Experimentelle Umfrageforschung mit der Randomized-Response-Technik

Inaugural-Dissertation

zur

Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Martin Stefan Ostapczuk

aus Breslau

April 2008

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

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

Referent: Prof. Dr. Jochen Musch

Koreferentin: Prof. Ute J. Bayen, Ph.D.

Tag der mündlichen Prüfung: 25.04.2008

„Wissenschaft ist als Erkenntnis verschwunden, wenn sie in Resultaten erstarrt.“

Karl Jaspers (1883-1969)

Danksagung

An erster Stelle möchte ich mich bei den vielen Patienten, Internetsurfern, Schülern und Studenten für die Teilnahme an den hier vorliegenden Untersuchungen bedanken.

Meinem Betreuer Prof. Jochen Musch danke ich *allerherzlichst* für die Gelegenheit zur Promotion über dieses Thema sowie für die uneingeschränkte und bereitwillige Zurverfügungstellung aller nur erdenklichen Ressourcen und Hilfen sowohl hinsichtlich der Promotion als auch hinsichtlich anderer (weniger psychologischer) Bereiche. Es ist nicht übertrieben zu behaupten, dass ohne ihn diese Promotion nicht möglich gewesen wäre. Prof. Ute Bayen danke ich für die Übernahme der Zweitbegutachtung trotz extremer (zeitlicher) Rahmenbedingungen.

Bei meinen Kollegen Morten Moshagen, Dr. Zengmei Zhao und Michael Wolf bedanke ich mich für zahlreiche Hilfestellungen bei der Datenerhebung sowie insbesondere bei Morten für das Korrekturlesen aller Artikel und die vielen fruchtbaren inhaltlichen Diskussionen. Unseren (Ex-)Hilfskräften Helen-Rose Cleveland, Anna Fligg und Sonja Willing danke ich für die Eingabe unzähliger Fragebögen. Dennis Winter habe ich die Programmierung von Experiment III zu verdanken und Sebastian Ullrich die Korrektur des einen oder anderen randomized Entwurfes. Unserer Sekretärin Sabine Hillebrandt danke ich für die stets reibungslose Organisation „im Hintergrund“.

Für weitere Unterstützung bei der extensiven Datenerhebung zu den Experimenten I und II möchte ich Barbara und Dr. Thilo Moshagen, Birgit und Dr. Klaus Scholz, Monika Undorf sowie meinen Eltern Drs. Anna-Maria und Stefan Ostapczuk danken.

Besonders freue ich mich, mich auch wieder bei meinem aus den Diplomarbeitszeiten bewährten Korrekturleseteam aus Christiane Federlin und Nicole Vahsen bedanken zu dürfen, die auch dieses Werk kritisch gelesen haben. Meiner Familie danke ich für die uneingeschränkte Unterstützung in allen Lebenslagen der letzten Jahre und insbesondere der letzten Monate.

Inhaltsverzeichnis

Zusammenfassung

Der Selbstbericht stellt bei vielen sozialwissenschaftlichen Fragestellungen eine wichtige und häufig die einzige Datenquelle dar. Ist das Thema der Befragung jedoch sensibel, muss man damit rechnen, dass einige Befragte – beispielsweise im Bestreben, sozial erwünschte Antworten zu geben – beschönigend antworten. Die Validität der Daten wird dadurch bedroht und ihre Interpretierbarkeit eingeschränkt.

Die *Randomized-Response-Technik* (RRT; Warner, 1965) wurde entwickelt, um dieses Problem zu lösen. Bei Anwendung des Forced-Response-Modells der RRT entscheidet ein Zufallsgenerator, ob der Befragte gebeten wird, ehrlich auf die kritische Frage zu antworten, oder ob er gebeten wird, unabhängig vom Frageninhalt das Vorhandensein des sensiblen Merkmals zu bejahen. Weil der Ausgang des Zufallsexperiments dem Fragesteller nicht bekannt ist, kann aus dem Antwortverhalten nicht auf den wahren Merkmalsstatus geschlossen werden. Dadurch fördert das Verfahren die Bereitschaft, auch sensible Fragen ehrlich zu beantworten. Bei bekannter Verteilung des Zufallsgenerators ist auf Gruppenebene eine Schätzung der Prävalenz des sensiblen Merkmals bei gleichzeitiger Wahrung der Vertraulichkeit individueller Antworten möglich. Obwohl die Anonymität dadurch erhöht wird, kann es vorkommen, dass sich manche Teilnehmer nicht an die RRT-Regeln halten und trotz der Aufforderung, inhaltsunabhängig mit „Ja“ zu antworten, das Vorhandensein des kritischen Merkmals abstreiten. In diesem Fall unterschätzt auch die RRT die Prävalenz des kritischen Merkmals, soweit es sich bei den Regelverweigerern um Merkmalsträger handelt. Mit Hilfe einer Erweiterung des Forced-Response-Modells kann versucht werden, durch die Verwendung einer unabhängigen zweiten Stichprobe, in der eine andere Randomisierungswahrscheinlichkeit verwendet wird, den Anteil der Verweigerer zu schätzen (Clark & Desharnais, 1998).

In der vorliegenden Dissertation wurde in vier experimentellen Umfragen, die in unterschiedlichen Erhebungskontexten durchgeführt wurden, ein multinomiales Modell der Verweigererdetektionsvariante der RRT von Musch, Bröder und Klauer (2001) validiert und erweitert. *Experiment I* zeigte in einer Papier-und-Bleistift-Untersuchung zur Non-Compliance bei der Medikamenteneinnahme, dass auf der Basis des multi-

nomialen Verweigererdetektionsmodells validere Prävalenzschätzungen für sozial unerwünschtes Verhalten als in einer direkten Befragung erzielt werden können. Die mit Hilfe der Verweigererdetektionsvariante der RRT geschätzte Lebenszeitprävalenz von Non-Compliance lag mit 33% deutlich über der Prävalenzschätzung der direkten Befragung (21%). Darüber hinaus zeigten die Ergebnisse von Experiment I, dass bei einem Verzicht auf die experimentelle Erweiterung des Versuchsdesigns zur Verweigererdetektion unbemerkt geblieben wäre, dass sich fast die Hälfte (47%) der unter RRT-Bedingungen befragten Teilnehmer nicht an die Regeln der Technik gehalten hat; auf der Basis des Modells konnte jedoch unter Berücksichtigung der Verweigererrate eine obere Schranke für die wahre Prävalenz der Non-Compliance bestimmt werden, welche noch einmal erheblich über der Prävalenzrate lag, welche mit der RRT geschätzt worden war.

Experiment II nutzte die vom multinomialen Modellierungsansatz gebotene Möglichkeit zur Prüfung auf Parametergleichheit in Subgruppen, um einen inhaltlich bedeutsamen Gruppenunterschied auf seine Gültigkeit hin zu untersuchen. Frühere Untersuchungen haben gezeigt, dass sich hinsichtlich ausländerfeindlicher Einstellungen ein deutlicher Bildungseffekt zeigt: Personen mit geringer Bildung geben bei Selbstauskünften regelmäßig negativere Einstellungen gegenüber Ausländern an als Personen mit höherer Bildung. Bei den bisherigen Studien konnte jedoch nicht ausgeschlossen werden, dass es sich bei diesem Bildungseffekt nur um ein Artefakt einer stärkeren Tendenz zur sozial erwünschten Antwort bei den Personen mit höherer Bildung handelt. In der vorliegenden Untersuchung zeigte ein Vergleich der direkten Befragungsergebnisse mit den mit Hilfe der der RRT geschätzten Prävalenzen, dass hoch gebildete Befragte nicht nur bei einer konventionellen Befragung, sondern auch unter dem Schutz der Zufallsverschlüsselung weniger ausländerfeindlich und vor allem ausländerfreundlicher als niedrige gebildete Befragte antworten. Vervollständigt durch die Betrachtung der verschiedenen oberen Schranken für ausländerfeindliche und ausländerfreundliche Einstellungen war dieses Ergebnismuster besser mit der Interpretation des Bildungseffektes im Sinne eines wahren Gruppenunterschiedes als eines Artefaktes vereinbar.

In *Experiment III* wurde das multinomiale Verweigererdetektionsmodell nochmals erweitert, um die Validität eines konkurrierenden Verfahrens zur Reduktion von

Antwortverzerrungen zu überprüfen. Untersucht wurde, ob die projektive Most-People-Technik (MPT; Alpert, 1971; Smith, 1954) zur Überschätzung der Prävalenz sensibler Merkmale führen kann. In einer im WWW durchgeführten Untersuchung konnte gezeigt werden, dass die Prävalenzschätzung der MPT nicht nur über derjenigen der direkten Befragung und der RRT lag, sondern auch über der mittels RRT bestimmten oberen Schranke für die Prävalenz negativer Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung. Die Untersuchung zeigte damit, dass die projektive MPT die Prävalenz sensibler Merkmale überschätzen kann. Dies stellt ihre Verwendbarkeit als Methode zur Reduktion von Antwortverzerrungen in Frage und unterstreicht die Nützlichkeit der Verweigererdetektion im Rahmen der Randomized-Response-Technik.

Experiment IV wurde wieder als Papier-und-Bleistift-Studie durchgeführt. An einer Stichprobe von chinesischen Studenten wurde überprüft, ob die Verweigererdetektionsvariante verbessert werden kann, indem die dem Modell inhärente Asymmetrie zwischen bejahenden und verneinenden Antworten aufgehoben wird. Im einfachen Verweigererdetektionsmodell bringt eine vom Zufallsgenerator erzwungene „Ja"-Antwort auch Nichtmerkmalsträger in den Verdacht der Merkmalsträgerschaft; dies ist notwendig, damit eine „Ja"-Antwort nicht länger den wahren Merkmalsstatus offenbart und so auch Merkmalsträger zu ehrlichen Antworten ermutigt werden. Durch eine „Nein"-Antwort kann der Befragte den Verdacht, Merkmalsträger zu sein, jedoch von vornherein von sich weisen. Um diese Asymmetrie zu vermeiden und dem von ihr ausgehenden Anreiz zur Nichtbefolgung der RRT-Instruktionen entgegenzuwirken, wurden in Experiment IV in neuerlicher Erweiterung des Befragungsmodells auch „Nein"-Antworten vom Zufallsgenerator erzwungen. Es zeigte sich, dass dadurch die Verweigererrate wirksam reduziert werden kann. Die obere Schranke für die Prävalenz des kritischen Merkmals kann auf diese Weise gesenkt und die Aussagekraft experimenteller RRT-Umfragen erhöht werden.

Zusammenfassend legen die Ergebnisse der Experimente I bis IV nahe, dass die auf einer Zufallsverschlüsselung der Antworten beruhende Verweigererdetektionsvariante der Randomized-Response-Technik Antwortverzerrungen wirksam und besser als konkurrierende Methoden zu kontrollieren vermag. Darüber hinaus konnte gezeigt werden, dass und wie zufallsverschlüsselte Umfragen im Rahmen eines multinomialen

Modellierungsansatzes flexibel an neue Erhebungssituationen angepasst und erfolgreich für die Untersuchung weiterführender, inhaltlicher wie methodischer Fragestellungen genutzt werden können. Der experimentellen Umfrageforschung steht damit ein wirksames und erfolgreich validiertes Instrument zur Modellierung und Kontrolle von Antwortverzerrungen bei Selbstauskünften zur Verfügung.

Summary

Self-report data on sensitive topics are often biased due to social desirability. The *randomized-response-technique* (RRT; Warner, 1965) attempts to reduce social desirability bias by randomizing the interviewees' answers in order to increase the validity of prevalence estimates of sensitive behaviours and attitudes. Using a randomization device ensures that an individual interviewee can no longer be associated with the sensitive attribute. Knowing the probability distribution of the randomization device, the proportion of affirmative responses that have not been prompted by the randomization device can be estimated at the group level. The technique therefore guarantees the confidentiality of responses and arguably encourages more honest responding.

Nevertheless, some interviewees may decide to cheat by disregarding the RRT instructions. To the extent that cheating respondents are in fact holding the sensitive attribute, the RRT underestimates the prevalence of sensitive attributes, too. An RRT cheating detection model by Clark and Desharnais (1998), however, allows for the assessment of the proportion of cheaters in a sample and thereby for controlling potential response bias. In the present dissertation, I successfully tested, validated and improved a *multinomial model of cheating detection* (Musch, Klauer & Bröder, 2001) in four experimental surveys.

Experiment I, a paper-and-pencil-study on medication non-adherence, showed that using the cheating detection model can improve the validity of prevalence estimates of socially undesirable behaviours as compared with direct questioning. The estimate of lifetime medication non-adherence obtained by the cheating detection model of RRT was significantly higher (33%) than the corresponding prevalence estimate obtained when questioned directly (21%). Moreover, the results of experiment I underscore the utility of the cheating detection extension of RRT, since almost half of the participants (47%) disregarded the RRT rules; this significant proportion of cheaters would have gone unnoticed if conventional self-report measures or traditional RRT-variants not considering cheating had been used instead of the cheating detection model.

In *Experiment II*, an extension of the multinomial modelling approach was used in order to test whether the frequently reported education effect in attitudes towards foreigners might be due to an artefact. Previous studies have not been able to exclude the assumption that highly educated interviewees are as xenophobic as less educated interviewees, but simply more receptive to the sensitive nature of the inquiry thus biasing their responses in surveys on xenophobia towards the socially desirable, i.e., the xenophile, alternative. This alternative interpretation could be rejected by comparing the RRT estimates with the direct questioning estimates. Even after controlling for potential response bias, highly educated interviewees gave more xenophile responses than less educated interviewees. This result suggests that the education effect is not due to a group specific, differential tendency to distort answers.

