 2013, p. 148).

To estimate trust toward an RRT question, some researchers asked participants whether they thought there was a trick to the RRT procedure. Since 20% to 40% (Abernathy, Greenberg, & Horvitz, 1970) and 15% to 37% (Chi et al., 1972) of respondents answered affirmatively to this statement, a considerable fraction of respondents appear to mistrust RRT despite a promise of confidentiality. When questioned indirectly, respondents estimated the probability of "the researcher [knowing] which questions [they] answered" at 55% to 72% (Soeken & Macready, 1982, p. 488), depending on the choice of the randomization probability. Consequently, participants estimated the probability of the procedure protecting their privacy at only 28 to 45%.

Few respondents (15% to 22%) believed RRT "guaranteed the anonymity of their answers" in a study from Coutts and Jann (2011, p. 179); for an unmatched count technique question, the rate was slightly higher though low at 29%.

Aside from assessment of total rates of trust and understanding, some studies compare perceived privacy protection of direct versus indirect questions. In one study, 91% of respondents "felt that the use of th[e] RRT would enhance the confidentiality of their responses" when compared to conventional direct questioning (Edgell et al., 1982, p. 97). In another, a rate of 72% of respondents trusting the RRT procedure was unexpectedly exceeded by a rate of 83% of trustful participants in a DQ condition (van der Heijden et al., 1998), implying RRT failed to establish higher trust. Only 29% of participants in a study from Hejri et al. (2013) perceived that the "RRT increased confidentiality" when compared to a direct question (p. 148). Other studies compare perceived privacy protection offered by various indirect questioning techniques, and evaluate the unmatched count technique as superior to RRT regarding trust and understanding (Coutts & Jann, 2011; James et al., 2013).

Few studies examine the influence of cognitive skill and education on comprehension and perceived privacy protection of indirect questioning designs. Judging from results of an early field trial, Abul-Ela et al. (1967) argue that RRT works only "on more intelligent respondents" (p. 990). Chi et al. (1972) also found a positive effect of education on rate of cooperative respondents. Although the rate of participants failing to understand RRT in a group with no formal education was estimated at 72%, the rate dropped to 27% for participants who graduated from primary school and to 2% for participants who held a junior high school degree. Landsheer et al. (1999) found no influence of participants' formal education on incidences of incorrect answers. Holbrook and Krosnick (2010a) report that the most implausible results in their study

occurred in a subgroup of highly educated participants, indicating that the "failure of the RRT was not due to the cognitive difficulty of the task" (p. 336).

Overall, results from studies that investigate participants' trust in and understanding of indirect questioning are inconclusive. Some studies report high rates of trust and understanding and others show that a substantial share of participants fail to understand indirect questions, or do not trust the procedures. Data do not allow separation of these factors, and thus independent assessment of trust and understanding is needed to identify indirect questioning techniques that are both comprehensible and trustful. The role of cognitive skill and education as moderators of trust and understanding is not yet understood.

Present Study

Various indirect questioning procedures promise to provide less biased and thus more valid prevalence estimates than conventional DQs do. It is however important that participants understand the complex instructions and trust protection of their privacy if the application of indirect measures is to be successful. In this study, four indirect questioning techniques used frequently in survey research that addresses sensitive questions were entered into an experimental comparison of comprehensibility and perceived privacy protection. The CDM (Clark & Desharnais, 1998) accounts for noncompliant participants by implementing an additional cheating parameter, but the true statuses of cheating participants regarding sensitive attributes in questions remain unknown, resulting in rough estimates of the proportion of carriers. With the SLD (Moshagen et al., 2012), an additional cheating parameter allows for separate estimation of the proportion of carriers responding untruthfully, and non-carriers are expected to obey instructions. However, a recent validation study found that SLD overestimates the prevalence of a non-sensitive control attribute, possibly resulting from some participants

experiencing difficulties following the procedure (Hoffmann & Musch, 2014). The CWM (Yu et al., 2008) is presumably easier to understand than other RRT models due to its simplified instructions; participants are confronted with a symmetric design without a safe answering category, which might facilitate honest responding. The UCT (Miller, 1984) is similarly easy to employ, and some participants prefer UCT over RRT questions concerning trust and understanding (Coutts & Jann, 2011). This study evaluates the comprehensibility and perceived privacy protection of these four indirect questioning techniques separately since these two factors might be intertwined though not linked causally in a unidirectional connection. Some participants might understand the instructions but not trust the protection of their privacy. Others might fail to comprehend the task but perceive that indirect questions offer more confidentiality than conventional direct questioning approaches do.

To allow an objective and rigorous evaluation of participants' instruction comprehension, we used a scenario-based design. To assess whether they understood the procedure, participants responded to a number of questions vicariously for various fictional characters. Participants were first given information regarding these characters (e.g., "Wilhelm has never cheated on an exam" or "Wilhelm was born on July") and were subsequently provided with instructions for one of the indirect questioning techniques, and finally indicated which answer the fictional character must give. This approach ensured participants would not respond untruthfully to conceal personal statuses regarding sensitive attributes. The true status for each fictional character was known to both the respondent and questioner, and thus served as an objective criterion for assessment of the correctness of a respondent's answers. The mean proportion of questions answered correctly in a test that assessed a respondent's understanding of the procedure was determined as an estimate of the comprehensibility of each questioning procedure. We also assessed how

participants estimated the privacy protection offered by various questioning techniques. Finally, by questioning two groups of participants with high versus low educations, we investigated moderation of cognitive skill.

This study addresses the following research questions: 1) Do indirect questions differ from conventional direct questions regarding comprehensibility? If so, which one of the four models under investigation is most comprehensible? 2) Do indirect questions offer higher perceived privacy protection than direct questions do? If so, what model is perceived as most protective? 3) Do cognitive skills, measured by respondents' education, moderate the influence of questioning technique on comprehension or perceived privacy protection? 4) Is there an association between comprehension and perceived privacy protection?

Methods

Participants

Seven-hundred sixty-six participants were recruited to participate in an online survey through a commercial online panel. Since education was part of the experimental design, an online quota ensured matching proportions of participants with lower versus higher educations. Participants in the lower-education group finished at most nine years of school (the German *Hauptschule*), and participants in the higher-education group finished at least twelve years of education (the German *Abitur*). To increase homogeneity in the sample, only respondents between 25 and 35 years of age were allowed to participate. Exclusion of participants who failed to complete the questionnaire resulted in a sample of 401 respondents, with a mean age of 30.72 years (SD = 3.35), including 211 (53%) females and 386 (97%) individuals with German as their first language. Education groups were represented evenly, with 199 lower- and 202 higher-education participants. Power analyses conducted using G*Power 3 software (Faul, Erdfelder,

Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) revealed that the sample size provided sufficient power for detection of medium effects during analysis of mean differences between groups ($f = 0.25$; $1\text{-}\beta = .99$) and correlations ($r = .30$; $1\text{-}\beta > .99$).

Design

The scenario-based experiment implemented a 5 by 2, quasi-experimental mixed design. Questioning technique varied within subjects, realized in five blocks: CDM (Clark & Desharnais, 1998), SLD (Moshagen et al., 2012), CWM (Yu et al., 2008), UCT (Miller, 1984) and a conventional DQ approach. Indirect questioning techniques were implemented as shown in Figures 1 through 4. Academic cheating served as the sensitive attribute, as used in several studies of indirect questioning techniques (e.g., Hejri et al., 2013; Lamb & Stem, 1978; Ostapczuk, Moshagen, Zhao, & Musch, 2009; Scheers & Dayton, 1987). The wording of the sensitive question was identical in all questioning technique conditions, reading "Have you ever cheated on an exam?" Three additional, non-sensitive attributes were used to employ indirect questioning techniques. First, month of birth was used as the randomization device for the CDM, SLD, and CWM questions. To allow application of the UCT format, we constructed a list of four items: the sensitive attribute, the non-sensitive month of birth, and two non-sensitive attributes (i.e., gender and a question concerning whether participants visited London). Each of the questioning techniques was applied to four fictional characters named Ludwig, Ernst, Hans, and Wilhelm, characterized differently regarding the sensitive and non-sensitive attributes. Ludwig and Ernst were presented as carriers of the sensitive attribute, and Hans and Wilhelm were described as non-carriers. The birthdays of Ludwig and Hans were chosen to fall into one of the outcome categories of the binary randomization procedure, and the months of birth for Ernst and Wilhelm were set to fall into the other category. All four characters were male, and none was

described to have visited London. The descriptions were chosen to avoid extreme counts in the UCT condition. Descriptions of the four fictional characters were accessible to participants at any time during the experiment. To control for effects of serial position, the sequence of presentation of the five questioning technique blocks was randomized among participants. Additionally, the four fictional characters were presented in random order within each of the questioning technique blocks. A second, quasi-experimental, between-subjects independent variable was the participants' education (high versus low), described above.

To examine the comprehensibility of the questioning techniques, participants vicariously indicated answers that the four fictional characters must give if confronted with each of the various questioning techniques. Descriptions of the characters were displayed along with the questions. As an example, a screenshot of a CWM question that had to be answered from the perspective of Wilhelm is shown in figure 5. The comprehensibility of the questioning techniques was operationalized as the mean proportion of correct answers concerning all four fictional characters.

TAKE IN FIGURE 5

To assess perceived privacy protection, participants rated perceived confidentiality offered by each questioning technique on a 7-point, Likert-type scale, ranging from -3 (no confidentiality) to 3 (perfect confidentiality). Scales were presented directly below the comprehension questions. Perceived privacy protection was operationalized as the mean score on these Likert-scales concerning all four fictional characters.