In *Experiment III*, the multinomial cheating detection model was extended once more to test the validity of an alternative measure attempting to reduce social desirability bias in surveys on sensitive topics. Specifically, I tested the assumption whether the projective most-people-technique (MPT; Alpert, 1971; Smith, 1954) overestimates the prevalence of negative attitudes towards people with disabilities. The online-study showed that the prevalence estimates obtained by the MPT were not only significantly higher than the corresponding direct questioning estimates, but also exceeded an upper bound of the prevalence estimates determined by the cheating detection variant of the RRT. These results severely question the validity of the MPT and demonstrate the superiority of the alternative RRT.

Experiment IV was conducted as a paper-and-pencil study again. In a sample of Chinese students, I tested whether the cheating detection model can be improved by eliminating the asymmetry between "yes"- and "no"-responses inherent to the original model. In order to reduce the appeal of cheating, this asymmetry was avoided by a final extension of the cheating detection model. The results demonstrated the effectiveness of this strategy: the proportion of cheaters was successfully reduced to a minimum, which no longer differed significantly from zero.

1 Einleitung

Wie Fox und Tracy (1986) anmerken, existieren für viele – wenn nicht die meisten – sozialwissenschaftlichen Fragestellungen keine Archivdaten, auf die zu ihrer Beantwortung unmittelbar zurückgegriffen werden könnte. Dem Forscher bleibt demnach häufig nichts anderes übrig, als selbst Daten zu erheben. Dabei ist die Selbstauskunft von Befragten aus Gründen der Erhebungsökonomie immer noch die am häufigsten und oft sogar ausschließlich verwendete Datenquelle. Sich auf Selbstauskünfte zu verlassen, ist jedoch gerade bei der Untersuchung sensibler Merkmale (z.B. Drogenmissbrauch, Abtreibung, Fahrerflucht, ausländerfeindliche Einstellungen usw.) problematisch: Eine Vielzahl von Gründen (wie beispielsweise Angst vor Prestigeverlust oder Strafe, soziale Sanktionierung, Unerwünschtheit und Schamgefühle) können dazu führen, dass die Befragten ihre Antworten verzerren, indem sie entweder gar nicht (*Refusal / Nonresponse Bias*) oder nicht ehrlich bzw. beschönigend (*Response Bias*) antworten (Fox & Tracy, 1986; Lensvelt-Mulders, Hox, van der Heijden & Maas, 2005; Scheers, 1992). Dass Antwortverzerrungen tatsächlich auftreten, konnte empirisch vielfach gezeigt werden (z.B. Edwards, 1957; Hyman, 1944; Lee, 1993; Maccoby & Maccoby, 1954; Zerbe & Paulhus, 1987). Die Folge solch eines Antwortverhaltens bei klassischen Befragungen liegt auf der Hand: Das Messergebnis ist wenig valide, die Prävalenz des sensiblen Merkmals wird unterschätzt und es ist bestenfalls die Schätzung einer Untergrenze der wahren Prävalenz möglich (Musch & Plessner, eingereicht).

Die naheliegendste Möglichkeit, dem Problem mangelnder Bereitschaft zur ehrlichen Antwort zu begegnen, ist die direkte Überprüfung der wahren Merkmalsausprägung des Befragten. Eine beispielsweise mit Hilfe einer Blut- oder Haarprobe getroffene Feststellung, ob der Befragte Drogenkonsument ist, könnte bei Verwendung hinreichend sensitiver und spezifischer Indikatoren sehr valide Daten liefern und würde eine fehlerbehaftete Selbstauskunft unnötig machen. Da dies jedoch häufig aus praktischen, ethischen oder rechtlichen Gründen nicht möglich ist, sind in den vergangenen Jahrzehnten verschiedene Techniken entwickelt und diskutiert worden, um Antwortverzerrungen auch ohne Verzicht auf die Selbstauskunft in den Griff zu

bekommen. Grob lassen sich die verschiedenen Verfahren in drei Klassen einteilen (vgl. Nederhof, 1985): (i) Techniken zur Sichtbarmachung der Tendenz, die Antwort zu verzerren (z.B. Verfahren zur Messung sozialer Erwünschtheit), (ii) Techniken, die versuchen, die Verzerrung durch Erhöhung des Drucks auf den Befragten[1] zu reduzieren (z.B. Lügendetektion), (iii) Techniken, die darauf abzielen, die Verzerrung durch Erhöhung der Anonymität zu reduzieren, ohne dabei die klassische Befragungsart grundlegend zu ändern (z.B. Versiegelung der Umschläge), sowie Techniken, die ebenfalls mit erhöhter Anonymität arbeiten, gleichzeitig aber die Modalitäten der Befragung verändern (z.B. die Randomized-Response-Technik).[2]

In der vorliegenden Dissertation habe ich mich mit der Frage beschäftigt, ob eine multinomial modellierte Verweigererdetektionsvariante der *Randomized-Response-Technik* (RRT; Clark & Desharnais, 1998; Musch, Bröder und Klauer, 2001; Warner, 1965) ihrem Anspruch gerecht wird, validere Prävalenzschätzungen sensibler Merkmale als andere Methoden zu ermöglichen. Weiterhin habe ich geprüft, ob und wie das Verfahren auf neue Befragungskontexte adaptiert und hinsichtlich seiner Methodik verbessert werden kann.

Im Folgenden werden zunächst die oben genannten Techniken zur Erhöhung der Validität von Selbstauskünften bei sensiblen Fragen näher beschrieben (Kapitel 2). Das daran anschließende Kapitel 3 widmet sich der Vorstellung der Verweigererdetektionsvariante der RRT, die sämtlichen hier vorgestellten Einzelarbeiten zugrundeliegt. In Kapitel 4 wird die Fragestellung für die vier Einzelarbeiten hergeleitet, die im folgenden Kapitel 5 zusammengefasst werden. Kapitel 6 diskutiert die Ergebnisse der Arbeiten, gefolgt von einem Ausblick auf zukünftige Forschungsfragen (Kapitel 7). Die Arbeit schließt mit fünf zusammenfassenden Thesen (Kapitel 8) sowie den angehängten Originalarbeiten.

---

[1] Im Folgenden wird zwecks besserer Lesbarkeit nur die männliche Form verwendet. Wenn nicht anders ausgewiesen, ist auch immer die weibliche Form mit eingeschlossen.

[2] Als eine weitere Möglichkeit zur Reduktion von Antwortverzerrungen wird zuweilen die Fremdauskunft genannt (z.B. Nederhof, 1985). Diese wird hier jedoch aus zwei Gründen nicht weiter behandelt: Erstens weil es sich bei der Befragung Dritter definitionsgemäß um keine Methode zur Reduktion der Antwortverzerrungen von *Selbst*auskünften handelt, und zweitens weil ein überzeugender Nachweis, dass Dritte validere Auskünfte als die Befragten selbst erteilen, bislang aussteht.

2 Techniken zur Erhöhung der Validität von Selbstauskünften

In den folgenden Unterkapiteln (2.1-2.3) werden drei große Klassen von Ansätzen zum Umgang mit Antwortverzerrungen bei Selbstauskünften beschrieben und evaluiert. Das Kapitel schließt mit einer ausführlicheren Beschreibung der Randomized-Response-Technik, dem Thema der vorliegenden Arbeit.

2.1 *Ansatz I*: Messung sozialer Erwünschtheit

Skalen zur Erfassung der *Tendenz, sozial erwünscht* zu antworten wie die deutsche Version des Balanced Inventory of Socially Desirable Responding (BIDR-D; Musch, Brockhaus & Bröder, 2002; Stöber, Dette & Musch, 2002) oder die Soziale-Erwünschtheits-Skala-17 (SES-17; Stöber, 1999) gehören zu den ältesten Techniken zur Erhöhung der Validität von Selbstauskünften. Früher wurden sie häufig als Lügen- oder Kontrollskalen bezeichnet. Ihr Vorteil liegt in der vergleichsweise einfachen Handhabung und voraussetzungsfreien Einsetzbarkeit. Ihr großer Nachteil liegt jedoch darin, dass sie keine tatsächliche Möglichkeit zur Kontrolle systematischer Antwortverzerrungen erlauben: Fehlende oder unehrliche Antworten werden nicht vermieden; mit Hilfe der Skalen können allenfalls interindividuelle Unterschiede in der Tendenz zur positiven Selbstdarstellung erfasst und bei der Betrachtung des interessierenden Merkmals berücksichtigt werden. Doch selbst dieses Vorgehen erweist sich als problematisch: Die Herauspartialisierung der Selbstdarstellungstendenz aus der Selbstauskunft führt nämlich oft nicht zu den erhofften Validitätsverbesserungen (McCrae & Costa, 1983), unter anderem wohl auch weil soziale Erwünschtheitsskalen selbst nicht vor intentionaler Verfälschung gefeit sind (Pauls & Crost, 2004). Zudem bringt der Ausschluss von Probanden mit „zu hohen Werten“ auf der Erwünschtheitsskala die Probleme eines schwer begründbaren und daher letztlich willkürlichen Cut-Off-Wertes für den Ausschluss sowie einer durch den Ausschluss unter Umständen reduzierten Repräsentativität der Stichprobe mit sich (Nederhof, 1985).

Analog zur Messung der sozialen Erwünschtheit auf Seiten der Befragten kann man auch das interessierende Merkmal selbst von den Befragten hinsichtlich seiner sozialen Erwünschtheit einschätzen lassen (Mummendey, 1987; Nederhof, 1985) und diese Einschätzung bei der eigentlichen Auswertung berücksichtigen. Leider ist dieser Ansatz mit ähnlichen Problemen behaftet wie die zuvor beschriebene Erfassung und Berücksichtigung der Selbstdarstellungstendenz der Befragten und daher ähnlich unfruchtbar; eine Möglichkeit, das Ausmaß von Antwortverzerrungen zu quantifizieren und Selbstauskünfte entsprechend zu korrigieren, bieten beide Ansätze nicht.

2.2 *Ansatz II*: Erhöhung des Drucks auf den Befragten

Anstatt soziale Erwünschtheit „nur" zu messen, kann man versuchen, die durch sie bedingten Verzerrungen zu reduzieren, indem man den Druck auf den Befragten erhöht. Dies kann beispielsweise mit Hilfe der psychophysiologischen Lügendetektion oder der Bogus-Pipeline-Technik geschehen.

2.2.1 Psychophysiologische Lügendetektion

Eine zumindest auf den ersten Blick recht vielversprechende Methode im Hinblick auf die Reduktion von sozial erwünschtem Antwortverhalten bietet die *psychophysiologische Lügendetektion* (Fiedler, Schmidt & Stahl, 2002; Iacono, 2000). Die grundlegende Idee bei der Lügendetektion ist, dass sich die psychophysiologischen Reaktionen von Menschen, die lügen und nicht lügen, unterscheiden lassen. Geeignete psychophysiologische Indikatoren sind beispielsweise die elektrodermale Aktivität, die Atmung oder der Blutdruck. Der *Control-Question-Test* (CQT) und der *Guilty-Knowledge-Test* (GKT) stellen die zwei am häufigsten verwendeten Varianten dar. Beim CQT werden dem Befragten sowohl Fragen zur eigentlichen Tat gestellt (z.B. „Haben Sie schon einmal illegale Drogen eingenommen?") als auch Kontrollfragen, die ebenfalls eine emotionale Reaktion hervorrufen sollen (z.B. „Haben Sie schon einmal so viel

Alkohol getrunken, dass Sie sich am nächsten Tag nicht an die Geschehnisse erinnern konnten?“). Bei den schuldigen Befragten wird eine stärkere psychophysiologische Reaktion auf die tatrelevanten als auf die Kontrollfragen erwartet. Bei unschuldigen Befragten erwartet man dagegen, dass sie keine allzu starke psychophysiologische Reaktion auf die tatrelevanten Fragen im Gegensatz zu einer starken Reaktion auf die emotional besetzten Kontrollfragen zeigen sollten. Der größte Kritikpunkt am CQT zielt auf seine bisher unbelegte Grundannahme, dass nur schuldige Befragte stärker auf die tatrelevanten als auf die Kontrollfragen reagieren. Der GKT vermeidet diese kritische Annahme und geht stattdessen davon aus, dass schuldige Befragte eine stärkere psychophysiologische Reaktion als unschuldige Befragte zeigen, wenn sie mit Informationen konfrontiert werden, die nur der Täter kennen kann. Bezogen auf das Drogenbeispiel könnte eine solche Frage lauten: „Welche von den folgenden illegalen Drogen haben Sie schon einmal eingenommen? a) Marihuana, b) Heroin, c) Kokain.“ Angenommen, der schuldige Befragte habe Heroin konsumiert, so sollte seine psychophysiologische Reaktion bei b) höher ausfallen als bei a) und c). Der unschuldige Befragte sollte hingegen auf alle drei Antwortalternativen etwa gleich stark reagieren. Doch auch der GKT ist kritisiert worden: So ist das zur Konstruktion der benötigten Items notwendige Wissen über den genauen Tathergang auf Seiten der Ermittler oft gar nicht vorhanden. Zudem ist tatrelevantes Wissen oft bereits an die Öffentlichkeit gelangt, so dass auch Unschuldige darüber verfügen können (Erdfelder & Musch, 2006).

2.2.2 Bogus-Pipeline-Technik

Die *Bogus-Pipeline-Technik* (Jones & Sigall, 1971; Mummendey, Bolten & Isermann-Gerke, 1982) versucht die Probleme psychophysiologischer Lügendetektion zu umgehen, indem sie sich lediglich den weit verbreiteten Glauben an die Validität der Lügendetektion zu Nutze macht: Der Befragte glaubt, er sei an ein psychophysiologisches Messgerät angeschlossen, welches im Stande ist, wahre von unwahren Antworten zu unterscheiden. Er fühlt sich somit ebenfalls unter Druck gesetzt, ehrlich zu antworten. Ein spezielles Problem dieser Technik besteht darin, dem Befragten glaubhaft zu machen, dass das Gerät tatsächlich in der Lage ist, Lügen zu erfassen.

Außerdem weist die Bogus-Pipeline-Technik zwei generelle Probleme auf, die auch für die tatsächliche Lügendetektion gelten: Der Einsatz beider Methoden ist mit einem großen apparativen Aufwand verbunden und unterliegt darüber hinaus oft einer Reihe ethischer oder rechtlicher Bedenken.