Results

Comprehensibility

Mean proportions and standard errors of correct responses as a function of questioning technique and education are shown in Table 1. Reliability analyses for the proportion of correct responses across all five questioning techniques revealed that the variable measured a homogenous construct (Cronbach's $\alpha = .75$). A univariate, 5 (questioning technique) by 2 (education), mixed-model ANOVA revealed a main effect for within-subjects questioning technique ($F(4,1596) = 75.46$, $p < .001$, $\eta^2 = .16$), a main effect for between-subjects education ($F(1,399) = 17.07$, $p < .001$, $\eta^2 = .04$), and an interaction of these two factors ($F(4,1596) = 4.13$, $p < .001$, $\eta^2 = .01$). A Bonferroni *post-hoc* test for within-subjects questioning technique showed that the mean proportion of correct answers in the DQ control condition was higher than with CDM ($\Delta M = 15.04\%$, $p < .001$; $r = .44$, $dz = 0.70$; according to Cohen, 1988), SLD ($\Delta M = 21.73\%$, $p < .001$; $r = .23$, $dz = 0.79$), CWM ($\Delta M = 7.07\%$, $p < .001$; $r = .49$, $dz = 0.33$), and UCT ($\Delta M = 13.38\%$, $p < .001$; $r = .52$, $dz = 0.49$) condition. Pairwise comparisons among indirect questioning techniques resulted in differences for all combinations (all $p < .001$; CDM versus SLD: $r = .38$, $dz = 0.26$; CDM versus CWM: $r = .39$, $dz = 0.33$; SLD versus CWM: $r = .29$, $dz = 0.52$; SLD versus UCT: $r = .25$, $dz = 0.24$; CWM versus UCT: $r = .52$, $dz = 0.23$), except for no difference between CDM and UCT conditions ($p > .99$; $r = .42$, $dz = 0.06$). Thus, participants demonstrated highest comprehension for direct questions, slightly though reduced comprehension for CWM questions, further reduced comprehension for CDM and UCT questions, and lowest comprehension for SLD questions. As hypothesized, higher education resulted in higher mean percentages of correct responses. To explore the interaction of questioning technique and education further, five pairwise t-tests for independent groups on a

Bonferroni-corrected α-level (corrected α = .05/5 = .01) were computed that compared participants with high versus low education separately within each questioning technique condition. The comparisons revealed no effect of education in the DQ condition (ΔM = 1.39%, $t(399) = -0.71$, $p = .48$; $d = 0.07$). Within the CDM condition, people with lower education had slightly lower scores, but the difference remained non-significant on the corrected α-level (ΔM = 4.98%, $t(399) = -2.37$, $p = .02$; $d = 0.24$). For SLD (ΔM = 9.70%, $t(399) = -4.07$, $p < .001$; d = 0.41), CWM (ΔM = 7.61%, $t(399) = -3.39$, $p < .001$; $d = 0.34$), and UCT (ΔM = 11.07%, $t(399)$ = -3.56, $p < .001$; $d = 0.36$) conditions, lower education resulted in lower scores. Hence, although comprehension was comparable between educational groups for a direct question, education moderated comprehension in three of four indirect questioning formats.

TAKE IN TABLE 1

Perceived privacy protection

Mean ratings and standard errors of perceived privacy protection as a function of questioning technique and education are shown in Table 2. Reliability analyses for mean ratings of perceived privacy protection across all five questioning techniques revealed that the variable measured a homogenous construct ($\alpha = .87$). A univariate 5 (questioning technique) by 2 (education), mixed-model ANOVA revealed a main effect for within-subjects questioning technique ($F(4,1596) = 18.76$, $p < .001$, $\eta^2 = .05$), but no effect for between-subjects education ($F(1,399) < 1$). However, the two factors showed an interaction ($F(4,1596) = 9.21$, $p < .001$, $\eta^2 = .02$). A Bonferroni *post-hoc* test of the factor questioning technique revealed that mean scores in the DQ control condition were lower than with CDM ($\Delta M = 0.26$, $p < .001$; $r = .57$, $dz = 0.19$), SLD ($\Delta M = 0.25$, $p < .01$; $r = .53$, $dz = 0.18$), CWM ($\Delta M = 0.39$, $p < .001$; $r = .39$, $dz = 0.25$), and UCT ($\Delta M = 0.52$, $p < .001$; $r = .40$, $dz = 0.33$) conditions. *Post-hoc* tests between the indirect

questioning techniques showed that the UCT format resulted in the highest scores, indifferent from scores in the CWM condition ($\Delta M = 0.13$, $p = .21$; $r = .64$, $dz = 0.12$) but higher than scores with CDM ($\Delta M = 0.26$, $p < .001$; $r = .61$, $dz = 0.22$) and SLD ($\Delta M = 0.27$, $p < .001$; $r = .64$, $dz = 0.24$) conditions. Mean scores in the CWM condition were comparable to scores in the CDM ($\Delta M = 13$, $p = .31$; $r = .61$, $dz = 0.11$) and SLD ($\Delta M = 0.14$, $p = .10$; $r = .67$, $dz = 0.13$) conditions. Finally, CDM and SLD scores showed no difference ($\Delta M = 0.01$, $p > .99$; $r = .65$, $dz = 0.01$). Combined, all indirect questioning techniques enhanced perceived privacy protection in comparison with a conventional DQ. Participants perceived the highest privacy protection when confronted with UCT and CWM questions, and perceived privacy ratings for CWM, CDM, and SLD questions did not differ. Since no main effect of education emerged, results are only presented for the interaction of education and questioning technique. Five pairwise t-tests for independent groups on a Bonferroni-corrected α-level (*corrected* $\alpha = .05 / 5 = .01$) were computed to compare participants with high versus low education separately within each questioning technique condition. The comparisons revealed an education effect only in the DQ condition ($\Delta M = 0.51$, $t(399) = 3.35$, $p < .001$; $d = 0.33$), while education groups did not differ on the corrected α within CDM ($\Delta M = 0.08$, $t(399) = 0.64$, $p = .53$; $d = 0.07$), SLD ($\Delta M = 0.10$, $t(399) = 0.78$, $p = .43$; $d = 0.08$), CWM ($\Delta M = 0.10$, $t(399) = 0.77$, $p = .44$; $d = 0.07$), and UCT ($\Delta M = 0.26$, $t(399) = 1.98$, $p = .05$; $d = 0.20$) conditions. Hence, participants with lower education perceived higher privacy protection when confronted with a direct question than participants with higher education, and perceived privacy protection did not differ between education groups within indirect questioning conditions.

TAKE IN TABLE 2

Association of comprehension and perceived privacy protection

To investigate whether participants' comprehension of a questioning technique was associated with perceived privacy protection, bivariate Pearson-correlations were computed for the total sample, and separately for the two education groups (Table 3). Comprehension and perceived privacy protection showed no associations, with the exception of one positive correlation in the SLD condition for participants with low education. Hence, comprehension and perceived privacy protection were not associated.

TAKE IN TABLE 3

Discussion

In the present study, we compared four indirect questioning procedures in terms of comprehensibility and perceived privacy protection. A conventional direct question served as a control condition. Moderating effects of participants' level of education were investigated.

Comprehensibility of indirect questioning techniques

All indirect questioning techniques showed lower comprehensibility in comparison to a DQ condition. Results accord with extant studies that suggest the instructions of indirect questions are more complex and thus more difficult to comprehend than direct questions (e.g., Böckenholt, Barlas, & van der Heijden, 2009; Coutts & Jann, 2011; Edgell et al., 1992; Landsheer et al., 1999; O'Brien, 1977). In a qualitative interview study, Boeije and Lensvelt-Mulders (2002) report that the reduced comprehensibility of indirect RRT questions might be explained partially by participants experiencing difficulties when "doing two things at the same time" (p. 30). Participants struggle to focus on RRT questions and the randomization procedure simultaneously. This experience applies to the present study since subjects had to integrate two types of information to identify the correct responses in all indirect questioning conditions: first

the status of the fictional characters regarding a sensitive attribute, and second their statuses concerning non-sensitive randomization attribute(s). Results suggest that some indirect questioning formats showed better comprehensibility than others did; CWM appears to have been the most comprehensible format, corroborating Yu et al.'s (2008) assertion that CWM is easier to operate. Integrating two types of information or "doing two things at the same time" (Boeije & Lensvelt-Mulders, 2002, p. 30; also see Lensvelt-Mulders & Boeije, 2007, p. 598) might have been easiest for participants in the CWM condition since this questioning format incorporates the randomization procedure and the response to the sensitive statement in a single step. Respondents must simply read two answer options and identify the appropriate one. In contrast, comprehension was lowest in the SLD condition. A more detailed inspection of the SLD's instructions revealed that participants must make three sequential decisions to identify the correct response: a) decide whether the fictional character is a carrier of the sensitive attribute, b) identify the question that must be answered as determined by the randomization procedure (if the character is a non-carrier), and c) identify the correct response to the respective question. Answering an SLD question appears more difficult, and more prone to errors, than answering a CWM question. However, this explanation is speculative and thus should be tested separately in future studies. A qualitative interview study similar to the one conducted by Boeije and Lensvelt-Mulders (2002) might reveal the exact mechanisms that account for differential comprehensibility of the four indirect questioning models investigated here.

The lower-education group demonstrated decreased comprehension of all indirect questioning techniques, with the exception of CDM. Researchers investigating the prevalence of sensitive personal attributes should consider that the comprehension of indirect questions might be reduced in samples that include less-educated participants, and should refrain from applying

indirect questioning techniques if less-educated individuals report difficulties while completing a survey.

Comprehension rates reported in this study are likely a lower boundary for the comprehensibility of questioning procedures under investigation. Since a within-subjects, scenario-based design was used, participants' comprehension was likely to improve if they had to deal with only one questioning technique, and if not required to respond vicariously about fictional characters but for themselves. However, mean comprehension in the DQ condition was high (> 90%) and unaffected by education, indicating participants were capable of answering questions from the perspective of the four fictional characters. Instructions for all indirect questioning procedures were kept as concise as possible. During real applications, more comprehensive instructions could be presented along with extended explanations, and can be combined with comprehension checks to ensure respondents understand the procedure. In contrast to many extant studies that use face-to-face questioning or paper-pencil tests, this study confronted participants with an online questionnaire that contained indirect questioning techniques. Although RRT yielded valid results in previous online studies (e.g., Musch, Bröder, & Klauer, 2001), a face-to-face setting offers better opportunities to assist participants who experience difficulties, and might help respondents achieve better comprehension and avoid errors when answering questions.

Perceived privacy protection

Regarding perceived privacy protection, all indirect questioning techniques showed higher mean scores than a conventional DQ, suggesting participants developed higher trust toward indirect questions. The highest mean score was achieved in the UCT condition, followed by a slightly but insignificantly reduced mean score with CWM. Scores under CWM, CDM, and

SLD were similar, though the latter two differed from the UCT condition. Education influenced perceived privacy protection only in the DQ condition, with lower-education participants reporting higher perceived protection. This education effect did not occur in any indirect questioning condition. Hence, the influence of education on perceived privacy protection reduces to failure to understand that direct questions provide poorer privacy protection. When sensitive questions are assessed using indirect questioning, education might be negligible concerning perceived protection.

Comprehension did not associate with perceived privacy protection for the entire sample, or in the two education groups. Only in the SLD condition was a small, positive association in the lower-education subsample detected. This pattern suggests that although participants understood the instructions, they did not necessarily trust the procedure. Results also suggest respondents developed trust despite failure to comprehend instructions fully. Lack of association between comprehension and perceived privacy protection suggests the importance of examining differential impacts of these two constructs separately when assessing sensitive topics with indirect questioning techniques. To allow valid assessment of the prevalence of sensitive personal attributes, participants should ideally both understand and trust the questioning technique.

This study supports application of indirect questioning since it increases perceived privacy protection. When selecting among techniques, the best advice is to use CWM (Yu et al., 2008) to assess sensitive personal attributes. This model had the highest comprehensibility among indirect questioning techniques, and substantially increased perceived privacy protection in comparison to direct questioning. This recommendation is supported further by findings from various extant studies that suggest CWM results in more valid prevalence estimates than

conventional direct questioning (e.g., Coutts et al., 2011; Hoffmann & Musch, 2014; Jann et al., 2012; Kundt et al., 2013; Nakhaee et al., 2013). If the attribute under investigation is extraordinarily sensitive (e.g., deviant sexual interests or severe criminal behavior), researchers may want to consider using the UCT (Miller, 1984) to maximize perceived privacy.

References

Abernathy, J. R., Greenberg, B. G., & Horvitz, D. G. (1970). Estimates of Induced Abortion in Urban North-Carolina. *Demography, 7*(1), 19-29.

Abul-Ela, A. L. A., Greenberg, B. G., & Horvitz, D. G. (1967). A Multi-Proportions Randomized Response Model. *Journal of the American Statistical Association, 62*, 990-1008.

Ahart, A. M., & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods, 7*, 101-114. doi: 10.1177/1094428103259557

Akers, R. L., Massey, J., Clarke, W., & Lauer, R. M. (1983). Are Self-Reports of Adolescent Deviance Valid? Biochemical Measures, Randomized-Response, and the Bogus Pipeline in Smoking-Behavior. *Social Forces, 62*, 234-251.

Antonak, R. F., & Livneh, H. (1995). Randomized-Response Technique - a Review and Proposed Extension to Disability Attitude Research. *Genetic, Social, and General Psychology Monographs, 121*, 97-145.

Böckenholt, U., Barlas, S., & van der Heijden, P. G. M. (2009). Do Randomized-Response Designs Eliminate Response Biases? An Empirical Study of Non-Compliance Behavior. *Journal of Applied Econometrics, 24*(3), 377-392. doi: Doi 10.1002/Jae.1052

Boeije, H., & 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.

Boruch, R. F. (1972). Relations among statistical methods for assuring confidentiality of social research data. *Social Science Research, 1*, 403-414.

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

Chaudhuri, A., & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

Chi, I., Chow, L. P., & Rider, R. V. (1972). Randomized Response Technique as Used in Taiwan Outcome of Pregnancy Study. *Studies in Family Planning, 3*, 265-269.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coutts, E., & Jann, B. (2011). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research, 40*, 169-193. doi: 10.1177/0049124110390768

Coutts, E., Jann, B., Krumpal, I., & Näher, A. F. (2011). Plagiarism in Student Papers: Prevalence Estimates Using Special Techniques for Sensitive Questions. *Jahrbücher für Nationalökonomie Und Statistik, 231*, 749-760.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the Unmatched Count Technique (Uct) to Estimate Base Rates for Sensitive Behavior. *Personnel Psychology, 47*, 817-828. doi: 10.1111/j.1744-6570.1994.tb01578.x

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]*. Göttingen: Hogrefe.

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

Erdfelder, E., & Musch, J. (2006). Experimental methods of psychological assessment. In M. Eid & E. Diener (Eds.), *Handbook of Multimethod Measurement in Psychology* (pp. 205-220). Washington, D.C.: American Psychological Association.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fidler, D. S., & Kleinknecht, R. E. (1977). Randomized Response Versus Direct Questioning - 2 Data-Collection Methods for Sensitive Information. *Psychological Bulletin, 84*, 1045-1049.

Fox, J. A., & Tracy, P. E. (1980). The Randomized-Response Approach - Applicability to Criminal-Justice Research and Evaluation. *Evaluation Review, 4*(5), 601-622. doi: Doi 10.1177/0193841x8000400503

Fox, J. A., & Tracy, P. E. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.

Franklin, L. (1998). Randomized Response Techniques. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (Vol. 5, pp. 3696-3703). New York: Wiley.

Goodstadt, M. S., & Gruson, V. (1975). Randomized Response Technique - Test on Drug-Use. *Journal of the American Statistical Association, 70*, 814-818.

Hejri, S. M., Zendehdel, K., Asghari, F., Fotouhi, A., & Rashidian, A. (2013). Academic disintegrity among medical students: a randomised response technique study. *Medical Education, 47*, 144-153. doi: 10.1111/Medu.12085

Hoffmann, A., & Musch, J. (2014). *Assessing the validity of two indirect questioning techniques: a Stochastic Lie Detector versus the Crosswise Model*. Manuscript under preparation.

Holbrook, A. L., & Krosnick, J. A. (2010a). 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

Holbrook, A. L., & Krosnick, J. A. (2010b). Social desirability bias in Voter Turnout Reports: Tests Using the Item Count Technique. *Public Opinion Quarterly, 74*, 37-67. doi: 10.1093/Poq/Nfp065

Horvitz, D. G., Greenberg, B. G., & Abernathy, J. R. (1976). Randomized Response - Data-Gathering Device for Sensitive Questions. *International Statistical Review, 44*, 181-196.

Horvitz, D. G., Shah, S., & Simmons, W. R. (1967). The Unrelated Question Randomized Response Model. *Proceedings of the Social Statistics Section, American Statistical Association*.

James, R. A., Nepusz, T., Naughton, D. P., & Petroczi, A. (2013). A potential inflating effect in estimation models: Cautionary evidence from comparing performance enhancing drug

and herbal hormonal supplement use estimates. *Psychology of Sport and Exercise, 14*, 84-96. doi: 10.1016/j.psychsport.2012.08.003

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

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

Kulka, R. A., Weeks, M. F., & Folsom, R. E. (1981). *A comparison of the randomized response approach and direct questioning approach to asking sensitive survey questions*. Working paper. NC: Research Triangle Institute.

Kundt, T. C., Misch, F., & Nerré, B. (2013). Re-assessing the merits of measuring tax evasions through surveys: Evidence from Serbian firms. ZEW Discussion Papers, No. 13-047. Retrieved Dec 12th, 2013, from http://hdl.handle.net/10419/78625

LaBrie, J. W., & Earleywine, M. (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. *Journal of Sex Research, 37*, 321-326.

Lamb, C. W., & Stem, D. E. (1978). An Empirical Validation of the Randomized Response Technique. *Journal of Marketing Research, 15*, 616-621. doi: Doi 10.2307/3150633

Landsheer, J. A., van der Heijden, P., & 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

Locander, W., Sudman, S., & Bradburn, N. (1976). An Investigation of Interview Method, Threat and Response Distortion. *Journal of the American Statistical Association, 71*, 269-275. doi: 10.2307/2285297

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

Mangat, N. S., & Singh, R. (1990). An Alternative Randomized-Response Procedure. *Biometrika, 77*, 439-442. doi: 10.1093/biomet/77.2.439

Marquis, K. H., Marquis, M. S., & Polich, J. M. (1986). Response Bias and Reliability in Sensitive Topic Surveys. *Journal of the American Statistical Association, 81*, 381-389. doi: 10.2307/2289227

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

Moors, J. J. A. (1971). Optimization of Unrelated Question Randomized Response Model. *Journal of the American Statistical Association, 66*, 627-629.

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., 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. M. (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*(3-4), 1-6.

Nepusz, T., Petroczi, A., Naughton, D. P., Epton, T., & Norman, P. (2014). Estimating the Prevalence of Socially Sensitive Behaviors: Attributing Guilty and Innocent Noncompliance With the Single Sample Count Method. *Psychological Methods, 19*, 334-355. doi: 10.1037/a0034961

O'Brien, D. (1977). *The Comprehension Factor in Randomized Response.* Ph.D. thesis, University of Wyoming, Laramie, Wyoming.

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.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and Denial in Socially Desirable Responding. *Journal of Personality and Social Psychology, 60*, 307-317. doi: 10.1037/0022-3514.60.2.307

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

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W. C., & Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology, 13*, 465-484. doi: 10.1002/(Sici)1099-0720(199910)13:5<465::Aid-Acp609>3.0.Co;2-Y

Scheers, N. J., & Dayton, C. M. (1987). Improved Estimation of Academic Cheating Behavior Using the Randomized-Response Technique. *Research in Higher Education, 26*(1), 61-69. doi: 10.1007/Bf00991933

Simon, P., Striegel, H., Aust, F., Dietz, K., & Ulrich, R. (2006). Doping in fitness sports: estimated number of unreported cases and individual probability of doping. *Addiction, 101*, 1640-1644. doi: 10.1111/j.1360-0443.2006.01568.x

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

Sudman, S., & Bradburn, N. (1974). *Response effects in surveys*. Chicago: Aldine.

Tian, G. L., & Tang, M. L. (2013). *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

Tracy, D. S., & Mangat, N. S. (1996). Some development in randomized response sampling during the last decade - a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science, 4*, 147-158.

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking Sensitive Questions: A Statistical Power Analysis of Randomized Response Models. *Psychological Methods, 17*(4), 623-641. doi: 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.

van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (1998). A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of social security fraud. *Kwantitatieve Methoden, 19*, 15-34.

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.

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

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology, 82*, 756-763.

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

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

Zdep, S. M., & Rhodes, I. N. (1977). Making the Randomized Response Technique Work. *Public Opinion Quarterly, 40*, 531-537.

Tables

Table 1

Mean percentage of correct responses as a function of questioning technique and educational level (standard errors in parentheses).

Group	Questioning Technique				
	DQ (control)	CDM	SLD	CWM	UCT
Total sample	90.40 (0.98)	75.37 (1.06)	68.70 (1.21)	83.35 (1.14)	77.06 (1.58)
High education	91.09 (1.37)	77.85 (1.53)	73.51 (1.79)	87.13 (1.45)	82.55 (1.96)
Low education	89.70 (1.39)	72.86 (1.43)	63.82 (1.56)	79.52 (1.72)	71.48 (2.42)

Note. DQ = Direct Question, CDM = Cheating Detection Model, SLD = Stochastic Lie Detector, CWM = Crosswise Model, UCT = Unmatched Count Technique.