2.3 *Ansatz III*: Erhöhung der Anonymität

Wahrscheinlich das beste Mittel, Befragte zu ehrlicheren Antworten zu motivieren, ist die Herstellung von *Anonymität* (Fisher, 1993; Ong & Weiss, 2000). Mit konventionellen Methoden sind in diesem Zusammenhang Ansätze gemeint, die die Anonymität der Befragungssituation erhöhen, ohne die Befragungsmodalitäten entscheidend zu verändern. Im Gegensatz dazu wird die Anonymität bei Ansätzen wie projektiven bzw. nominativen Techniken, der Unmatched-Count- sowie der Randomized-Response-Technik, erhöht, indem die Art der Fragen oder Antworten verändert wird.

2.3.1 Konventionelle Techniken

Die Verwendung von nicht namentlich gekennzeichneten Fragebögen, die gesammelte Rückgabe von Fragebögen in einer verschlossenen Schachtel bzw. in versiegelten Umschlägen sowie die Verringerung des persönlichen Kontaktes durch telefonische, postalische oder computergestützte Befragung verändern weder die Art und Weise, wie die sensiblen Fragen gestellt werden (wie bei der Lügendetektion bzw. der Bogus-Pipeline-Technik und den projektiven Techniken, 2.3.2), noch die Art und Weise, wie sie zu beantworten sind (wie bei der Unmatched-Count- oder Randomized-Response-Technik, 2.3.3-2.3.4). Auch müssen die Antworten nicht nachträglich korrigiert oder Probanden ausgeschlossen werden (wie bei der Erfassung sozialer Erwünschtheit). Dennoch haben sich all diese Vorkehrungen zur Erhöhung der Anonymität als

nützliche Methoden erwiesen, Befragte zu ehrlicheren Antworten zu ermuntern (Fisher, 1993; Nederhof, 1985).

Entgegen ursprünglichen Hoffnungen scheint durch Maßnahmen zur Anonymisierung jedoch nur eine graduelle Verbesserung der Bereitschaft zu ehrlichem Antworten erreichbar zu sein, die zudem – ohne den Vergleich mit noch valideren Methoden – hinsichtlich ihres Ausmaßes nicht quantifiziert werden kann (van der Heijden, van Gils, Bouts & Hox, 2000). Abgesehen davon haben einige Autoren darauf hingewiesen, dass zu intensive Hinweise auf die Gewährleistung von Anonymität sogar das Gegenteil bewirken können, nämlich eine Zunahme von sozial erwünschtem Antwortverhalten, wenn Befragte dadurch erst auf die Sensitivität des Themas aufmerksam gemacht und folglich misstrauisch werden (Reamer, 1979; Singer, Hippler & Schwarz, 1992).

2.3.2 Projektive und nominative Techniken

Der psychoanalytischen Theorie zufolge handelt es sich bei der Projektion, dem Namensgeber für die so genannten *projektiven Techniken* bei der Umfrageerstellung, um einen Abwehrmechanismus des Ichs (Freud, 1938): Werden Individuen mit Angst auslösenden und daher wenig wünschenswerten Impulsen, Gefühlen oder Einstellungen konfrontiert, entsteht unangenehme Spannung. Um diese zu reduzieren, attribuiert das Individuum die Gefühle oder Einstellungen unbewusst auf die äußere Welt. Im Kontext der Umfrageforschung versuchen projektive Verfahren dies auszunutzen, indem sie dem Befragten Konflikt auslösende Stimuli darbieten, die ihnen eine Projektion der oben genannten Gefühle oder Einstellungen auf andere ermöglichen. Die *Most-People-Technik* stellt eine strukturierte projektive Methode dar (Alpert, 1971; Smith, 1954): Anstatt ihre eigene Meinung kundzutun, werden Befragte darum gebeten, anzugeben, was ihrer Ansicht nach die meisten Menschen auf eine bestimmte sensible Frage antworten würden (z.B. „Glauben Sie, dass die meisten Menschen schon einmal illegale Drogen eingenommen haben?“). Aus der Antwort wird dann jedoch – der psychoanalytischen Vorstellung des Projektionsvorgangs folgend – auf die eigene Einstellung der Befragten geschlossen. Damit bietet die Methode völlige

Anonymität, da die Befragten glauben, dass sie gar nichts von sich preisgeben, und sich dadurch sicher fühlen sollten. Offensichtlich sind die theoretischen Annahmen des Verfahrens jedoch durchaus problematisch und der Verdacht liegt nahe, dass die Most-People-Technik zu Prävalenzüberschätzungen führen kann (Bégin & Boivin, 1980).

Nominative Techniken (Miller, 1985; Bradburn & Sudman, 1979) sind nach einem ähnlichen Prinzip aufgebaut, kommen jedoch ohne einige der kritischen Annahmen projektiver Verfahren aus: Die Befragten sollen angeben, ob ein bestimmtes Verhalten oder eine Einstellung in ihrem Freundeskreis auftritt oder nicht. Der Unterschied zur projektiven Most-People-Technik besteht darin, dass hier nicht abstrakt nach den *meisten*, sondern nach *konkreten* Menschen gefragt und überdies nicht davon ausgegangen wird, dass die Befragten ihr eigenes Verhalten auf andere projizieren; sie sollen vielmehr über das tatsächliche Verhalten ihrer Freunde berichten. Damit bleiben sowohl die Befragten selbst als auch die Freunde vollständig anonym, da keine Namen genannt werden müssen. Auch nominative Techniken kommen jedoch nicht ohne die problematisierbare Annahme aus, dass man über das Verhalten und die Einstellungen seiner Freunde gut genug informiert ist, um zuverlässige Informationen liefern zu können.

2.3.3 Unmatched-Count-Technik

Eine Methode, bei der versucht wird, durch die experimentelle Herstellung von Anonymität mehr ehrliche Antworten auf sensible Fragen zu erhalten, hat Miller (1984) entwickelt. Im Rahmen ihrer *Unmatched-Count-Technik* (auch Randomized-List-Technik genannt) werden die Befragten nicht nur nach dem kritischen Merkmal, sondern zusätzlich auch nach einer Reihe harmloser Merkmale gefragt. Dabei werden sie jedoch gebeten, nicht jede Einzelfrage zu beantworten, sondern lediglich über alle Fragen hinweg die Summe der „Ja"-Antworten (z.B. für die Fragen A + B + C + D + E + F) zu bilden. Diese Summe (*S1*) anschließend zu berichten, sollte den Umfrageteilnehmern leicht fallen, denn sie sagt – bei geeigneter Wahl der harmlosen Fragen A, B, C, D und E – nichts über ihre Antwort auf die kritische Frage F aus. Nach Zufall kann jedoch eine Hälfte der Umfrageteilnehmer einer zweiten Versuchs-

bedingung zugewiesen werden, in der wieder nach der Summe der „Ja"-Antworten, diesmal aber nur für die harmlosen Fragen (*S2* = A + B + C + D + E) gefragt wird. In der Fragenliste der Kontrollbedingung fehlt also die kritische Frage nach dem sensiblen Merkmal (*F*). Deshalb kann der Anteil der Träger des kritischen Merkmals F durch Bildung der Differenz der Summe der „Ja"-Antworten in den beiden Bedingungen geschätzt werden (*F = S1 − S2*). Die individuellen Antworten der Befragten auf die einzelnen Fragen bleiben dabei geschützt und ihre Anonymität dadurch gewahrt. Tabelle 1 veranschaulicht das Prinzip der Unmatched-Count-Technik.

Tabelle 1

Experimental- und Kontrollbedingung in einem Unmatched-Count-Technik-Design.

Experimentalbedingung	*Kontrollbedingung*
Harmlose Frage A	Harmlose Frage A
Harmlose Frage B	Harmlose Frage B
Harmlose Frage C	Harmlose Frage C
Harmlose Frage D	Harmlose Frage D
Harmlose Frage E	Harmlose Frage E
Kritische Frage F	–
S1 = A + B + C + D + E + F	*S2* = A + B + C + D + E

Trotz der einleuchtenden Logik und vergleichsweise einfachen Durchführung der Unmatched-Count-Technik ist sie bisher insgesamt sehr selten und im europäischen Raum noch gar nicht validiert worden. In den wenigen durchgeführten Untersuchungen führte die Verwendung des Verfahrens jedoch zu höheren Prävalenzschätzungen als eine direkte Befragung (LaBrie & Earleywine, 2000; Wimbush & Dalton, 1997). Probleme beim Einsatz der Technik ergeben sich dann, wenn die harmlosen Fragen nach Merkmalen mit hoher Prävalenz fragen, weil es dann dazu kommen kann, dass die zu berichtende Summe der „Ja"-Antworten die Zahl der überhaupt gestellten Fragen erreicht, was dann direkt den Schluss auf das Zutreffen

auch des sensiblen Merkmals ermöglichen würde. Auch ist das Verfahren in der von Miller (1984) vorgeschlagenen Form von geringer Effizienz bei der Parameterschätzung gekennzeichnet, weil die harmlosen Fragen ganz erhebliche, für die Frage nach dem sensiblen Merkmal jedoch irrelevante Varianz zum zu berichtenden Summenwert beitragen.

2.3.4 Randomized-Response-Technik

Eine andere experimentelle Möglichkeit der Herstellung von Anonymität bietet die von Warner (1965) entwickelte *Randomized-Response-Technik* (RRT). Das Prinzip des Warnerschen Original-Modells ist wie folgt: Bei der Befragung entscheidet ein Zufallsgenerator (z.B. ein Würfel oder der Geburtsmonat des Befragten), ob der Befragte gebeten wird, die kritische Frage („Haben Sie schon einmal illegale Drogen eingenommen?“) oder das durch Verneinung gebildete Komplement zur kritischen Frage („Haben Sie noch nie illegale Drogen eingenommen?“) zu beantworten. So könnte beispielsweise der Befragte aufgefordert werden, die kritische Frage zu beantworten, wenn er im Januar oder Februar geboren wurde (Randomisierungswahrscheinlichkeit $p = 2/12 = 0.17$) und auf die Verneinung der kritischen Frage zu antworten, wenn er im März bis Dezember geboren wurde ($1 - p = 10/12 = 0.83$). Der Ausgang des Zufallsexperimentes (d.h. hier der Geburtsmonat des Befragten und damit, welche Frage er beantwortet hat) ist dem Fragesteller nicht bekannt; er weiß also nicht, ob sich eine „Ja“-Antwort auf die kritische Frage, und damit das Vorhandensein des kritischen Merkmals, oder auf ihre Verneinung, und damit das Nicht-Vorhandensein des kritischen Merkmals, bezieht. Das individuelle Antwortverhalten wird dadurch geschützt und bleibt anonym. Auf aggregierter Ebene kann dennoch bei bekannter Verteilung der Zufallsvariable (d.h. der Geburtsmonate in der Bevölkerung) rechnerisch bestimmt werden, wie viele Befragte – im Schutze der Zufallsverschlüsselung – das sensible Merkmal eingeräumt haben, d.h. wie viele Konsumenten illegaler Drogen sich in der Stichprobe befinden.

Seit der Einführung der Technik durch Warner ist eine Vielzahl von RRT-Modellen entwickelt worden (vgl. Antonak & Livneh, 1995; Campbell & Joiner, 1973;

Chaudhuri & Mukerjee, 1988; Dawes & Smith, 1985; Scheers, 1992; Umesh & Peterson, 1991, für Übersichten). Dabei hat sich das *Forced-Response-Modell* (Dawes & Moore, 1980; Greenberg, Abul-Ela, Simmons & Horvitz, 1969) als das unter Normalbedingungen, d.h. bei niedriger bis mittelhoher Prävalenz des sensiblen Merkmals, effizienteste erwiesen: Im Vergleich zu konventionellen direkten Befragungen mit gleicher Stichprobengröße zeigt das Modell in diesen Fällen die geringste Varianzerhöhung in der Prävalenzschätzung des sensiblen Merkmals. Nur bei höheren Prävalenzraten ist ein RRT-Modell von Mangat (1994) etwas effizienter (Lensvelt-Mulders, Hox & van der Heijden, 2005).

Beim Forced-Response-Modell entscheidet der Zufallsgenerator, ob der Befragte gebeten wird, ehrlich auf die kritische Frage („Haben Sie schon einmal illegale Drogen eingenommen?“) zu antworten („Sagen Sie die Wahrheit, wenn Sie im März bis Dezember geboren wurden.“), oder ob er unabhängig vom Frageninhalt aufgefordert wird, das Vorhandensein des sensiblen Merkmals zu bejahen („Sagen Sie ‚Ja‘, wenn Sie im Januar oder Februar geboren wurden.“). Auch in diesem Modell kann bei bekannter Verteilung der Zufallsvariable die Prävalenz des sensiblen Merkmals auf aggregierter Ebene geschätzt werden, ohne die Anonymität des Einzelnen aufzuheben. In dem gerade genannten Beispiel beträgt die Randomisierungswahrscheinlichkeit, vom Zufallsgenerator aufgefordert zu werden, inhaltsunabhängig mit „Ja“ zu antworten, $p = 2/12 = 0.17$.[3]

Die verschiedenen RRT-Modelle sind vielfach eingesetzt worden, um die Prävalenz von so unterschiedlichen Verhaltensweisen wie Steuerhinterziehung, illegalem Drogenkonsum, Ladendiebstahl oder Abtreibungen zu schätzen. Es gibt Hinweise darauf, dass die Befragten die Technik tatsächlich als anonymer erleben und daher im Vergleich zu direkten Befragungen bereitwilliger sind, ehrlich zu antworten (Edgell, Himmelfarb & Duchan, 1982). Vor kurzem hat sich eine Metaanalyse systematisch mit der Frage beschäftigt, ob die RRT das Versprechen, validere Prävalenzschätzungen als direkte Befragungstechniken zu liefern, einzulösen vermag. Lensvelt-Mulders, Hox, van

[3] Man beachte, dass der Begriff Randomisierungswahrscheinlichkeit (p) im Original-Modell von Warner (1965) die Wahrscheinlichkeit, die kritische Frage beantworten zu müssen, meint, während er sich im Forced-Response-Modell (Dawes & Moore, 1980; Greenberg et al., 1969) auf die Wahrscheinlichkeit, die kritische Frage inhaltsunabhängig mit „Ja“ zu beantworten, bezieht.

der Heijden und Maas (2005) unterschieden dabei zwischen zwei Arten von Validierungsstudien, die sie als „weich“ und „hart“ bezeichneten. Bei „weichen“ Validierungsstudien ist die wahre Prävalenz des sensiblen Merkmals nicht bekannt, und die RRT wird mit einer direkten Befragungstechnik verglichen. Es wird davon ausgegangen, dass die Technik, die zu einer höheren Schätzung der Prävalenz des sensiblen Merkmals führt, die validere Schätzung liefert. Bei „harten“ Validierungsstudien ist die wahre Prävalenz des sensiblen Merkmals bekannt, und die RRT wird ebenfalls mit einer direkten Befragungstechnik verglichen. Hier liefert die Technik, deren Prävalenzschätzung weniger von der wahren Prävalenz abweicht, die validere Schätzung. In der Metaanalyse von Lensvelt-Mulders, Hox, van der Heijden und Maas (2005) zeigte sich, dass die RRT in beiden Arten von Studien direkten Befragungsmethoden überlegen war.