Table 2

Mean rating of perceived privacy protection as a function of questioning technique and educational level (standard errors in parentheses).

Group	Questioning Technique				
	DQ (control)	CDM	SLD	CWM	UCT
Total sample	4.03 (0.08)	4.29 (0.07)	4.28 (0.07)	4.42 (0.07)	4.55 (0.07)
High education	3.78 (0.12)	4.25 (0.10)	4.23 (0.10)	4.37 (0.10)	4.68 (0.10)
Low education	4.28 (0.09)	4.34 (0.09)	4.33 (0.08)	4.47 (0.09)	4.42 (0.09)

Note. DQ = Direct Question, CDM = Cheating Detection Model, SLD = Stochastic Lie Detector, CWM = Crosswise Model, UCT = Unmatched Count Technique.

Table 3

Parametric correlation coefficients (Pearson's r) measuring the association of comprehension and perceived privacy protection.

Group	Questioning Technique				
	DQ (control)	CDM	SLD	CWM	UCT
Total sample	-.08	-.01	.07	.04	.09
High education	-.12	-.09	.03	.02	.03
Low education	-.01	.09	.16 *	.07	.13

Note. DQ = Direct Question, CDM = Cheating Detection Model, SLD = Stochastic Lie Detector, CWM = Crosswise Model, UCT = Unmatched Count Technique.
* $p < .05$

Figures

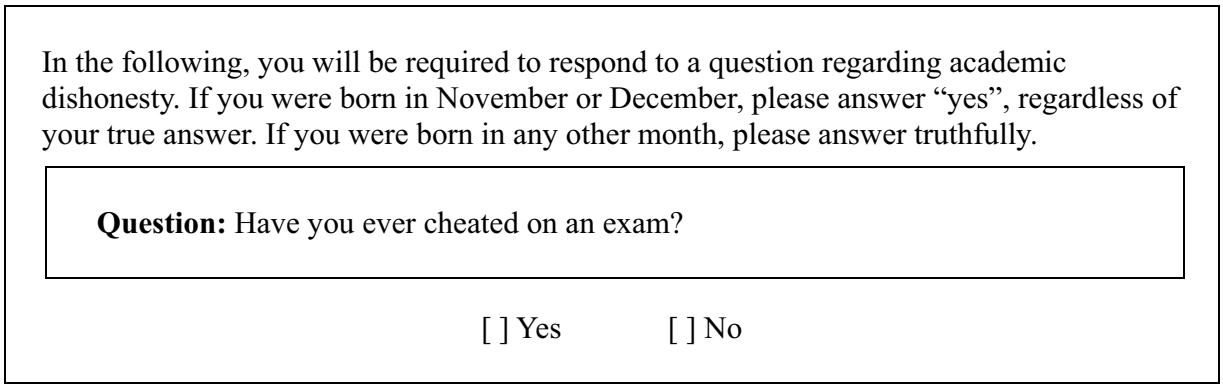

Figure 1. Example of a question regarding academic dishonesty as presented in surveys employing the Cheating Detection Model (Clark & Desharnais, 1998). The respondent’s month of birth is used as a randomization device with randomization probability $p = 2/12 = .17$.

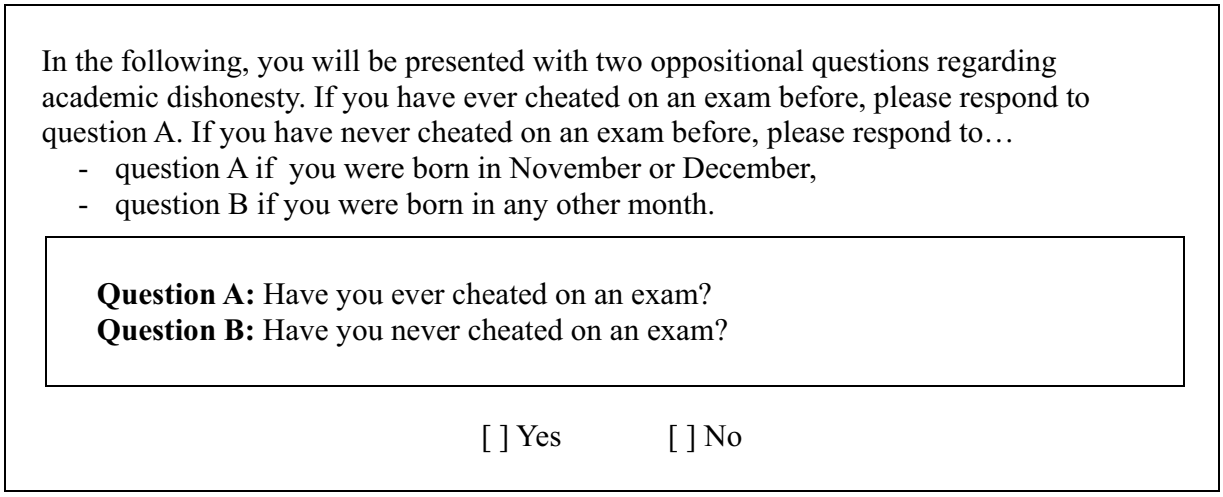

Figure 2. Example of a question regarding academic dishonesty using the Stochastic Lie Detector (Moshagen et al., 2012). The respondent's month of birth is used as a randomization device with randomization probability $p = 2/12 = .17$.

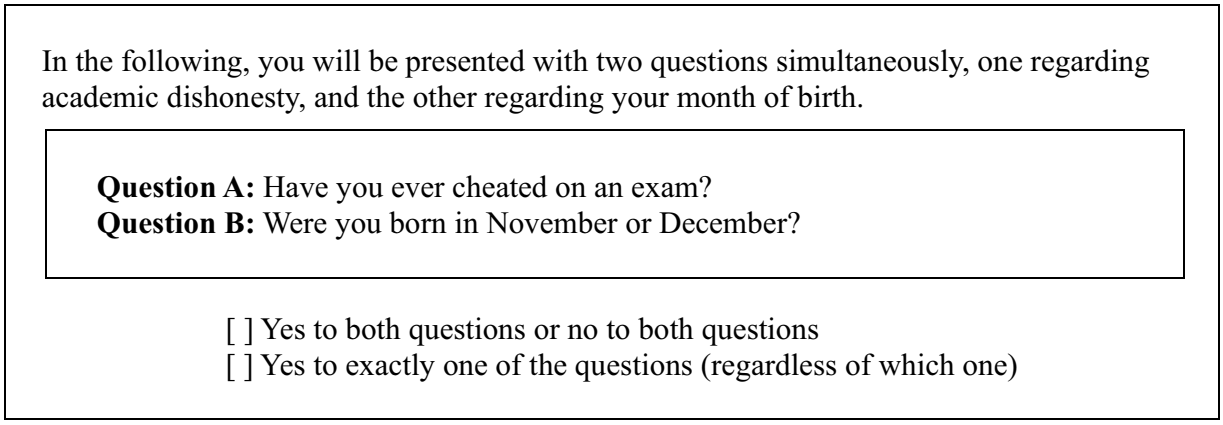

Figure 3. Example of a question regarding academic dishonesty using the Crosswise Model (Yu et al., 2008). The respondent's month of birth is used as a randomization device with randomization probability $p = 2/12 = .17$.

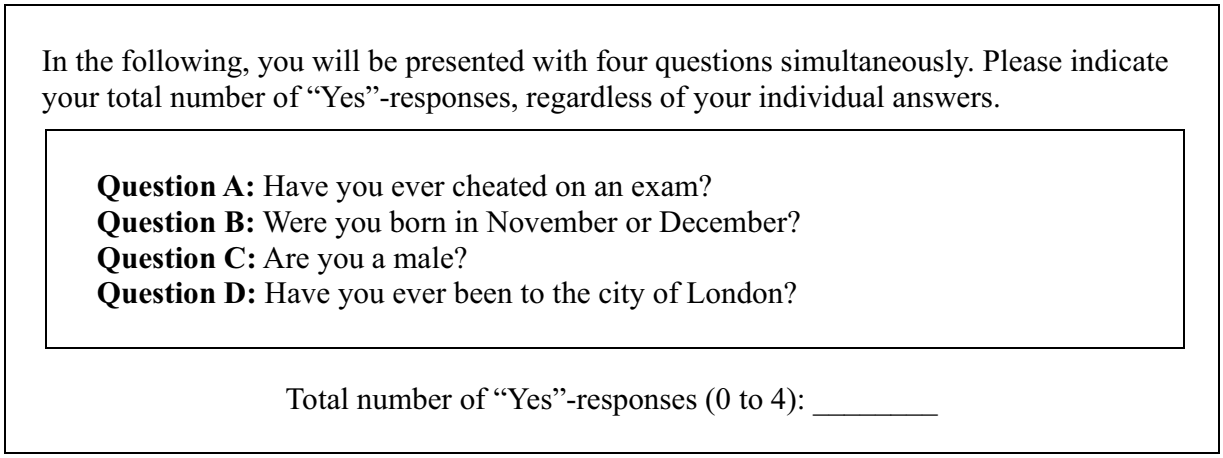

Figure 4. Example of a question regarding academic dishonesty using the Unmatched Count Technique (Miller, 1984) with one sensitive (A) and three non-sensitive questions (B to D).

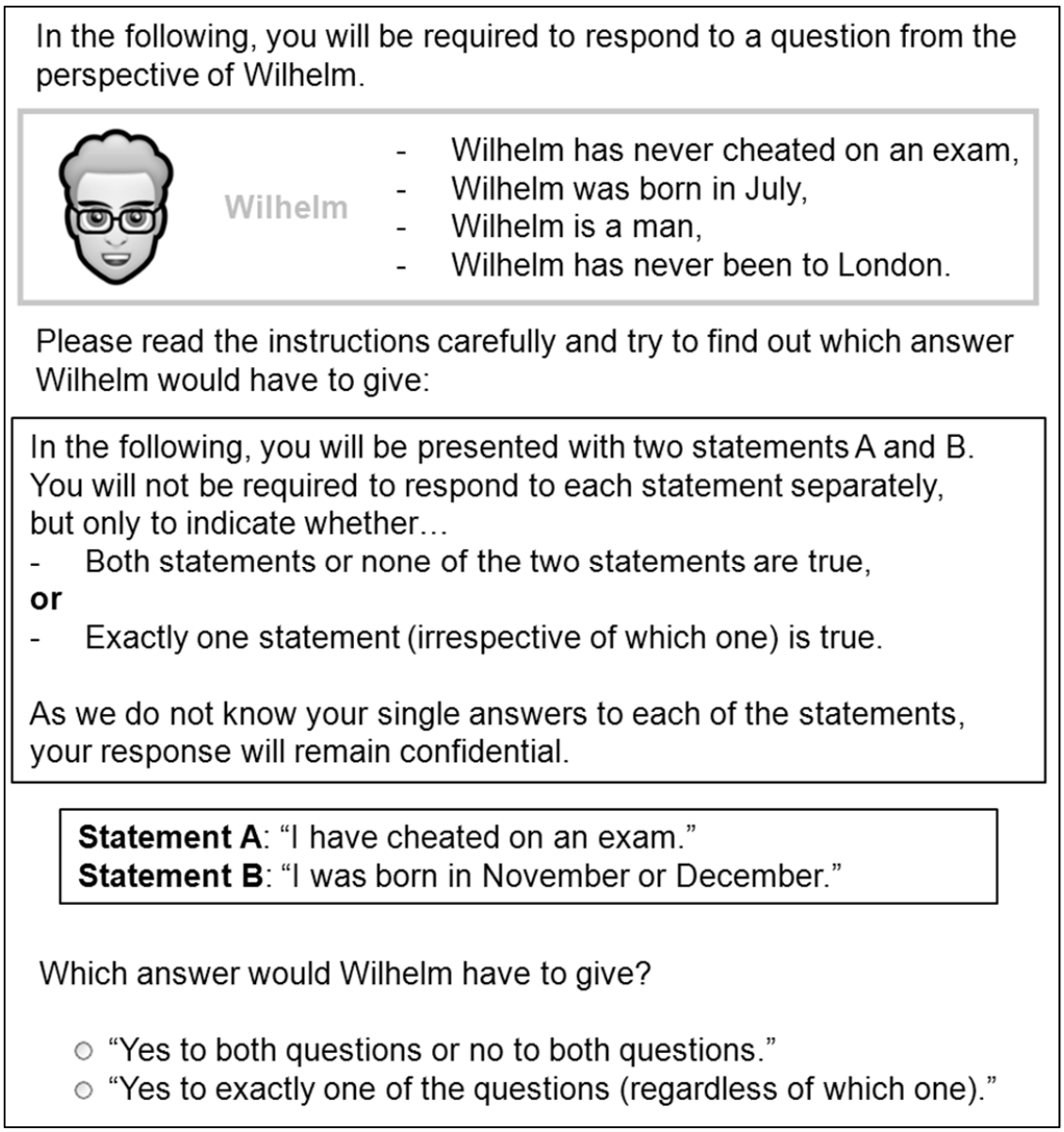

Figure 5. Screenshot of a CWM question that had to be answered from the perspective of the fictional character Wilhelm. As Wilhelm never cheated on an exam and was born in July, the first answer option ("Yes to both questions or no to both questions.") would have been correct.

A strong validation of the Crosswise Model using experimentally induced cheating behavior

Adrian Hoffmann[1], Birk Diedenhofen[1], Bruno J. Verschuere[2], & Jochen Musch[1]

[1]University of Duesseldorf, [2]University of Amsterdam, Ghent University, Maastricht University

Author Note

Adrian Hoffmann, Birk Diedenhofen, and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf, Germany.

Bruno J. Verschuere, Department of Clinical Psychology, University of Amsterdam, The Netherlands; Department of Psychology, Ghent University, Ghent, Belgium; Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands.

Correspondence concerning this article should be addressed to Adrian Hoffmann, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, 40225 Duesseldorf, Germany. E-mail: adrian.hoffmann@uni-duesseldorf.de

Abstract

We constructed an online cheating paradigm that could be used to validate the Crosswise Model (CWM; Yu, Tian, & Tang, 2008), a promising indirect questioning technique designed to control for socially desirable responding on self-reports. Participants qualified for a reward only if they could identify the target words from three anagrams, one of which was virtually unsolvable as shown on a pretest. Of the 664 participants, 15.51% overreported their performance and were categorized as cheaters. When participants were asked to report whether they had cheated, a conventional direct question resulted in a substantial underestimate (5.07%) of the known prevalence of cheaters. Using a CWM question resulted in a more accurate estimate (13.03%). This result shows that the CWM can be used to control for socially desirable responding and provides estimates that are much closer to the known prevalence of a sensitive personal attribute than those obtained using a direct question.

Keywords: Crosswise Model, Randomized Response Technique, Strong validation, Cheating

A strong validation of the Crosswise Model using experimentally induced cheating behavior

On surveys that ask for direct self-reports to assess sensitive personal attributes, some individuals tend to refrain from answering truthfully in order to present themselves in a socially desirable light (Paulhus, 1991; Tourangeau & Yan, 2007). Such socially desirable responding may result in biased prevalence estimates of socially desirable or undesirable attitudes and behaviors (Krumpal, 2013; Phillips & Clancy, 1972). The validity of prevalence estimates is particularly threatened when employing a question that directly asks whether the respondent embodies the sensitive attribute (*direct questioning*; DQ).

Scientific research has spawned various approaches that aim to measure or control for the influence of the social desirability bias (Nederhof, 1985; Paulhus, 1991). To increase the validity of prevalence estimates based on self-reports, indirect questioning procedures such as the Randomized Response Technique (RRT; Warner, 1965) have been proposed. In the original RRT procedure, respondents are instructed to answer either a positively ("Do you embody the sensitive attribute?") or negatively worded sensitive question ("Do you not embody the sensitive attribute?"), depending on the outcome of an external randomization process (e.g., the cast of a die). The individual outcome of the randomization process remains unknown to the questioner, but the distribution of the randomization outcomes and thus the probability of selecting either of the questions are known at sample level. Therefore, neither a "yes" nor a "no" response will allow any conclusions to be drawn about an individual respondent's true status with respect to the sensitive attribute. The perceived confidentiality of responses and the respondent's probability of answering truthfully is thereby increased. Despite the increased confidentiality at the level of the individual, the known distribution of the randomization outcome allows an estimate to be made of the prevalence of a sensitive attribute in the sample with presumably less

of a bias from socially desirable responding. However, unsystematic variance is added to individual responses; estimates obtained using indirect questions therefore suffer from increased variance compared with more conventional direct questions. This decrease in efficiency is supposed to be compensated for by an increase in validity, especially if the attribute in question is highly sensitive in nature (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). However, a loss in efficiency is considered tolerable only to the extent to which the use of an indirect question actually results in more valid prevalence estimates than the use of a direct question.

As a promising new approach for controlling for socially desirable responding, Yu et al. (2008) proposed the Crosswise Model (CWM). This technique allows questions to be asked indirectly without requiring participants to operate an external randomization device (e.g., a die). Questions asked in the CWM format require participants to respond to two statements simultaneously. One of these statements asks for the sensitive attribute ("I embody the sensitive attribute X") with unknown prevalence π; the other statement asks about a second nonsensitive attribute for which the prevalence p is known in the population (e.g., "I was born in November or December"). Participants are then instructed to indicate whether (a) "either statement or neither of the two statements is true," or whether (b) "exactly one of the two statements (regardless of which one) is true." Neither of the two answer options (a) or (b) allows inferences to be made about a respondent's individual status with regard to the sensitive attribute. For the entire sample, however, Yu et al. (2008) showed that π can be estimated by

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

with *n'* representing the total number of "both true/both false" responses and *n* reflecting the sample size. An estimate of the variance of $\hat{\pi}$ is given by

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

Whereas the CWM is mathematically equivalent to the original RRT model proposed by Warner (Ulrich, Schröter, Striegel, & Simon, 2012), its questioning format offers potential advantages over previous approaches. As the randomization procedure is integrated into the question itself, the instructions are simpler than in competing RRT procedures; this makes the CWM presumably "easy to operate for both interviewer and interviewee" (Yu et al., 2008, p. 255). Therefore, Tian and Tang (2013) suggested that the CWM be referred to as a "Nonrandomized Response Technique." Indeed, a recent study found the CWM to be superior to various competing indirect questioning techniques with regard to comprehensibility and perceived privacy protection (Hoffmann, Schmidt, Waubert de Puiseau, & Musch, 2014). Moreover, the CWM offers response symmetry in the sense that none of the answer options offers a "safe alternative" to which respondents might turn to dispel any connection to the sensitive attribute. Response symmetry has been shown to increase compliance with instructions and hence to increase the validity of prevalence estimates (Ostapczuk, Moshagen, Zhao, & Musch, 2009). However, whether the use of CWM questions does, indeed, lead to more valid prevalence estimates has yet to be shown in empirical validation studies.

Validation studies for indirect questioning techniques can be roughly divided into two categories (Moshagen, Hilbig, Erdfelder, & Moritz, 2014): "Weak" validation studies compare prevalence estimates obtained with different questioning techniques (e.g., direct vs. indirect questions). For socially undesirable attributes, higher prevalence estimates are usually considered to have higher validity, assuming that the social desirability bias leads to underestimates of the true value ("more-is-better" criterion; Umesh & Peterson, 1991). However, higher prevalence estimates may still under- or overestimate the true rate at which people embody a sensitive

attribute if the prevalence in the sample and the prevalence in the population ("ground truth") remain unknown. Consequently, such studies can offer only weak evidence for the validity of a questioning technique. By contrast, "strong" validation studies rely on a sensitive attribute for which the true prevalence in the sample is known. If a questioning technique provides a prevalence estimate close to the known prevalence in the sample, this accordance is considered particularly strong evidence for its validity. Strong validation studies are considered the gold standard in the evaluation of methods that aim to control for social desirability bias (Lensvelt-Mulders et al., 2005; Moshagen et al., 2014). Unfortunately, the assessment of the true proportion of respondents in a sample who carry a sensitive attribute is usually costly and often impossible. For example, assessing a sample of individuals known to have been convicted of social welfare fraud (van der Heijden, van Gils, Bouts, & Hox, 2000) is possible only if legal regulations grant researchers access to the respective contact data. Moreover, accessing such sensitive information consumes time and money and may raise some ethical concerns because the individual status of respondents with respect to the sensitive attribute will be uncovered. Consequently, strong validation studies implementing indirect questioning techniques are very scarce; in Lensvelt-Mulders et al. (2005) meta-analysis, only six of 38 validation studies met this criterion. With respect to the CWM, some weak validation studies using the "more-is-better" approach have shown that the application of the CWM results in higher prevalence estimates than a conventional direct question for plagiarism in student papers (Jann, Jerke, & Krumpal, 2012), tax evasion (Korndörfer, Krumpal, & Schmukle, 2014; Kundt, Misch, & Nerré, 2013), steroid use (Nakhaee, Pakravan, & Nakhaee, 2013), and xenophobia and islamophobia (Hoffmann & Musch, 2014). However, only a strong validation study using a sensitive attribute with known prevalence can provide firm evidence for the validity of prevalence estimates

obtained using the CWM. To the best of our knowledge, a strong validation study has yet to be conducted to investigate the CWM.