Trotz ihrer Vorteile und ihrer zahlreichen erfolgreichen Feldeinsätze ist die RRT aus hauptsächlich vier Gründen kritisiert worden (Antonak & Livneh, 1995; Umesh & Peterson, 1991). Erstens kann die RRT nicht verwendet werden, um den individuellen Status von Befragten zu erfassen, was z.B. das Berechnen von Korrelationen des sensiblen Merkmals mit Hintergrundvariablen erschwert. Doch gerade die Anonymität auf individueller Ebene macht den vertraulichen Charakter der Technik aus, wenngleich Rittenhouse (1996a, 1996b) gezeigt hat, dass selbst RRT-Modelle einige, wenn auch nur probabilistische, teilnehmerspezifische Informationen liefern. Andere Autoren haben logistische Regressionstechniken entwickelt, die es – auch ohne das individuelle Antwortverhalten zu kennen – durch Berücksichtigung der durch die Zufallsverschlüsselung verursachten Varianz ermöglichen, Korrelationen mit Hintergrundvariablen zu berechnen, jedoch mit hohem Schätzfehler (z.B. Lensvelt-Mulders, van der Heijden, Laudy & van Gils, 2006; Maddala, 1983; van der Heijden, van Gils, Bouts & Hox, 2000). Zweitens ist der Einsatz der Technik zeit- und damit kostenintensiver als eine direkte Befragung, weil den Befragten das Prinzip der Technik erst erklärt werden muss. Drittens sind alle RRT-Modelle wegen der Zufallsverschlüsselung im Vergleich zu konventionellen Befragungstechniken wenig effizient und damit wiederum kostenintensiver: Sogar im effizientesten RRT-Design, dem oben genannten Forced-Response-Modell, werden doppelt so viele Befragte wie bei einer direkten Befragung benötigt, um eine ähnlich niedrige Varianz des Schätzers, und damit ein vergleichbares Konfidenz-

intervall für die Prävalenzschätzung zu erhalten (Lensvelt-Mulders, Hox & van der Heijden, 2005). Man muss dabei jedoch bedenken, dass die – durch längere Durchführungszeiten oder größeren Probandenbedarf verursachte – höhere Kostenintensivität von einer erhöhten Validität mehr als kompensiert werden kann. Viertens beruht die Prävalenzschätzung bei allen RRT-Modellen auf der impliziten Annahme, dass sich die Befragten an die Regeln der verwendeten Zufallsverschlüsselung halten und stets so antworten, wie es der Zufallsgenerator vorsieht (Campbell, 1987). Dass dies nicht zwingend der Fall ist, konnte allerdings wiederholt gezeigt werden (z.B. Lensvelt-Mulders & Boeije, 2007; Locander, Sudman & Bradburn, 1976). Missachten Befragte die RRT-Regeln, wird die Prävalenz des sensiblen Merkmals unterschätzt, sofern es sich bei den Regelverweigerern um Merkmalsträger handelt, und der Vorteil der RRT gegenüber einer direkten Befragungstechnik schwindet. Clark und Desharnais (1998) haben deshalb eine RRT-Erweiterung entwickelt, die es ermöglichen soll, neben der herkömmlichen Prävalenzschätzung auch den Anteil der Befragten, der sich nicht an die Regeln hält, d.h. den Anteil der hier so genannten „Verweigerer“, zu bestimmen. Die vorliegende Arbeit beschäftigt sich mit einer Weiterentwicklung dieser Erweiterung, die im folgenden Kapitel näher beschrieben wird.

3 Verweigererdetektion im Rahmen der Randomized-Response-Technik

In dem folgenden Unterkapitel (3.1) wird die ursprüngliche Formulierung der Verweigererdetektionsvariante der RRT nach Clark und Desharnais (1998) vorgestellt. Das anschließende Unterkapitel (3.2) beschäftigt sich mit einer Reformulierung des Modells im Rahmen eines multinomialen Ansatzes durch Musch et al. (2001). Das letzte Unterkapitel (3.3) widmet sich schließlich der Darstellung von Vorarbeiten, aus denen sich die Fragestellung für die in dieser Dissertation vorliegenden Einzelarbeiten ableitet.

3.1 Ursprüngliche Formulierung

Clark und Desharnais (1998) haben für das Forced-Response-Modell der RRT (Dawes & Moore, 1980; Greenberg et al., 1969) eine Taxonomie unterschiedlicher Antwortmuster entwickelt, die auch die Möglichkeit einer Nichtbefolgung der RRT-Regeln berücksichtigt. Tabelle 2 veranschaulicht diese Taxonomie am Beispiel des sensiblen Merkmals „Konsum illegaler Drogen“.

Traditionelle RRT-Modelle unterteilen die Stichprobe in zwei disjunkte Klassen: π (den Anteil der Merkmalsträger, d.h. Drogenkonsumenten) und β (= 1 – π, den Anteil der Nicht-Merkmalsträger, d.h. Nicht-Drogenkonsumenten). Im Gegensatz dazu unterteilt die Taxonomie von Clark und Desharnais (1998) die Stichprobe in drei disjunkte Klassen: π (den Anteil der Befragten, der das sensible Merkmal aufweist und dieses auch – möglicherweise aufgrund der durch die Zufallsverschlüsselung gewährten Anonymität – zugibt; hier also ehrliche Drogenkonsumenten), β (den Anteil der Befragten, der das sensible Merkmal nicht aufweist, es deshalb wahrheitsgemäß abstreitet, gleichzeitig jedoch bereit ist, auf eine entsprechende Aufforderung durch den Zufallsgenerator mit „Ja“ zu antworten, d.h. ehrliche Nicht-Drogen-Konsumenten) und γ (= 1 – π – β, den Anteil der Befragten, der die Befolgung der Spielregeln verweigert und –

unabhängig vom Ausgang des Zufallsexperimentes – mit „Nein“ auf die kritische Frage antwortet; die hier so genannten „Verweigerer“).

Tabelle 2

Taxonomie möglicher Antwortmuster in Randomized-Response-Technik-Untersuchungen nach Clark und Desharnais (1998).

	Ehrlicher Drogen-konsument	Ehrlicher Nicht-Drogen-konsument	Verweigerer
Wahre Merkmalsausprägung (Attribut: „Drogenkonsument“)	*Ja*	*Nein*	*Unbekannt*
Populationsanteil	π	β	$\gamma\ (= 1 - \pi - \beta)$
Antwort auf die Aufforderung: „Sagen Sie die Wahrheit“	„Ja“	„Nein“	„Nein“
Antwort auf die Aufforderung: „Sagen Sie ‚Ja‘!“	„Ja“	„Ja“	„Nein“

Es ist wichtig anzumerken, dass keine Annahme über das Motiv getroffen wird, das einer mit der Wahrscheinlichkeit γ auftretenden Regelverweigerung zugrundeliegt: Es ist möglich, dass sie auf Drogenkonsum zurückgeht, den der Befragte unter gar keinen Umständen einräumen möchte, weil er vielleicht der Zufallsverschlüsselung nicht traut oder die Regeln nicht verstanden hat. Genauso ist es aber denkbar, dass sich einige Nicht-Drogen-Konsumenten entscheiden, mit einer „Nein“-Antwort auf der vermeintlich sicheren Seite zu bleiben, da sie dadurch mit dem sensiblen Merkmal gar nicht erst in Verbindung gebracht werden können. Es ist unmöglich, diese beiden Fälle empirisch zu unterscheiden, weswegen auch im Rahmen der Taxonomie keine Aussage über den wahren Status von Verweigerern getroffen wird. Dennoch stellt sie einen grundsätzlichen Fortschritt gegenüber traditionellen RRT-Modellen dar, da sie eine quantitative Bestimmung der Verweigererrate vorsieht.

Diese quantitative Bestimmung ist jedoch nicht ohne weiteres möglich: Durch das Hinzufügen eines dritten Parameters γ liegen nämlich nun mit π und β zwei

unabhängige Parameter vor, da π, β und γ gemeinsam 1 ergeben müssen. Die zwei unabhängigen Parameter können im Gegensatz zu traditionellen RRT-Modellen, in denen nur ein unabhängiger Parameter π (mit $\beta = 1 - \pi$) vorliegt, nicht mehr aus *einer* relativen Häufigkeit von „Ja"-Antworten geschätzt werden. Um eine ausreichende Datenbasis zu erhalten, kann man jedoch einen experimentellen Ansatz verfolgen und eine *zweite* unabhängige Stichprobe befragen. Dieser ist eine numerisch andere Randomisierungswahrscheinlichkeit *p* zuzuweisen, mit der die Befragten vom Zufallsgenerator aufgefordert werden, die kritische Frage inhaltsunabhängig mit „Ja" zu beantworten. Die abweichende Randomisierungswahrscheinlichkeit in der zweiten Gruppe p_2 kann – muss jedoch nicht notwendigerweise – als $1 - p_1$ gewählt werden. Unter der Annahme, dass bei randomisierter Gruppenzuteilung π, β und γ in beiden Substichproben gleich sind, liefert die Verweigererdetektionsvariante zwei unabhängige beobachtbare Häufigkeiten von „Ja"-Antworten. Diese wiederum genügen, um die beiden unabhängigen Parameter π und β für die Gesamtstichprobe zu schätzen; γ ergibt sich aus $1 - \pi - \beta$.

In der ursprünglichen Formulierung ihres Modells leiteten Clark und Desharnais (1998) analytische Formeln für die Maximum-Likelihood-Schätzung der Parameter π, β und γ ab und entwickelten einen Signifikanztest zur Prüfung der Nullhypothese, dass sich keine Verweigerer in der Gesamtstichprobe befinden ($\gamma = 0$). Dies bedeutete einen erheblichen Fortschritt gegenüber sowohl konventionellen direkten Befragungstechniken als auch gegenüber früheren RRT-Modellen: Im Idealfall einer vollständigen Regelbefolgung ($\gamma = 0$) liefert die Verweigererdetektionsvariante eine exakte Schätzung der Prävalenz des sensiblen Merkmals. Liegt dagegen ein signifikanter Anteil an Verweigerern in der Gesamtstichprobe vor ($\gamma > 0$), kann dieser Anteil zumindest bestimmt und auf dieser Basis eine untere sowie eine obere Schranke für die Prävalenz des sensiblen Merkmals angegeben werden. Zur Berechnung der unteren Schranke wird einfach davon ausgegangen, dass *kein einziger* Verweigerer – dessen wahrer Status im Rahmen des Modells unbekannt bleibt – Träger des sensiblen Merkmals ist; die untere Schranke entspricht dann π, also dem Anteil der Befragten, die das sensible Merkmal aufweisen und ohnehin einräumen. Zur Berechnung der oberen Schranke wird dagegen in einer *worst-case*-Betrachtung davon ausgegangen, dass viele oder sogar alle Verweigerer tatsächlich Träger des sensiblen Merkmals sind; ihr Anteil muss dann

schlimmstenfalls dem Anteil der Befragten, die das Merkmal ohnehin einräumen, zugeschlagen werden, was einer oberen Schranke für die Prävalenz des sensiblen Merkmals von $\pi + \gamma$ entspricht. Dazu muss allerdings die Zusatzannahme getroffen werden, dass sich niemand freiwillig als Träger eines stigmatisierenden Merkmals identifiziert, der es in Wirklichkeit gar nicht aufweist. Diese Annahme dürfte bei eindeutig unerwünschten Merkmalen in der Regel erfüllt sein, und ohne sie wären auch die Ergebnisse einer herkömmlichen direkten Befragung oder einer traditionellen RRT-Umfrage nicht interpretierbar.

3.2 Multinomiale Reformulierung

Musch et al. (2001) haben die Verweigererdetektionsvariante der RRT von Clark und Desharnais (1998) multinomial modelliert (vgl. Batchelder & Riefer, 1999; Hu, 1999). Abbildung 1 veranschaulicht diese Reformulierung in Form eines multinomialen Verarbeitungsbaums.

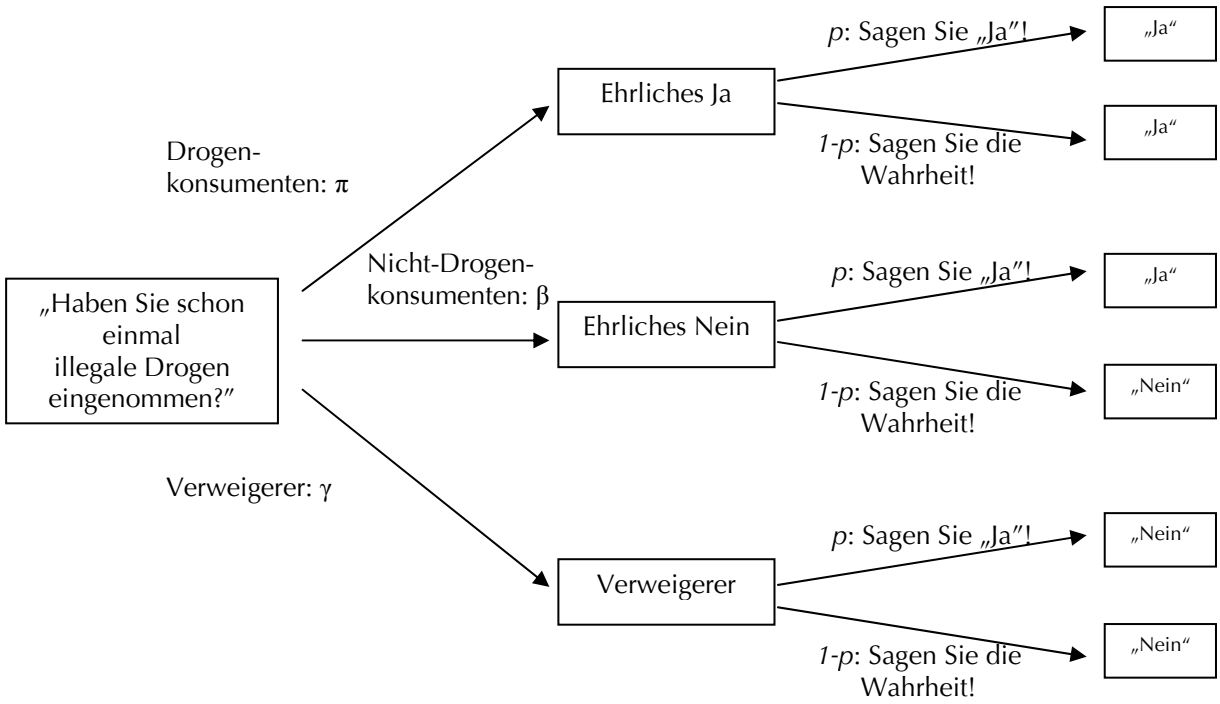


Abbildung 1: Multinomiales Modell der Verweigererdetektionsvariante der Randomized-Response-Technik.

Gemäß der Taxonomie von Clark und Desharnais (1998) wird die Stichprobe auch im multinomialen Verarbeitungsbaum in die drei disjunkten Gruppen der ehrlichen Merkmalsträger (Drogenkonsumenten, π), der ehrlichen Nicht-Merkmalsträger (Nicht-Drogenkonsumenten, β) und der Verweigerer mit unbekanntem wahrem Status ($\gamma = 1 - \pi - \beta$) unterteilt; p steht für die gewählte Randomisierungswahrscheinlichkeit, mit der die Befragten aufgefordert werden, die kritische Frage inhaltsunabhängig mit „Ja“ zu beantworten.

Die entscheidende Besonderheit der Verweigererdetektionsvariante besteht darin, dass eine zweite unabhängige Substichprobe mit einer anderen Randomisierungswahrscheinlichkeit befragt wird, was die Schätzung des neuen Parameters γ ermöglicht. Abbildung 2 zeigt einen entsprechend erweiterten, verbundenen multinomialen Verarbeitungsbaum mit zwei Gruppen von Befragten, in denen zwei unterschiedliche Randomisierungswahrscheinlichkeiten p_1 und p_2 verwendet werden.

Da die Randomisierungswahrscheinlichkeiten p_1 und p_2 vom Forscher gewählt werden und damit bekannt sind, können nun ausgehend von den beobachteten Häufigkeiten der „Ja“- und „Nein“-Antworten in den zwei Gruppen mit Hilfe des Expectation-Maximization-(EM) Algorithmus (Hu & Batchelder, 1994) die Maximum-Likelihood-Schätzer der Parameter π, β und γ bestimmt werden. Dies geschieht üblicherweise unter Verwendung von spezialisierter Statistik-Software, wie z.B. HMMTree (Stahl & Klauer, 2007).[4]

[4] Genau genommen muss hierfür der in Abbildung 1 bzw. 2 dargestellte multinomiale Verarbeitungsbaum zunächst in einen binären multinomialen Verarbeitungsbaum reparametrisiert werden, d.h. in einen Baum mit nur zwei Ästen pro Wurzel. In einem solchen binären Baum wird der übrig gebliebene Parameter als bedingte Wahrscheinlichkeit eines anderen formuliert. Hierbei handelt es sich lediglich um einen mathematischen Zwischenschritt, der zu Gunsten der anschaulicheren Darstellung mit drei Ästen, d.h. einem Ast pro Parameter, in der Abbildung nicht berücksichtigt wurde.

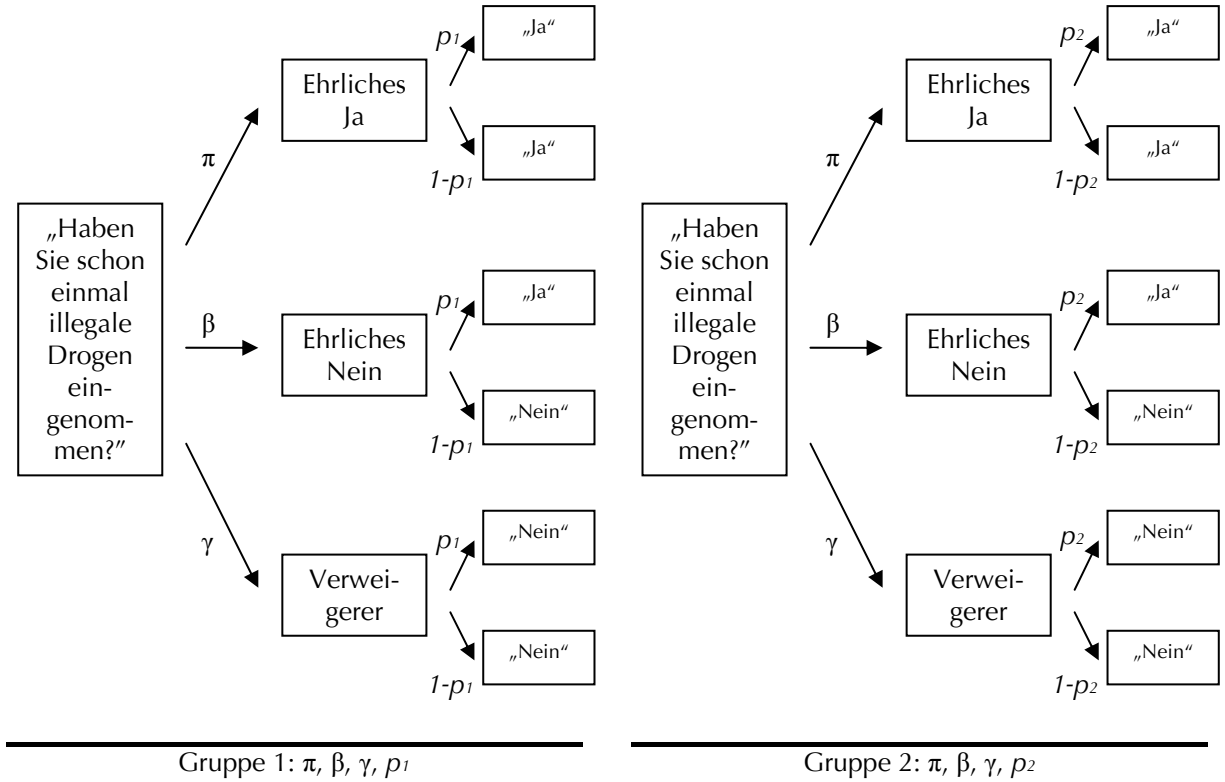


Abbildung 2: Multinomiales Modell der Verweigererdetektionsvariante der Randomized-Response-Technik mit zwei Gruppen von Befragten. Die Befragten werden mit unterschiedlichen Randomisierungswahrscheinlichkeiten p_1 und p_2 aufgefordert, die kritische Frage inhaltsunabhängig mit „Ja" zu beantworten.

Die so vorgenommene Bestimmung der Parameter bietet drei Vorteile gegenüber der ursprünglich von Clark und Desharnais (1998) vorgeschlagenen Berechnungsweise. Erstens bleibt der Parameterraum bei der Schätzung auf das allein sinnvoll interpretierbare Intervall von 0 bis 1 beschränkt. Zweitens sind auf einfachem Wege flexible Modellerweiterungen möglich, für die nicht in jedem Einzelfall erst ein Schätzer auf analytischem Wege bestimmt werden muss. So können beispielsweise weitere Befragungsbedingungen (etwa eine direkte Befragung zur Kontrolle) als zusätzliche Bäume in das verbundene Modell integriert werden. Oder es können nach Moderatorvariablen (beispielsweise nach Geschlecht) unterteilte Subgruppen untersucht werden, wobei jede Subgruppe einen eigenen Baum erhält. Drittens erlauben multinomiale Modelle eine flexible Testung von Parameterrestriktionen. So kann beispielsweise die Annahme geprüft werden, dass keine Verweigerung auftritt ($\gamma = 0$), oder dass sich ein bestimmter

Parameter in den verschiedenen Subgruppen nicht unterscheidet (z.B. $\pi_{Männer} = \pi_{Frauen}$). Hierzu wird einfach der jeweilige Parameter mit Null bzw. dem Parameter, mit dem er verglichen werden soll, gleichgesetzt, und man untersucht, ob sich infolge dessen die Passung des Modells verschlechtert. Die Passung wird über die asymptotisch χ^2-verteilte G^2-Statistik inferenzstatistisch abgesichert und eine Veränderung der Passung dementsprechend über ΔG^2. Zur Bestimmung der Anzahl der Freiheitsgrade (*df*) von ΔG^2 muss die Differenz aus der Anzahl der Parameter (*S*) und der Anzahl der eingeführten Restriktionen (*R*) berechnet werden ($df = S - R$). Man beachte, dass das in Abbildung 2 dargestellte verbundene multinomiale Modell saturiert ist und daher perfekt passt ($G^2 = 0$ bei $df = 0$), womit klar wird, dass die gesamte Fehlpassung in einem restringierten Modell auf die Restriktionen zurückzuführen ist.

3.3 Bisherige Vorarbeiten

Obwohl die Verweigererdetektionsvariante der RRT (Clark & Desharnais, 1998; Musch et al., 2001) auf theoretischer Ebene eine entscheidende Verbesserung gegenüber traditionellen RRT-Modellen darstellt, liegen bislang erst wenige Validierungsstudien zu ihr vor. Diese werden im Folgenden kurz dargestellt.

In der ersten Studie von Musch et al. (2001) wurde in einer Stichprobe von 568 Internetsurfern die Prävalenz von Steuerhinterziehung mit Hilfe der multinomialen Verweigererdetektionsvariante und einer direkten Befragung vergleichend geschätzt. In der direkten Kontrollbedingung beantworteten nur 28% der Befragten die kritische Frage („Haben Sie schon einmal Steuern hinterzogen?“) mit „Ja“; 72% der Befragten antworteten mit „Nein“. Unter RRT-Bedingungen dagegen bekannten sich $\pi = 44\%$ der Befragten zur Steuerhinterziehung bei einem Verweigereranteil von $\gamma = 32\%$ und einer Anteil ehrlicher Steuerzahler von $\beta = 24\%$. Die Prävalenzschätzung in der RRT-Bedingung lag signifikant über der Schätzung aus der direkten Befragung, und der Verweigereranteil in der RRT-Befragung wich signifikant von Null ab. Das Prävalenzminimum für Steuerhinterziehung wurde also in dieser Stichprobe auf $\pi = 44\%$ geschätzt, und das entsprechende Maximum auf $\pi + \gamma = 44\% + 32\% = 76\%$. Bei

alleinigem Einsatz der direkten Befragung wäre also selbst das Prävalenzminimum erheblich unterschätzt werden.

In einer weiteren Online-Untersuchung von Musch und Plessner (eingereicht) wurden 467 Wettkampfsportler sowohl unter RRT-Bedingungen als auch direkt zum Thema Doping befragt. Wiederum bekannte sich in der direkten Bedingung nur ein Bruchteil der Sportler zur Einnahme von verbotenen leistungssteigernden Substanzen (8%). Dieser Anteil war unter RRT-Bedingungen mit $\pi = 42\%$ signifikant höher. Auch in dieser Untersuchung entschied sich ein signifikanter Anteil der Befragten ($\gamma = 16\%$), die RRT-Regeln zu verweigern, so dass das Prävalenzmaximum für Doping 50 Prozentpunkte über der Prävalenzschätzung aus der direkten Befragungsbedingung lag ($\pi + \gamma = 58\%$). In dieser Untersuchung wurde von der durch die multinomiale Modellierung gebotenen Möglichkeit Gebrauch gemacht, die Gesamtstichprobe in Subgruppen zu unterteilen. Dabei zeigte sich, dass die Prävalenz von Doping unter Bodybuildern ($\pi_{Bodybuilder} = 58\%$) viel höher war als unter Vertretern anderer Sportarten ($\pi_{andere\ Sportler} = 33\%$), und dass Bodybuilder seltener die Befolgung der RRT-Regeln verweigerten ($\gamma_{Bodybuilder} = 7\%$ versus $\gamma_{andere\ Sportler} = 25\%$). Wie eine genaue Betrachtung der beiden Prävalenzmaxima ($\pi_{Bodybuilder} + \gamma_{Bodybuilder} = 65\%$ versus $\pi_{andere\ Sportler} + \gamma_{andere\ Sportler} = 58\%$) zeigt, kann sogar die Vermutung, dass Bodybuilder möglicherweise nicht (viel) häufiger auf Dopingmittel zurückgreifen als andere Sportler, dafür jedoch bereiter sind, dies zuzugeben, nicht ohne weiteres zurückgewiesen werden.