As was recently shown by Moshagen et al. (2014), a promising approach that can be applied to overcome the notorious difficulties associated with strong validation studies is the experimental induction of a socially undesirable attribute. Moshagen et al.'s participants were instructed to secretly roll a die and to report the outcome, under the condition that certain outcomes were financially rewarded (an adjusted "die-under-the-cup-paradigm"; e.g., Hilbig & Hessler, 2013). The proportion of respondents claiming a reward was about twice as high as would be expected if all participants had honestly reported the outcome of their die roll. Consequently, about half of the respondents claiming a reward had to be categorized as cheaters (Moshagen et al., 2014). When respondents were asked whether they had cheated, a direct question substantially underestimated the actual rate of cheaters in the sample. By contrast, an application of the RRT that was based on the Stochastic Lie Detector (SLD; Moshagen, Musch, & Erdfelder, 2012) resulted in a prevalence estimate that was fairly close to the true prevalence. Therefore, the SLD was evaluated favorably as a method that was capable of controlling for socially desirable responding (Moshagen et al., 2014).

Whereas the rationale of the die-under-the-cup task appears compelling, the procedure requires space, material, preparation, and a personal interaction with every respondent, making it difficult to employ in an online setting (but see, e.g., Shalvi, Dana, Handgraaf, & De Dreu, 2011). Online surveys, however, offer an attractive environment for indirect questioning studies as the large sample sizes that are required to compensate for the increased variance in indirect prevalence estimates can be efficiently assessed using the World Wide Web (Musch, Bröder, & Klauer, 2001). Therefore, we used a validation method similar to the die-under-the-cup-paradigm

presented in Moshagen et al. (2014) but developed it further into a format that could be employed in online assessments. Our approach was based on the "word-jumble task" proposed by Wiltermuth (2011). In this paradigm, participants are presented with a list of nine words with a jumbled letter sequence. They are instructed to rearrange the letter sequence of each jumble in their minds until the letters form a valid word; instead of having to report the solution, however, they are asked merely to report the number of consecutive jumbles they were able to solve from the beginning. They are told that a higher reported number of successively solved jumbles will be remunerated with a higher financial reward. However, the third consecutive anagram is chosen to be practically unsolvable. On pretests of several studies using the word-jumble task, 0 out of 10 (Halevy, Shalvi, & Verschuere, 2013), 0 out of 30 (Wiltermuth, 2011), and 0 out of 42 (Gino & Mogilner, 2014) participants were able to identify the solution of the third anagram. Hence, any participant reporting to have solved three or more consecutive scrambled words is most likely a cheater. In the original study by Wiltermuth (2011), about 29% of the respondents claimed to have solved three or more anagrams in a row. Remarkably, participants "perceived the act of over-reporting [their] performance as unethical and greedy" (p. 162), strongly indicating the socially undesirable nature of such behavior. Other studies reported cheating rates of 23% (Halevy et al., 2013), 52% (Ruedy, Moore, Gino, & Schweitzer, 2013), and 40% to 73% under various experimental conditions (Gino & Mogilner, 2014). These results suggest that the word-jumble task is reliably capable of experimentally generating a substantial rate of cheaters in a given sample, even though cheating on the task is perceived as socially undesirable by the participants.

In the present study, we pursued two goals: First, on the basis of a variant of the anagram cheating task proposed by Wiltermuth (2011), we constructed an experimental procedure that

allowed for an online application of the validation method used by Moshagen et al. (2014). Second, we conducted the first strong validation of the CWM (Yu et al., 2008) using cheating behavior induced via the anagram-cheating task as an external criterion. Because dark personality traits such as psychopathy have been discussed as potential moderators of cheating behavior (Halevy et al., 2013), we included the "dirty dozen" items (Jonason & Webster, 2010; Küfner, Dufner, & Back, 2014) as brief measures of psychopathy, Machiavellism, and narcissism.

Method

Participants and Design

A total of 698 registered members of a noncommercial research panel run by scientific employees of the University of Düsseldorf accessed the online questionnaire. Pretest participants were not eligible for the main study. Respondents who did not complete the experiment (34) were excluded from the analyses, resulting in a final sample size of $N = 664$ (95.1% of the participants who had accessed the survey). The mean age of the respondents was 36.7 years ($SD = 13.73$, $Min = 18$, $Max = 94$); 380 (57.2%) of the participants were female.

Using a single-factor between-subjects design, we randomly assigned participants to respond to a question about their cheating behavior either in a DQ ($n = 138$; 20.8% of the final sample) or CWM format ($n = 526$). A skewed allocation ratio of about 1:4 was chosen to compensate for the comparably low efficiency of prevalence estimates obtained via the CWM (Ulrich et al., 2012).

Measures

Anagram cheating task. Previous versions of the word-jumble task had been constructed and tested only in English (Gino & Mogilner, 2014; Ruedy et al., 2013; Wiltermuth, 2011) or Dutch (Halevy et al., 2013). We therefore created a new set of anagrams from scratch

and conducted a pretest to identify German items with suitable item difficulties. Our objective was to find two anagrams that would be solved by virtually every participant ($p > .99$) and one anagram that was practically unsolvable ($p < .01$). To this end, we made use of three anagram characteristics that had been found to be associated with item difficulty in previous studies. First, anagrams with frequently used target words are usually easier to solve than anagrams with rarely used target words (e.g., Dominowski, 1967; Lemay, 1972; Mayzner & Tresselt, 1958, 1959; Mendelsohn, 1976). Second, when a larger number of moves are needed to transform the anagram letters into the target word, the difficulty of an anagram usually increases (e.g., Dominowski, 1966; Mendelsohn & Obrien, 1974). Finally, anagrams with a high initial similarity to their respective target word are presumably easier to solve. Previous studies have measured this similarity using the number of letters in the anagram that remain in the correct sequence of the target word (Gilhooly & Johnson, 1978, p. 61) or by considering the correlations between the letter sequences of the anagram and the target word (e.g., Johnson, 1966; Terakoa, 1959). We considered all three of these variables while compiling our preliminary item pool. As a basis for the identification of potential target words, we used the German corpus provided by the *Deutscher Wortschatz* project (Quasthoff, Richter, & Biemann, 2006). This corpus contains information on the frequency classes of over 1,000,000 words in the German language based on a random sample of 1,000,000 sentences drawn from the German Wikipedia (http://de.wikipedia.org) in 2010. In this corpus, the frequency class of a word denotes its relative frequency in relation to the most frequent word in the corpus and is computed as

$$N = 0.5 - \log_2\left(\frac{F_i}{F_{max}}\right) \tag{3}$$

with F_i denoting the frequency of the word in question and F_{max} denoting the frequency of the most common word in the corpus (cf. Zipf, 1935). Hence, higher values indicate lower frequencies. Aside from providing linguistic corpora, the *Deutscher Wortschatz* project also offers various services, some of which were used in this study. For example, words can be checked for grammatical properties and for anagrams in particular. To avoid enabling participants to derive hypotheses about item difficulties based on word length and to allow the use of clear-cut instructions, we extracted a set of six-letter German nouns in their basic form from the corpus but did not include any personal names or proper nouns. Words that had another German word as an anagram and words with double letters were also excluded to ensure that all anagrams would have only one solution. From the remaining word pool, we chose 32 words with a high frequency in the German language (frequency classes seven to 11) as possible target words for easy anagrams, and 13 with a low frequency (frequency classes 18 to 22) as possible target words for difficult anagrams. To create potentially easy anagrams, we jumbled the letters in a way that ensured that participants would need to move only one letter to solve it, that five letters would already be in the correct order of the target word, and that the rank correlation between the letter sequences of the target word and the anagram would be high ($\tau > .50$). Difficult anagrams were constructed by jumbling the letters in a way that ensured that the maximum possible number of five letter moves would be needed to find the solution, that none of the letters would remain in the correct order of the target word, and that the rank correlation between the letter sequences of the target word and the anagram would be close to zero ($-.10 < \tau < .10$). A full list of all 45 anagrams thus identified for our pilot study, along with their respective target words and item characteristics is presented in Appendix A. These anagrams were pretested on 136 individuals recruited online via the noncommercial research panel run by members of the

Department of Experimental Psychology at the University of Düsseldorf. This panel was also used for the subsequent main study. Items were presented in a random order. For each anagram, participants were given a maximum of 20 s to identify the target word. If participants identified the target word, they were asked to press a button labeled “I found the target word and want to enter the solution.” After the 20-s presentation time had elapsed or when the respective button was pressed, the anagram was masked and respondents were given another 20 s to enter the solution into an input box. Answers were forced to have exactly six letters and were scored as correct only when they provided a perfect match with the target word. Three anagrams showed item characteristics that made them perfectly suitable for the construction our cheating task (see Table 1): Two very easy anagrams were solved by every single pretest participant, and one difficult anagram was solved by none of the 136 respondents.

TAKE IN TABLE 1

In the final anagram-cheating task and as in the original paradigm used by Wiltermuth (2011), we simply asked participants to indicate the number of anagrams they were able to unscramble. Deviating from the original paradigm, however, only three anagrams were presented in a fixed order, and respondents were not asked to consider whether they had solved the anagrams successively. The two very easy anagrams “APRXIS” (target word “PRAXIS”; translating to “practice”; solved by every respondent on the pretest) and “UMELTW” (target word “UMWELT”; translating to “environment”; also solved by every respondent on the pretest) were presented first, followed by the virtually unsolvable anagram “IERTLO” (target word “TRIOLE”; translating to “triplet”; not solved by any respondent on the pretest). The three anagrams were presented for a maximum of 20 s each; participants could abort the presentation of each anagram if they believed they had identified the solution. On the basis of the pretest, we

expected that all or virtually all participants would be able to unscramble two of the anagrams but that virtually no participant would be able to identify all three target words. With a very low expected rate of false alarms, respondents claiming to have solved all three anagrams were therefore categorized as cheaters. Following the paradigm proposed by Moshagen et al. (2014), we offered an incentive for claiming that all three anagrams were solved, expecting that this would motivate a substantial number of participants to cheat knowingly.