Nach dem gleichen Prinzip haben Musch und Bröder (eingereicht) in einer WWW-Umfrage die Prävalenz von Software-Piraterie untersucht, und Moshagen, Musch, Ostapczuk und Zhao (in Vorbereitung) haben chinesische Studenten zum Thema Zahnhygiene befragt. Beide Studien erbrachten signifikant höhere Prävalenzschätzungen des sensiblen Merkmals (Software-Piraterie bzw. mangelnde Zahnhygiene) unter RRT- als unter direkten Befragungsbedingungen. Darüber hinaus unterschied sich der Anteil der Verweigerer in beiden Studien signifikant von Null, es war also im Rahmen des Modells möglich nachzuweisen, dass sich bei weitem nicht alle Befragten an die RRT-Regeln hielten.

Die Ergebnisse der ersten Anwendungen der multinomialen Verweigererdetektionsvariante der RRT (Clark und Desharnais, 1998; Musch et al., 2001) lassen sich wie folgt zusammenfassen: Die Technik führte in vier – der Unterscheidung von

Lensvelt-Mulders, Hox, van der Heijden und Maas (2005) folgend – „weichen“ Validierungsstudien über verschiedene sensible Themenbereiche (Steuerhinterziehung, Doping, Software-Piraterie, mangelnde Zahnhygiene) hinweg zu konsistent höheren Prävalenzschätzungen als eine einfache direkte Befragung. Mit ihrer Hilfe konnte auf eine vergleichsweise einfache Art und Weise bestätigt werden, dass sich über verschiedene Themen hinweg signifikante Anteile der Befragten nicht an die RRT-Regeln halten und die nicht bei allen Teilnehmern vorhandene Bereitschaft, bei RRT-Befragungen die Instruktionen zu befolgen, ein zu beachtendes Problem darstellt. Zudem verhalf sie zur Aufdeckung erster inhaltlich relevanter Gruppenunterschiede.

4 Fragestellung

Die Verweigererdetektionsvariante der RRT ist im Vergleich zu traditionellen RRT-Modellen weniger effizient, da zur Schätzung des Verweigereranteils eine zweite Substichprobe benötigt wird. Das der Verweigererdetektion zugrundeliegende Forced-Response-Modell der RRT ist wiederum weniger effizient als konventionelle direkte Befragungsmethoden. Wie weiter oben ausgeführt wurde (vgl. 2.3.4), lassen sich die dadurch verursachten Mehrkosten rechtfertigen, wenn sie von einem konkreten Nutzen für den (Anwendungs-)Forscher kompensiert werden. Im vorigen Kapitel wurden die Ergebnisse erster vielversprechender Anwendungen der Technik vorgestellt. Bevor man jedoch guten Gewissens, d.h. empirisch fundiert, dazu raten kann, die aufwändige multinomiale Verweigererdetektionsvariante anstelle eines traditionellen RRT-Modells oder der noch einfacher durchführbaren direkten Befragung zu verwenden, sollte das Verweigererdetektionsmodell einer gründlichen Überprüfung und Validierung unterzogen werden. Dies ist das Ziel der vorliegenden Arbeit.

In einem ersten Experiment wurde im Rahmen einer Validierungsstudie geprüft, ob sich die bislang nur in Online-Studien belegte Überlegenheit der Verweigererdetektionsvariante der RRT auch in einer Papier-Bleistift-Variante replizieren und dabei auf einen neuen Inhaltsbereich – die Non-Compliance bei der Medikamenteneinnahme – generalisieren lässt. Die Ergebnisse dieses Experiments sind in der folgenden Arbeit zusammengefasst:

- Ostapczuk, M., Musch, J. & Moshagen, M. (eingereicht a). *Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique*.

In einem zweiten Experiment wurde die Verwendbarkeit der Verweigererdetektionsvariante für Vergleiche zwischen Gruppen geprüft, bei denen vermutet werden muss, dass sie sich entweder hinsichtlich der Prävalenz des sensiblen Merkmals, oder hinsichtlich der Tendenz, den Selbstbericht über dieses Merkmal in Richtung sozialer Erwünschtheit zu verfälschen, oder sogar in beiderlei Hinsicht

voneinander unterscheiden. Dazu wurde mit Hilfe des Verweigererdetektionsmodells geprüft, ob sich ein in der Literatur wiederholt dokumentierter Gruppenunterschied – der Effekt der Bildung auf die Stärke von Vorurteilen gegenüber Ausländern – als Artefakt einer gruppenspezifisch unterschiedlich stark ausgeprägten Tendenz zur sozial erwünschten Selbstauskunft bei tatsächlich gleicher Merkmalsprävalenz in beiden Gruppen erklären lässt.

- Ostapczuk, M., Musch, J. & Moshagen, M. (eingereicht b). *A randomized-response investigation of the education effect in attitudes towards foreigners*.

In einem dritten Experiment wurde die Validität des Verweigererdetektionsmodells mit der eines konkurrierenden Verfahrens zur Kontrolle von Antwortverzerrungen verglichen. Erstmals wurde dazu die Möglichkeit genutzt, auf der Basis einer mit dem Verweigererdetektionsmodell bestimmten oberen Schranke für die Prävalenz des sensiblen Merkmals zu prüfen, ob eine projektive Befragungstechnik zu einer Überschätzung der Prävalenz dieses Merkmals führt.

- Ostapczuk, M. & Musch, J. (eingereicht). *Projective questioning overestimates the prevalence of negative attitudes towards people with physical and mental disabilities*.

In einem vierten Experiment wurde geprüft, ob sich das Verweigererdetektionsmodell verbessern lässt, indem die dem Modell inhärente Asymmetrie zwischen bejahenden und verneinenden Antworten aufgehoben wird. Im einfachen Verweigererdetektionsmodell bringt eine vom Zufallsgenerator erzwungene „Ja“-Antwort auch Nichtmerkmalsträger in den Verdacht der Merkmalsträgerschaft; dies ist notwendig, damit eine „Ja“-Antwort nicht länger den wahren Merkmalsstatus offenbart und so auch Merkmalsträger zu ehrlichen Antworten ermutigt werden. Durch eine „Nein“-Antwort kann der Befragte den Verdacht, Merkmalsträger zu sein, jedoch von vornherein von sich weisen. Um diese Asymmetrie zu vermeiden und dem von ihr ausgehenden Anreiz zur Nichtbefolgung der RRT-Instruktionen entgegenzuwirken, wurden deshalb im

vierten Experiment auch „Nein“-Antworten vom Zufallsgenerator erzwungen. Geprüft wurde, ob dadurch die Verweigererrate wirksam reduziert werden kann.

- Ostapczuk, M., Moshagen, M., Zhao, Z. & Musch, J. (eingereicht). *Assessing sensitive attributes using the randomized-response-technique: Evidence for the importance of response symmetry*.

5 Zusammenfassung der Einzelarbeiten

In den folgenden Unterkapiteln (5.1-5.4) werden die vier oben genannten Einzelarbeiten skizziert. Die Daten wurden in sämtlichen Experimenten mit Hilfe der Programme SPSS 13.0 (2004) und HMMTree (Stahl & Klauer, 2007) analysiert.

5.1 *Experiment I*: Non-Compliance bei der Medikamenteneinnahme

In dem ersten Experiment wurde geprüft, ob sich die mit Hilfe der multinomialen Verweigererdetektionsvariante der RRT in bisherigen Online-Studien erzielten Effekte einer höheren Prävalenz sozial unerwünschter Merkmale unter RRT- als unter direkten Befragungsbedingungen auch in einer Papier-und-Bleistift-Untersuchung finden lassen. Darüber hinaus wurde mit der Non-Compliance bei der Medikamenteneinnahme ein neuer Themenbereich, in dem Antwortverzerrungen in Folge von sozialer Erwünschtheit zu erwarten sind, untersucht.

Non-Compliance bei der Medikamenteneinnahme, d.h. die Einnahme von Medikamenten in anderer Frequenz oder Dosierung, als sie vom Arzt vorgeschrieben wurde, stellt sowohl eine Herausforderung für die medizinische Forschung als auch für die Praxis und damit nicht zuletzt ein großes gesundheitsökonomisches Problem dar. Forschungsergebnisse zur Wirksamkeit von Medikamenten, die unter Bedingungen völliger Compliance erzielt werden, können nicht ohne weiteres auf die Praxis übertragen werden, wenn dort im Gegensatz zum Labor oder kontrollierten Studien mit Abweichungen von den Einnahmevorschriften zu rechnen ist. Bleibt der Behandlungserfolg aus, weiß der behandelnde Arzt nicht, ob dies an dem Medikament oder dem Einnahmeverhalten des Patienten liegt. Das kann zusätzliche Arztbesuche, das Ausprobieren von weiteren Medikamenten und Behandlungen oder gar Notfall-Einweisungen zur Folge haben, wodurch große und unnötige volkswirtschaftliche Kosten entstehen (Düsing, 2003; Farmer, 1999; Granger et al., 2004; Thieda, Beard, Richter & Kane, 2003). Eines der größten Probleme der Compliance-Forschung stellt die Messung des Phänomens dar: Objektive Methoden – wie z.B. „Pill Count“, „Medication Event

Monitoring Systems“ oder die Messung von Medikamentenmetaboliten im Blut oder Urin – sind zwar in der Regel valider, aber auch umständlicher und teurer in der Durchführung als subjektive Methoden, wie z.B. die Befragung des Pflegepersonals, der Angehörigen oder des Patienten selbst (DiMatteo, 2004; Düsing, 2003; Farmer, 1994; Garber, Nau, Erickson, Aikens & Lawrence, 2004; Gagné & Godin, 2005). Entsprechend variieren die in der Literatur berichteten Non-Compliance-Prävalenzraten in Abhängigkeit von der verwendeten Erfassungsmethode zwischen 0% und 95% (bei einem Median von 24%); das wahre Ausmaß des Problems ist also schwer einzuschätzen (DiMatteo, 2004; Farmer, 1999; Kravitz & Melnikow, 2004). Wünschenswert wäre deshalb die Entwicklung bzw. Verwendung einer vergleichsweise einfachen Erfassungsmethode, welche dennoch valide Prävalenzschätzungen liefert, die nicht von sozialer Erwünschtheit verzerrt sind. Damit handelt es sich bei der Non-Compliance um einen idealen neuen Themenbereich zur Erprobung der multinomialen Verweigererdetektionsvariante der RRT. Trotz wiederholter Aufrufe, die RRT zur Verbesserung der epidemiologischen Erfassung von Non-Compliance einzusetzen (Rittenhouse 1996a, 1996b; Soeken, 1987), ist dies bisher erst in einer einzigen Studie geschehen (Volicer & Volicer, 1982), in der allerdings ein älteres RRT-Modell, das keine Möglichkeit zur Verweigererentdeckung aufwies, in einer kleinen Stichprobe eingesetzt wurde, so dass die Ergebnisse wenig aussagekräftig ausfielen.

In der vorliegenden Untersuchung wurden 597 Patienten in zwei Arztpraxen und einem Krankenhaus zu ihren Gewohnheiten bei der Medikamenteneinnahme befragt. Die Antworten wurden im Wartezimmer oder während eines präoperativen Krankenhausaufenthaltes erhoben. Neben einer Reihe von demographischen und gesundheitsbezogenen Fragen mussten die Patienten die folgende kritische Frage zur Erfassung von Non-Compliance beantworten: „Haben Sie schon einmal ein Medikament, das Ihnen der Arzt verschrieben hat, absichtlich und für längere Zeit anders als vorgeschrieben eingenommen (indem Sie es z.B. deutlich zu kurz oder zu lange, zu häufig oder zu selten, zu früh oder zu spät am Tag eingenommen haben)?“. Die Teilnehmer wurden randomisiert in einem Verhältnis von 2:2:1 einer der drei Gruppen „RRT mit Randomisierungswahrscheinlichkeit p_1“, „RRT mit Randomisierungswahrscheinlichkeit p_2“ und „direkte Befragung“ zugewiesen. Das Verhältnis trug der im Vergleich zur direkten Befragung geringeren Effizienz der Randomized-Response-

Technik Rechnung (vgl. Lensvelt-Mulders, Hox, & van der Heijden, 2005). Die drei Fragebogenversionen unterschieden sich lediglich in den Anweisungen zur Beantwortung der kritischen Frage. In der direkten Kontrollbedingung (n = 124) wurde sie ohne weitere Erklärungen im Rahmen der gesundheitsbezogenen Fragen gestellt. In der RRT-Gruppe mit der niedrigen Randomisierungswahrscheinlichkeit (RRT1; n = 241) lautete die Instruktion: „Wurde Ihr Vater im Januar oder Februar geboren, dann antworten Sie bitte auf die folgende Frage unabhängig vom Inhalt mit ‚Ja'. Wurde Ihr Vater jedoch in einem anderen Monat geboren, so antworten Sie bitte auf die folgende Frage wahrheitsgemäß." Wie die Geburtsstatistiken des Statistischen Bundesamtes belegen, betrug damit die Randomisierungswahrscheinlichkeit, inhaltsunabhängig zu einer „Ja"-Antwort aufgefordert zu werden, $p_1 = 2/12 = 1/6 = 0.17$. In der RRT-Gruppe mit der hohen Randomisierungswahrscheinlichkeit (RRT2; n = 232) wurde die kritische Frage wie folgt eingeleitet: „Wurde Ihr Vater im März, April, Mai, Juni, Juli, August, September, Oktober, November oder Dezember geboren, dann antworten Sie bitte auf die folgende Frage unabhängig vom Inhalt mit ‚Ja'. Wurde Ihr Vater jedoch in einem anderen Monat geboren, so antworten Sie bitte auf die folgende Frage wahrheitsgemäß." Hier betrug die Randomisierungswahrscheinlichkeit also $p_2 = 1 - p_1 = 10/12 = 5/6 = 0.83$. In beiden RRT-Bedingungen folgte eine Erklärung, inwiefern die Verwendung des Geburtsmonats als Randomisierungsprozess dazu beiträgt, dass die individuellen Antworten anonym bleiben. Das verbundene multinomiale Modell, welches dieses Design abbildete, bestand damit im vorliegenden Experiment aus drei Bäumen, nämlich je einem Baum für die RRT1-, RRT2- und die direkte Befragungsgruppe.

Als unabhängige Variable diente in diesem Experiment der Befragungsmodus mit den beiden Realisierungen „zufallsverschlüsselte Befragung unter RRT-Bedingungen" und „direkte Befragung". Die geschätzte Lebenszeitprävalenz der Non-Compliance bei der Medikamenteneinnahme wurde als abhängige Variable betrachtet. Die Hypothese lautete, dass eine mögliche Non-Compliance bei der Medikamenteneinnahme unter RRT-Bedingungen bereitwilliger und damit häufiger als bei einer direkten Befragung eingeräumt wird.

Tabelle 3

Lebenszeitprävalenz von Non-Compliance bei der Medikamenteneinnahme in Abhängigkeit vom Befragungsmodus.

„Haben Sie schon einmal ein Medikament, das Ihnen der Arzt verschrieben hat, absichtlich und für längere Zeit anders als vorgeschrieben eingenommen (indem Sie es z.B. deutlich zu kurz oder zu lange, zu häufig oder zu selten, zu früh oder zu spät am Tag eingenommen haben)?“ (N = 597)		
Direkte Befragung (n = 124)		
	% „Ja“	21%
	% „Nein“	79%
Randomized-Response-Befragung (n = 473)		
	Ehrliches „Ja“ (π)	33%
	Ehrliches „Nein“ (β)	20%
	Verweigerer (γ)	47%
	$\Delta G^2(1)$: $\gamma = 0$[†]	174.19**
	$\Delta G^2(1)$: % „Ja“ = π[‡]	4.59*

Bemerkungen: [†]Hohe Werte bedeuten, dass sich die Modellpassung verschlechtert, wenn man annimmt, dass es keine Verweigerer in der Stichprobe gibt ($\gamma = 0$). [‡]Hohe Werte bedeuten, dass sich der Anteil der non-complianten Patienten in der DB-Bedingung (% „Ja“) nicht vom Anteil der non-complianten Patienten in den RRT-Bedingungen (π) unterscheidet. *$p < .05$, **$p < .01$.

Tabelle 3 zeigt die Parameterschätzungen für das saturierte Modell, $G^2(0) = 0$. Bei der direkten Befragung räumten nur 21% der befragten Patienten Non-Compliance bei der Medikamenteneinnahme ein. Unter Verwendung des Verweigererdetektionsmodells wurde dieser Anteil auf $\pi = 33\%$ geschätzt; das ist ein zufallskritisch absicherbar höherer Wert, denn die Annahme, dass sich die beiden Parameterschätzungen gleichsetzen lassen, führte zu einer signifikanten Verschlechterung der Modellpassung, $\Delta G^2(1) = 4.59$, $p < .05$. Darüber ergab die Auswertung, dass sich mit $\gamma = 47\%$ fast die Hälfte der unter RRT-Bedingungen befragten Patienten nicht an die RRT-Regeln hielt. Dieser Anteil unterschied sich signifikant von Null, $\Delta G^2(1) = 174.19$, $p < .01$. In Abhängigkeit davon, ob die Verweigerer ihre Medikamente tatsächlich eingenommen hatten oder nicht, wurde für die Lebenszeitprävalenz der Non-Compliance eine untere Schranke von $\pi = 33\%$ und eine obere Schranke von $\pi + \gamma = 33\% + 47\% = 80\%$

bestimmt. Damit lag die untere Schranke etwas über dem Median der bisher in der Literatur berichteten Prävalenzraten für Non-Compliance (24%) und die obere Schranke etwas unter dem bisher berichteten Maximum (95%).

Insgesamt zeigt das erste Experiment, dass sich das in den Vorarbeiten berichtete Ergebnismuster von unter RRT-Bedingungen höheren – und damit wahrscheinlich valideren – Prävalenzschätzungen für sozial unerwünschtes Verhalten auch in einem neuen Darbietungsmodus (Papier-und-Bleistift statt online) und in einem neuen Themenbereich (Non-Compliance bei der Medikamenteneinnahme) replizieren lässt. Die mit Hilfe der Verweigererdetektionsvariante der RRT geschätzte Lebenszeitprävalenz von Non-Compliance lag mit 33% deutlich über der Prävalenzschätzung der direkten Befragung (21%). Weiterhin unterstreichen die Ergebnisse den Nutzen der Verweigererentdeckung im Rahmen von RRT-Modellen, da bei Verwendung eines älteren RRT-Modells ohne Verweigererdetektion nicht aufgefallen wäre, dass sich fast die Hälfte (47%) der unter RRT-Bedingungen befragten Teilnehmer nicht an die Regeln der Technik hielt, so dass mit einem herkömmlichen RRT-Modell die Prävalenz der Non-Compliance erheblich unterschätzt worden wäre.

5.2 *Experiment II*: Bildungseffekt bei ausländerfeindlichen Einstellungen

Im zweiten Experiment wurde das untersuchte multinomiale Modell flexibel erweitert, um es für Mehrgruppenuntersuchungen anzupassen. Die vom multinomialen Modellierungsansatz gebotene Möglichkeit zur Prüfung auf Parametergleichheit in Subgruppen wurde genutzt, um einen inhaltlich bedeutsamen Gruppenunterschied auf seine Gültigkeit hin zu untersuchen.

Frühere Untersuchungen haben gezeigt, dass sich hinsichtlich ausländerfeindlicher Einstellungen ein deutlicher Bildungseffekt zeigt; Personen mit geringer Bildung geben bei Selbstauskünften regelmäßig negativere Einstellungen gegenüber Ausländern an als Personen mit höherer Bildung (in Deutschland: Bergmann & Erb, 1991; Mielke & Mummendey, 1995; Silbermann & Hüsers, 1995; in den USA: Pass, 1988; Photiadis & Biggar, 1962, Robinson & Rohde, 1946; Weiner, 1974; in Kanada:

Jerabek & de Man, 1994; in Mexiko: Cohen Shabat, 1993; in Australien und Südafrika: Ray, 1990; in den Niederlanden, Frankreich und Großbritannien: Wagner & Zick, 1995; in Österreich: Jimenez, 1999; in der Schweiz: Fend, 1994; in Schweden: Abraham, 1966). Einige Erklärungsansätze vermuten, dass dies einen echten Einstellungsunterschied widerspiegelt. Nicht ausgeschlossen werden kann jedoch, dass der Bildungseffekt nur ein Artefakt einer stärkeren Tendenz zur sozial erwünschten Antwort bei den Personen mit höherer Bildung ist. Ein anderer Ansatz nimmt deshalb an, dass gebildete Menschen möglicherweise gar keine positiveren Einstellungen gegenüber Ausländern haben, sondern in Befragungen lediglich sensibler für die sozial erwünschte, in der Regel ausländerfreundliche Antwort sind und deshalb ihre Antwort in diese Richtung verzerren (Hopf, 1999; Mielke & Mummendey, 1995; Wagner & Zick, 1995). Die bisherigen Arbeiten zu diesem Thema legen eher nahe, dass es sich bei dem Bildungseffekt um einen substantiellen Effekt handelt, sind diesbezüglich jedoch nicht völlig eindeutig (Mielke & Mummendey, 1995; Robinson & Rohde, 1946; Wagner & Zick, 1995). Deswegen wurde mehrfach vorgeschlagen, zu überprüfen, ob der Effekt auch bei Kontrolle möglicher Antwortverzerrungen erhalten bleibt (Hopf, 1999; Mielke & Mummendey, 1995). Cobb (2002) hat sogar explizit vorgeschlagen, die RRT zu diesem Zweck zu verwenden. Dieser Empfehlung wurde hier gefolgt.

Um eine möglichst bildungsheterogene Stichprobe zu erhalten, wurden die 606 Teilnehmer in zwei Universitäten, drei Arztpraxen, einem Krankenhaus und einer Berufsschule rekrutiert. Die Versuchsanordnung glich dem Design von Experiment I: Die Befragten bekamen einen Fragebogen vorgelegt, in dem sie neben mehreren demographischen Fragen auch Items zu ihren Erfahrungen mit und Einstellungen gegenüber dunkelhäutigen Afrikanern bearbeiten mussten, von denen jedoch nur ein Item sensibler Natur war. Dieses kritische Item wurde nach Literatursichtung und einem Online-Vortest mit $N = 63$ Teilnehmern nach den folgenden drei Kriterien ausgewählt: Es sollte aus einer bewährten Skala zur Erfassung von Einstellungen gegenüber Ausländern stammen, die bereits zur Untersuchung des Bildungseffektes verwendet wurde. Es sollte von mittlerer sozialer (Un-)Erwünschtheit sein, um keine Extremverteilung, d.h. nur „Ja“- oder nur „Nein“-Antworten, zu provozieren. Ferner sollte das kritische Item die Einstellung gegenüber einer konkreten Gruppe von Ausländern, die sich in Deutschland lediglich niedriger bis mittlerer Beliebtheit erfreut (anstatt der

abstrakten Zielgruppe „die Ausländer"), erfassen. Das Item, das alle drei Kriterien am besten erfüllte, lautete: „Angenommen, Sie hätten eine 20 Jahre alte Tochter. Würde es Sie stören, wenn diese eine Beziehung mit einem nigerianischen Staatsbürger schwarzer Hautfarbe eingehen würde?".[5] Wie in Experiment I wurden die Befragten in einem Verhältnis von 2:2:1 zufällig den Gruppen RRT1 (niedrige Randomisierungswahrscheinlichkeit p_1, n = 246), RRT2 (hohe Randomisierungswahrscheinlichkeit p_2, n = 230) und direkte Befragung (n = 130) zugewiesen. In der Kontrollbedingung wurde die kritische Frage direkt und ohne weitere Erläuterungen gestellt. In den RRT-Gruppen wurde die Zufallsverschlüsselung über den Geburtsmonat der Mutter verwirklicht. In RRT1 wurden die Befragten aufgefordert, die kritische Frage inhaltsunabhängig mit „Ja" zu beantworten, wenn ihre Mutter im Januar oder Februar geboren wurde (p_1 = 0.17), sonst jedoch ehrlich zu antworten. In RRT2 lautete die Anweisung komplementär zu RRT1, die kritische Frage unabhängig vom Inhalt mit „Ja" zu beantworten, wenn die Mutter des Befragten im März bis Dezember geboren wurde (p_2 = 0.83), und ansonsten ehrlich zu antworten. Wiederum wurde den Umfrageteilnehmern in den RRT-Bedingungen erklärt, auf welche Weise diese Zufallsverschlüsselung zur Erhöhung ihrer Anonymität beiträgt. Der Bildungsstand wurde der Forschungstradition folgend als dichotome Variable operationalisiert (vgl. Mielke & Mummendey, 1995; Wagner & Zick, 1995): Befragte, die das Abitur oder einen höheren Abschluss abgelegt haben, galten als hoch gebildet (n = 282), alle restlichen Befragten wurden als niedrig gebildet eingestuft (n = 324). Das entsprechend erweiterte verbundene multinomiale Modell der Verweigererdetektionsvariante der RRT bestand damit aus sechs Verarbeitungsbäumen: je einem Baum für die RRT1- (n = 113), RRT2- (n = 104) und die direkte Befragungsgruppe (n = 65) innerhalb der Substichprobe der hoch gebildeten Teilnehmer und je einem Baum für die RRT1- (n = 133), RRT2- (n = 126) und die

[5] Dieses Item, das in einer ähnlichen Form in den Untersuchungen von Silbermann und Hüsers (1995) sowie Jimenez (1999) zum Einsatz kam, wurde der Social Distance Scale (Bogardus, 1925, 1933) entnommen. In der Voruntersuchung erzielte es auf einer Skala von 1 bis 5 (mit 1 = weder erwünscht noch unerwünscht und 5 = sehr unerwünscht) einen Mittelwert von M = 3.79 (SD = 1.23). Bezüglich der gewählten Ausländergruppe zeigt eine Übersicht von Bergmann und Erb (1991), dass dunkelhäutige Afrikaner innerhalb der wenig beliebten religiösen und ethnischen Minderheiten in Deutschland eine Mittelstellung zwischen den noch weniger beliebten Arabern und Türken und den etwas beliebteren Juden und Israelis einnehmen.

direkte Befragungsgruppe (n = 65) innerhalb der Substichprobe der niedrig gebildeten Teilnehmer.

Die beiden unabhängigen Variablen waren der Befragungsmodus mit den Realisierungen „RRT“ versus „direkte Befragung“ und der Bildungsstand mit den Realisierungen „hoch gebildet“ versus „niedrig gebildet“. Als abhängige Variable wurde die Einstellung gegenüber Ausländern erhoben. Als eindeutig ausländerfeindlich wurden diejenigen Befragten klassifiziert, welche die diskriminierende Frage ehrlich bejahten. Als eindeutig ausländerfreundlich wurden diejenigen Befragten klassifiziert, welche die diskriminierende Frage ehrlich verneinten. Die Hypothesen lauteten, dass die hoch gebildeten Befragten sowohl unter direkten Befragungsbedingungen als auch unter RRT-Bedingungen weniger ausländerfeindliche bzw. mehr ausländerfreundliche Einstellungen als die niedrig gebildeten Befragten berichten sollten, falls es sich bei dem Bildungseffekt um einen wahren Einstellungsunterschied handelt. Handelt es sich bei dem Effekt jedoch um ein Artefakt differentieller Sensibilität für sozial unerwünschte Antworten, so sollten die hoch gebildeten Befragten zwar unter direkten Befragungsbedingungen weniger ausländerfeindliche bzw. mehr ausländerfreundliche Einstellungen als die niedrig gebildeten Befragten berichten, unter RRT-Bedingungen sollte der Unterschied jedoch verschwinden oder sich möglicherweise sogar umkehren.