Direct question. Participants in the DQ condition were directly questioned about whether the following statement was true or false: "On the anagram task, I claimed that I had solved more anagrams than I had actually solved."

Crosswise Model. If assigned to the CWM condition, participants were simultaneously confronted with two statements. The wording of the sensitive statement was identical to the DQ condition. A second nonsensitive statement read: "I was born in November or December." The probability of being born in these months is about 15.8% according to official birth statistics provided by the German Federal Agency for Statistics (cf. Moshagen et al., 2012). Respondents were instructed to choose one of the two available answer options that stated (a) "Either statement or neither of the two statements is true" versus (b) "Exactly one statement (regardless of which one) is true," respectively.

Procedure

The study was administered as an online questionnaire using the EFS survey 10.2 software (Questback, 2014). The first page welcomed participants and obtained their informed consent. After participants were asked to provide demographic information, they were given the instructions for the anagram-cheating task and were informed that they would be given the chance to partake in a lottery for three gift certificates worth 20€, 30€, and 100€ if they

performed well. Before the anagram-cheating task began, respondents completed two quite easy example anagrams to familiarize themselves with the paradigm. After completing the anagram-cheating task, respondents reported on a separate page how many anagrams they had solved. In the upper part of this page, all three anagrams were displayed again, along with their respective target words. In the middle part, participants were reminded that only a high performance would allow them to participate in the lottery. At the lower end of the page, the question about their performance on the anagram task asked: "How many of the anagrams did you solve in the available time?" The answer options were: "I did not solve any of the three anagrams," "I solved one of the three anagrams," "I solved two of the three anagrams," and "I solved all three anagrams." Only the last of these potential answers was identified as allowing the participant to partake in the lottery at the end of the survey. On the next page, depending on the experimental condition, a question in either a DQ or CWM format asked about whether respondents had just cheated on the anagram task by overreporting their results. Subsequently, the "dirty dozen" items (Jonason & Webster, 2010; Küfner et al., 2014) were presented on a single page to assess the dark personality constructs of psychopathy, Machiavellism, and narcissism. On the last two pages, participants were debriefed, thanked for their cooperation, and offered the opportunity to enter the lottery. Because no one should be discriminated against for being honest, we offered all participants the opportunity to participate in the lottery regardless of their performance.

Statistical Analysis

In the CWM condition, an estimate for the prevalence of cheating can be obtained by using Equations 1 and 2. In the DQ condition, the proportion of respondents answering "true" to a direct question provides a direct estimate of the prevalence of cheating. Following previous studies implementing indirect questioning techniques (e.g., Moshagen et al., 2012; Ostapczuk et

al., 2009; Ostapczuk & Musch, 2011), we translated both the DQ and CWM conditions into a combined multinomial processing tree model (MPT; Batchelder, 1998; Hu & Batchelder, 1994). This approach offers more flexibility in parameter estimation and offers convenient statistical tests of parameter restrictions (Moshagen et al., 2012). We obtained estimates for the prevalence of cheating in the DQ and CWM conditions via maximum likelihood procedures using the *multiTree* v0.41 software (Moshagen, 2010). Model equations and empirically observed answering frequencies that were entered into the analysis are reproduced in Appendices B and C. To compare parameter estimates with each other and with the known true proportion of cheaters, differences in model fit between an unrestricted baseline model and a restricted alternative model (e.g., in which the DQ and CWM prevalence estimates were equalized) were assessed via differences in the asymptotically X^2-distributed log-likelihood ratio statistic G^2 (Read & Cressie, 1988).

Results

Cheating Behavior

Of the 664 participants, 103 (15.51%) claimed that they had solved all three anagrams. As the probability of achieving a score of three anagrams had been shown to be virtually zero on the pretest, we used this percentage as a proxy for the "true" prevalence of cheating in our sample. Of the remaining 561 respondents who were classified as noncheaters, 558 (99.47% of the noncheaters, 84.04% of the sample) reported that they had solved exactly two anagrams, whereas only two participants (0.36% of the noncheaters, 0.30% of the sample) indicated that they had solved one anagram, and one participant (0.18% of the noncheaters, 0.15% of the sample) reported being unable to unscramble any words. These findings confirmed the results of

our pretest and met our expectations that virtually all participants would solve two anagrams but that some would choose to overreport their performance to maximize their personal benefit.

Prevalence Estimates

As Table 2 shows, the application of a direct question resulted in a significant underestimation (5.07%) of the known "true" prevalence of cheating (15.51%) by about 10.34%, ΔG^2 ($df = 1$) = 14.88, $p < .001$. Thus, about two thirds of the cheaters in the DQ condition apparently refrained from telling the truth, perhaps to conceal that they had engaged in a socially undesirable behavior. By contrast, the prevalence estimate obtained via the indirect question in the CWM format (13.03%) deviated by only about 2.48% from the "true" prevalence, and this small difference was not reliably different from zero, ΔG^2 ($df = 1$) = 0.79, $p = .37$. Thus, the CWM seemed to be able to obtain an unbiased estimate for the prevalence of cheating in the sample. Finally, the estimates obtained in the DQ and CWM conditions (5.07% vs. 13.03%) differed significantly from each other, ΔG^2 ($df = 1$) = 5.29, $p < .05$.[1]

TAKE IN TABLE 2

Discussion

The present study aimed to achieve two objectives: First, we wanted to develop an online paradigm that could induce a socially undesirable attribute and provide information about whether a participant embodied this attribute. To this end, we created an anagram task that incited participants to overreport their performance, thus inducing undesirable cheating behavior. In a second step, we used this anagram-cheating task to conduct a strong validation of the Crosswise Model (CWM; Yu et al., 2008), an indirect questioning technique designed to control for socially desirable responding. The known rate of cheaters in the sample served as an objective external criterion for the validity of the CWM prevalence estimate. In light of the

present results, the application of the anagram-cheating task seemed to have successfully motivated a substantial proportion of participants to engage in a socially undesirable behavior. About 15.51% of the participants claimed that they had obtained a result that was considered virtually impossible to achieve according to the findings of the pretest. However, only 5.07% of the respondents honestly admitted their cheating when asked using a conventional direct question. We attribute this difference between the levels of reported and actual behavior to the influence of the social desirability bias. Hence, our results further support the notion that the assessment of sensitive attributes via direct self-reports may lead to an underestimation of their prevalence, as some carriers of the sensitive characteristic tend to conceal their true status (Krumpal, 2013; Phillips & Clancy, 1972). With respect to the prevalence estimate obtained in the CWM condition, the present study also provides evidence for the validity of this questioning technique according to the more-is-better criterion (Jann et al., 2012; Kundt et al., 2013). This is because the CWM estimate of about 13.03% substantially exceeded the estimate that was based on direct questioning, and therefore was presumably less biased by socially desirable responding. Most importantly, however, our data provide the first strong evidence for the validity of the CWM as the prevalence estimate obtained using the CWM question concurred with the known prevalence of cheating. This finding suggests that the application of the CWM will indeed allow researchers to obtain prevalence estimates that are unbiased by social desirability.

Previous studies implementing similar anagram tasks have reported even higher cheating rates of 23% (Halevy et al., 2013) to 73% (Gino & Mogilner, 2014). The lower prevalence of cheating behavior in our sample may have been due to differences in the format of the task. In contrast to the original word-jumble task (Wiltermuth, 2011), our anagram-cheating task consisted of three items, and participants were asked merely to report the number of anagrams

they had solved regardless of whether the anagrams were solved consecutively. Furthermore, our target words were presented just before participants reported their personal performance. Participants were thus well aware of the performance they had just given. Consequently, respondents had to lie rather blatantly to claim that they had solved all three anagrams. Another possible explanation for the somewhat lower cheating rate we observed may lie in the reward respondents were offered in the present study. In Wiltermuth (2011), Halevy et al. (2013), and Gino and Mogilner (2014), participants were guaranteed a fixed amount of money for each anagram they solved, and this money was presumably paid out directly after the experiment. In the present study, a better performance allowed participants only to participate in a gift-certificate lottery, which might have had a weaker effect on their motivation to cheat. Finally, differences in cheating rates may also be explained by differences in the samples and the survey setting.

In conclusion, we developed and employed a new online anagram-cheating task that can be used to induce a socially undesirable attribute with a known prevalence. This task is simple, efficient, and can easily be adopted, offering researchers the potential to conduct further strong validation studies with a reasonable amount of effort. Future studies investigating sensitive attributes and the influence of socially desirable responding might therefore profit from employing this online anagram task. Most importantly, however, the present investigation is the first to provide strong evidence that the CWM is convincingly capable of obtaining valid prevalence estimates of sensitive attitudes and behaviors. We therefore conclude that the CWM appears to be a very promising indirect questioning technique that can be used to successfully control for social desirability on surveys of sensitive behavior.

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

Dominowski, R. L. (1966). Anagram Solving as a Function of Letter Moves. *Journal of Verbal Learning and Verbal Behavior, 5*, 107-111. doi: 10.1016/S0022-5371(66)80002-6

Dominowski, R. L. (1967). Anagram Solving as a Function of Bigram Rank and Word Frequency. *Journal of Experimental Psychology, 75*, 299-306. doi: 10.1037/H0025060

Gilhooly, K. J., & Johnson, C. E. (1978). Effects of solution word attributes on anagram difficulty - a regression analysis. *Quarterly Journal of Experimental Psychology, 30*, 57-70. doi: 10.1080/14640747808400654

Gino, F., & Mogilner, C. (2014). Time, Money, and Morality. *Psychological Science, 25*, 414-421. doi: 10.1177/0956797613506438

Halevy, R., Shalvi, S., & Verschuere, B. (2013). Being Honest About Dishonesty: Correlating Self-Reports and Actual Lying. *Human Communication Research, 40*, 54-72.

Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology, 49*, 263-266. doi: 10.1016/j.jesp.2012.11.010

Hoffmann, A., & Musch, J. (2014). *Assessing the validity of two indirect questioning techniques: a Stochastic Lie Detector versus the Crosswise Model*. Manuscript under preparation.

Hoffmann, A., Schmidt, A. F., Waubert de Puiseau, B., & Musch, J. (2014). *On the comprehensibility and perceived privacy protection of indirect questioning techniques*. Manuscript under preparation.

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

Johnson, D. M. (1966). Solution of Anagrams. *Psychological Bulletin, 66*, 371-384. doi: 10.1037/H0023886

Jonason, P. K., & Webster, G. D. (2010). The Dirty Dozen: A Concise Measure of the Dark Triad. *Psychological Assessment, 22*, 420-432. doi: 10.1037/A0019265

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

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

Küfner, C. P., Dufner, M., & Back, M. D. (2014). Das Dreckige Dutzend und die Niederträchtigen Neun – Kurzskalen zur Erfassung von Narzissmus, Machiavellismus und Psychopathie. *Diagnostica*, 1-16. doi: 10.1026/0012-1924/a000124

Kundt, T. C., Misch, F., & Nerré, B. (2013). Re-assessing the merits of measuring tax evasions through surveys: Evidence from Serbian firms. ZEW Discussion Papers, No. 13-047. Retrieved Dec 12th, 2013, from http://hdl.handle.net/10419/78625

Lemay, E. H. (1972). Anagram Solutions as a Function of Task Variables and Solution Word Models. *Journal of Experimental Psychology, 92*, 65-68. doi: 10.1037/H0032164

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

Mayzner, M. S., & Tresselt, M. E. (1958). Anagram Solution Times: A Function of Letter Order and Word-Frequency. *Journal of Experimental Psychology, 56*, 376-379. doi: 10.1037/H0041542

Mayzner, M. S., & Tresselt, M. E. (1959). Anagram Solution Times: A Function of Transition-Probabilities. *Journal of Psychology, 47*, 117-125.

Mendelsohn, G. A. (1976). An Hypothesis Approach to the Solution of Anagrams. *Memory & Cognition, 4*, 637-642. doi: 10.3758/Bf03213228

Mendelsohn, G. A., & Obrien, A. T. (1974). The Solution of Anagrams - A Reexamination of the Effects of Transition Letter Probabilities, Letter Moves, and Word Frequency on Anagram Difficulty. *Memory & Cognition, 2*, 566-574. doi: 10.3758/Bf03196922

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

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

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*(3-4), 1-6.

Nederhof, A. J. (1985). Methods of Coping with Social Desirability Bias - a Review. *European Journal of Social Psychology, 15*, 263-280. doi: 10.1002/ejsp.2420150303

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

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

Quasthoff, U., Richter, M., & Biemann, C. (2006). *Corpus Portal for Search in Monolingual Corpora.* Paper presented at the fifth international conference on Language Resources and Evaluation, LREC, Genoa.

Questback. (2014). Unipark EFS Survey 10.2. Retrieved from http://www.unipark.de

Read, T. R., & Cressie, N. A. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

Ruedy, N. E., Moore, C., Gino, F., & Schweitzer, M. E. (2013). The Cheater's High: The Unexpected Affective Benefits of Unethical Behavior. *Journal of Personality and Social Psychology, 105*, 531-548. doi: 10.1037/A0034231

Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes, 115*, 181-190. doi: 10.1016/j.obhdp.2011.02.001

Terakoa, T. (1959). Effects of letter-orders and material words on the anagram solution. *Japanese Journal of Psychology, 30*, 253-263.

Tian, G. L., & Tang, M. L. (2013). *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*(4), 623-641. doi: 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.

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.

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

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes, 115*, 157-168. doi: 10.1016/j.obhdp.2010.10.001

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

Zipf, G. K. (1935). *The psycho-biology of language*. Oxford, England: Houghton-Mifflin.

Footnotes

[1] None of the three dark personality traits we assessed (psychopathy, Machiavellism, and narcissism) showed any association with cheating behavior or respondents' answering behavior in the DQ condition. Therefore, we will not report them in detail. However, for the purpose of future meta-analyses and to avoid a potential publication bias, we will report the correlation coefficients we found: psychopathy and cheating ($N = 644$): $r = .04$, $p = .33$; Machiavellism and cheating ($N = 644$): $r = .07$, $p = .10$; narcissism and cheating ($N = 644$): $r = .03$, $p = .52$. In the DQ condition, we also found no significant correlations between psychopathy and lying ($N = 137$), $r = .09$, $p = .32$; Machiavellism and lying ($N = 137$): $r = .07$, $p = .42$, and narcissism and lying ($N = 137$): $r < .01$, $p > .99$.

Table 1

Item characteristics of the three anagrams that were chosen for the anagram-cheating task on the basis of the pretest results (N = 136)

	Target word	Anagram	Translation	% solved	M_{pt} (SD)	M_{rt} (SD)
1	PRAXIS	APRXIS	practice	100.00	1.99 (1.36)	3.51 (1.43)
2	UMWELT	UMELTW	environment	100.00	1.90 (0.93)	3.09 (1.36)
3	TRIOLE	IERTLO	triplet	0.00	18.73 (3.60)	5.30 (5.90)

Note. M_{pt} = mean presentation time in s; M_{rt} = mean response time in s.

Table 2

Parameter estimates (standard errors in parentheses) for the prevalence of cheating on the anagram task

Mode	$\hat{\pi}$	Test against "true" prevalence of 15.51%		Test against CWM prevalence estimate	
		ΔG^2 (df=1)	p	ΔG^2 (df=1)	p
DQ (n = 138)	5.07% (1.87)	14.88	< .001	5.29	< .05
CWM (n = 526)	13.03% (2.75)	0.79	.37	–	

Note: $\hat{\pi}$ = Prevalence of cheating as estimated using a direct question and the Crosswise Model.

Appendix A

All 45 anagrams that were presented on the pretest.

Item	German Target word	Anagram	English Translation	Frequency class	Letter Sequence	Moves required	Number of letters in correct order	τ (letter sequences of target word and anagram)
Easy items (envisaged difficulty: p > .99)								
1	KIRCHE	CKIRHE	church	7	412356	1	5	0.60
2	MONTAG	NMOTAG	Monday	7	312456	1	5	0.73
3	HERBST	HERSTB	autumn	8	123564	1	5	0.73
4	GEFAHR	GERFAH	danger	8	126345	1	5	0.60
5	MENSCH	MENHSC	human	8	123645	1	5	0.60
6	PRAXIS	APRXIS	practice	9	312456	1	5	0.73
7	URTEIL	URLTEI	verdict	9	126345	1	5	0.60
8	SCHULD	USCHLD	guilt	9	412356	1	5	0.60
9	HEIMAT	EIMHAT	home	9	234156	1	5	0.60
10	SCHUTZ	SCHTZU	protection	9	123564	1	5	0.73
11	DIENST	IEDNST	service	9	231456	1	5	0.73
12	WUNSCH	NWUSCH	wish	9	312456	1	5	0.73
13	FREUND	UFREND	friend	9	412356	1	5	0.60
14	JUGEND	JUGNDE	youth	9	123564	1	5	0.73
15	UMWELT	UMELTW	environment	9	124563	1	5	0.60
16	TERMIN	ERTMIN	appointment	10	231456	1	5	0.73
17	UMFANG	UMFGAN	coverage	10	123645	1	5	0.60
18	SCHLAG	SCLAGH	beat	10	124563	1	5	0.60
19	ABWEHR	ABWREH	defense	10	123645	1	5	0.60
20	BEZIRK	BEZKIR	district	10	123645	1	5	0.60
21	FLUCHT	CFLUHT	escape	10	412356	1	5	0.60
22	JUSTIZ	SJUTIZ	justice	10	312456	1	5	0.73
23	INHALT	INHLTA	content	10	123564	1	5	0.73
24	GIPFEL	PGIFEL	peak	10	312456	1	5	0.73
25	BESITZ	BESTZI	property	10	123564	1	5	0.73
26	VERBOT	ERBVOT	prohibition	10	234156	1	5	0.60
27	STRICH	STRHIC	line	10	123645	1	5	0.60
28	SCHILD	ISCHLD	shield	11	412356	1	5	0.60
29	OBJEKT	OBJTEK	object	11	123645	1	5	0.60
30	GEHALT	EHGALT	salary	11	231456	1	5	0.73
31	SYMBOL	BSYMOL	symbol	11	412356	1	5	0.60
32	EXPORT	EXTPOR	export	11	126345	1	5	0.60

Difficult items (envisaged difficulty: p < .01)								
33	URINAL	IRLAUN	urinal	18	326514	4	0	-0.07
34	ALBINO	LOANIB	albino	18	261543	4	0	-0.07
35	GLUCKE	CUGEKL	hen	18	431652	4	0	-0.07
36	FUNZEL	ZULFEN	dim light	18	426153	4	0	-0.07
37	EILZUG	LIGEUZ	fast train	19	326154	4	0	0.07
38	BARIUM	IRAMBU	barium	19	432615	4	0	-0.07
39	KARIBU	RAUKBI	caribou	19	326154	4	0	0.07
40	WOMBAT	AOWTBM	wombat	19	521643	4	0	-0.07
41	METRIK	TEKIMR	metric	20	326514	4	0	-0.07
42	TALKUM	LTMUKA	talcum	20	316542	4	0	-0.07
43	AZIMUT	MIATUZ	azimuth	20	431652	4	0	-0.07
44	NUKLID	IKUNDL	nuclide	22	532164	4	0	-0.07
45	TRIOLE	IERTLO	triplet	22	362154	4	0	-0.07

Appendix B

Multinomial model equations used for parameter estimation in multiTree (Moshagen, 2010):

CWM	CWM_bothtrueorfalse	PiCWM*p
CWM	CWM_onetrue	PiCWM*(1-p)
CWM	CWM_onetrue	(1-PiCWM)*p
CWM	CWM_bothtrueorfalse	(1-PiCWM)*(1-p)
DQ	DQ_true	PiDQ
DQ	DQ_false	(1-PiDQ)

Appendix C

Observed answering frequencies used for parameter estimation in multiTree (Moshagen, 2010):

CWM_bothtrueorfalse	396
CWM_onetrue	130
DQ_true	7
DQ_false	131

Versicherung an Eides Statt

Hiermit versichere ich an Eides Statt, dass die Dissertation mit dem Titel

Indirekte Befragungstechniken zur Kontrolle sozialer Erwünschtheit in Umfragen

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,

Adrian Hoffmann