Tabelle 4 sind die Ergebnisse für das saturierte Modell mit $G^2(0) = 0$ zu entnehmen. Man erkennt, dass der Bildungseffekt unter direkten Befragungsbedingungen repliziert werden konnte: Während von den hoch gebildeten Teilnehmern nur 25% eine ausländerfeindliche (bzw. 75% eine ausländerfreundliche) Antwort gaben, bekannten sich unter den niedrig gebildeten Befragten 45% zu einer ausländerfeindlichen Einstellung (bzw. 55% zu einer ausländerfreundlichen). Die Annahme, dass sich der Anteil der ausländerfeindlichen „Ja“-Antworten in beiden Gruppen nicht voneinander unterscheidet, verschlechterte die Modellpassung signifikant, $\Delta G^2(1) = 5.81$, $p < .05$, musste also zurückgewiesen werden. Die Betrachtung der Antworten unter RRT-Bedingungen ergab, dass sich die hoch gebildeten Befragten unter Bedingungen erhöhter Anonymität mit $\pi_{\text{hoch gebildet}} = 30\%$ nur noch deskriptiv etwas weniger ausländerfeindlich als die niedrig gebildeten Befragten mit $\pi_{\text{niedrig gebildet}} = 38\%$ zeigten; signifikant war dieser Unterschied nicht, $\Delta G^2(1) = 1.00$, *ns*. Hinsichtlich ausländerfreundlicher Einstellungen trat jedoch ein bedeutsamer Bildungsunterschied auf: Die hoch

gebildeten Befragten wurden mit $\beta_{\text{hoch gebildet}} = 53\%$ signifikant häufiger als der ausländerfreundlichen Gruppe zugehörig klassifiziert als die niedrig gebildeten Befragten mit $\beta_{\text{niedrig gebildet}} = 24\%$, $\Delta G^2(1) = 4.74$, $p < .05$. Weitere Analysen zeigten, dass in beiden Gruppen Regelverweigerung in einem nicht zu vernachlässigenden Ausmaß auftrat, $\gamma_{\text{hoch gebildet}} = 17\%$, $\Delta G^2(1) = 13.23$, $p < .01$, bzw. $\gamma_{\text{niedrig gebildet}} = 38\%$, $\Delta G^2(1) = 67.29$, $p < .01$. Die niedrig gebildeten Teilnehmer entschlossen sich dabei deutlich häufiger dazu, die RRT-Regeln zu missachten, $\Delta G^2(1) = 6.86$, $p < .01$. Als untere Schranke für ausländerfeindliche Einstellungen unter den hoch gebildeten Befragten wurde $\pi_{\text{hoch gebildet}} = 30\%$ ermittelt, als obere Schranke $\pi_{\text{hoch gebildet}} + \gamma_{\text{hoch gebildet}} = 30\% + 17\% = 47\%$. Die entsprechenden Schranken bei den niedrig gebildeten Teilnehmern betrugen $\pi_{\text{niedrig gebildet}} = 38\%$ (untere Schranke) und $\pi_{\text{niedrig gebildet}} + \gamma_{\text{niedrig gebildet}} = 38\% + 38\% = 76\%$ (obere Schranke). Sie waren damit jeweils höher als bei den hoch Gebildeten. Für ausländerfreundliche Einstellungen bei den hoch gebildeten Teilnehmern ergab sich als untere Schranke $\beta_{\text{hoch gebildet}} = 53\%$ und $\beta_{\text{hoch gebildet}} + \gamma_{\text{hoch gebildet}} = 53\% + 17\% = 70\%$ als obere Schranke. Bei den niedrig gebildeten Befragten dagegen lag die untere Schranke bei $\beta_{\text{niedrig gebildet}} = 24\%$ und die obere bei $\beta_{\text{niedrig gebildet}} + \gamma_{\text{niedrig gebildet}} = 24\% + 38\% = 62\%$. Im Hinblick auf ausländerfreundliche Einstellungen waren also diese beiden Schranken bei den niedrig Gebildeten niedriger als bei den hoch Gebildeten.

Zusammenfassend legt das Ergebnismuster eine Interpretation des Bildungseffektes im Sinne eines wahren Einstellungsunterschiedes nahe: Hoch gebildete Befragte zeigten sich in Experiment II nicht nur in der direkten Befragung ausländerfreundlicher (75%) bzw. weniger ausländerfeindlich (25%) als niedrig gebildete Befragte (55% bzw. 45%); auch unter RRT-Bedingungen wurden die hoch gebildeten Teilnehmer häufiger als der eindeutig ausländerfreundlichen Gruppe zugehörig klassifiziert (53%) als die niedrig gebildeten (24%). Der Anteil der als eindeutig ausländerfeindlich klassifizierten Umfrageteilnehmer war bei den hoch gebildeten Befragten (30%) zwar nur deskriptiv niedriger als bei den niedrig gebildeten (38%); dies könnte jedoch durch den wesentlich höheren Verweigereranteil bei den niedrig gebildeten (38%) im Vergleich zu den hoch gebildeten Teilnehmern (17%) bedingt gewesen sein. Auch die Betrachtung der verschiedenen Schranken für ausländerfeindliche bzw. ausländerfreundliche Einstellungen in den beiden Subgruppen ist mit

der Annahme, dass es sich bei dem Einstellungseffekt um einen wahren Gruppenunterschied und nicht um ein Artefakt handelt, besser vereinbar.

Tabelle 4

Ausländerfeindliche und ausländerfreundliche Einstellungen in Abhängigkeit vom Befragungsmodus und vom Bildungsstand.

„Angenommen, Sie hätten eine 20 Jahre alte Tochter. Würde es Sie stören, wenn diese eine Beziehung mit einem nigerianischen Staatsbürger schwarzer Hautfarbe eingehen würde?"

(N = 606)

		Niedrig gebildet (n = 324)	Hoch gebildet (n = 282)
Direkte Befragung (n = 130)		(n = 65)	(n = 65)
	% „Ja"	45%	25%
	% „Nein"	55%	75%
Randomized-Response-Befragung (n = 476)		(n = 259)	(n = 217)
	Ehrliches „Ja" (π)	38%	30%
	Ehrliches „Nein" (β)	24%	53%
	Verweigerer (γ)	38%	17%
	$\Delta G^2(1)$: $\gamma_{\text{niedrig gebildet}} = 0$[†]	67.29**	
	$\Delta G^2(1)$: $\gamma_{\text{hoch gebildet}} = 0$[†]		13.23**
	$\Delta G^2(1)$: $\gamma_{\text{niedrig gebildet}} = \gamma_{\text{hoch gebildet}}$[‡]	6.86**	
	$\Delta G^2(1)$: % „Ja"$_{\text{niedrig gebildet}}$ = % „Ja"$_{\text{hoch gebildet}}$ (bzw. % „Nein"$_{\text{niedrig gebildet}}$ = % „Nein"$_{\text{hoch gebildet}}$)[††]	5.81*	
	$\Delta G^2(1)$: $\pi_{\text{niedrig gebildet}} = \pi_{\text{hoch gebildet}}$[‡‡]	1.00	
	$\Delta G^2(1)$: $\beta_{\text{niedrig gebildet}} = \beta_{\text{hoch gebildet}}$[‡‡]	4.74*	

Bemerkungen: [†]Hohe Werte bedeuten, dass sich die Modellpassung verschlechtert, wenn man annimmt, dass es keine Verweigerer in dieser Substichprobe gibt ($\gamma_{\text{niedrig gebildet}} = 0$ bzw. $\gamma_{\text{hoch gebildet}} = 0$). [‡]…dass sich der Anteil der Verweigerer unter den niedrig gebildeten Befragten in den RRT-Bedingungen ($\gamma_{\text{niedrig gebildet}}$) nicht vom Anteil der Verweigerer unter den hoch gebildeten Befragten in den RRT-Bedingungen ($\gamma_{\text{hoch gebildet}}$) unterscheidet. [††]…dass sich der Anteil der ausländerfeindlichen (% „Ja"$_{\text{niedrig gebildet}}$) bzw. ausländerfreundlichen Befragten (% „Nein"$_{\text{niedrig gebildet}}$) unter den niedrig gebildeten Befragten in der DB-Bedingung nicht vom Anteil der ausländerfeindlichen (% „Ja"$_{\text{hoch gebildet}}$) bzw. ausländerfreundlichen Befragten (% „Nein"$_{\text{hoch gebildet}}$) unter den hoch Gebildeten in der DB-Bedingung unterscheidet. [‡‡]…dass sich der Anteil der ausländerfeindlichen ($\pi_{\text{niedrig gebildet}}$) bzw. ausländerfreundlichen Befragten ($\beta_{\text{niedrig gebildet}}$) unter den niedrig gebildeten Befragten in den RRT-Bedingungen nicht vom Anteil der ausländerfeindlichen ($\pi_{\text{hoch gebildet}}$) bzw. ausländerfreundlichen Befragten ($\beta_{\text{hoch gebildet}}$) unter den hoch gebildeten Befragten in den RRT-Bedingungen unterscheidet. *$p < .05$, **$p < .01$.

5.3 *Experiment III*: Vergleich mit der projektiven Befragung

Experiment III galt der Beurteilung der Validität einer alternativen Methode zur Reduktion von Antwortverzerrungen mit Hilfe der multinomialen Verweigerer-detektionsvariante der RRT. Wie unter 2.3.2 beschrieben wurde, handelt es sich bei der Most-People-Technik (MPT; Alpert, 1971; Smith, 1954) um eine strukturierte projektive Befragungsmethode, die ähnlich wie die RRT versucht, Befragte zu ehrlicheren Antworten auf sensible Fragen zu bewegen, indem sie ihnen mehr Anonymität bietet. Dies wird im Rahmen der MPT verwirklicht, indem der Befragte gar nicht nach seinen eigenen Einstellungen, sondern nach denen der meisten anderen Menschen gefragt wird. Der psychoanalytischen Erklärung von Projektion folgend (Freud, 1938) wird anschließend jedoch aus der Antwort des Befragten auf seine eigene Einstellung geschlossen. Die zugrundeliegende Annahme des Modells ist durchaus umstritten; Smith hat bereits 1954 angemerkt, dass es naiv wäre, jedes Mal, wenn ein Befragter in der dritten Person spricht, anzunehmen, er gebe damit etwas über sich selbst preis. Man braucht jedoch nicht zwingend mit psychoanalytischen Annahmen zu arbeiten, um sich vorstellen zu können, dass die MPT Antwortverzerrungen besser kontrollieren könnte als eine direkte Befragung. Wird man danach gefragt, was die meisten Menschen über ein sensibles Thema denken, könnte es sein, dass man einfach darüber nachdenkt, wie viele Menschen man selbst kennt, die die jeweilige unerwünschte Meinung vertreten. Davon könnte man dann auf die meisten Menschen abstrahieren. Die Annahme einer Projektion der eigenen Einstellung auf die anderen würde damit unnötig (Miller, 1985; Bradburn & Sudman, 1979).[6] Da die Befragten damit tatsächlich nicht ihre eigene Einstellung preisgeben, ist auch nicht davon auszugehen, dass sie ihre Schätzungen verfälschen. Diese Annahme bringt jedoch ebenfalls Probleme mit sich (Krueger & Clement, 1994; Marks & Miller, 1987), da die Genauigkeit einer solchen Prävalenzschätzung stark durch das tatsächliche Wissen des Befragten über das Verhalten und die Einstellungen anderer sowie die Fähigkeit oder Bereitschaft des Befragten, von seinem Wissen auf andere zu abstrahieren, beschränkt ist: Menschen

[6] Genau dieses Prinzip versuchen nominative Techniken zu nutzen (vgl. 2.3.2).

schätzen häufig die Prävalenz insbesondere negativer Einstellungen und Verhaltensweisen bei anderen höher als bei sich selbst ein, beispielsweise infolge selbstwertdienlicher Verzerrung (Lewicki, 1983) oder aufgrund von Overconfidence (Svenson, 1981). Damit bleibt die Annahme, dass die MPT validere Prävalenz-schätzungen als direkte Befragungsmethoden liefert (vgl. Fisher, 1993; Snir & Harpaz, 2002), unabhängig von der theoretischen Herangehensweise problematisch. Obwohl die Technik in den vergangenen Jahrzehnten vielfach eingesetzt wurde, um eine Vielzahl von sozial unerwünschten oder vermeintlich wenig bewussten Einstellungen zu untersuchen (z.B. Davoli, Perucci, Sangalli, Brancato & Dell'Uomo, 1992; Jo, Nelson & Kiecker, 1997; Saenger & Gilbert, 1950; Yesalis & Courson, 1991), sind die Ergebnisse der wenigen Validierungsstudien aufgrund methodischer Mängel wenig überzeugend (vgl. Alpert, 1971; Weitz & Nickols, 1953). Die einzigen beiden Studien, in denen die MPT mit der RRT und einer direkten Befragung verglichen wurde, lieferten das gleiche Ergebnis: Die Schätzungen der MPT lagen weit über denen der direkten Befragung und der RRT. Interessanterweise interpretierten die Autoren die Ergebnisse jedoch unterschiedlich: Während Bégin und Boivin (1980) davon ausgingen, dass es sich bei den MPT-Schätzungen um Überschätzungen der wahren Prävalenz handelt, merkten Armacost, Hosseini, Morris und Rehbein (1991) an, dass es sich hierbei um realistische obere Schranken der Prävalenz der jeweiligen unerwünschten Einstellung handeln könnte. Da in beiden Untersuchungen ein älteres RRT-Modell verwendet wurde, das über keine Möglichkeit zur Aufdeckung von Regelverweigerern und damit zur Berechnung von Prävalenzmaxima verfügt, sind beide Interpretationen legitim. In Experiment III wurde ein multinomiales Modell der MPT und der Verweigererdetektionsvariante der RRT eingesetzt, um diese Kontroverse zu klären.

Als Gegenstand der Befragung wurde ein weiteres von Antwortverzerrungen geplagtes Forschungsfeld gewählt, nämlich die Untersuchung von Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung. Trotz der bekannten Problematik selbstberichteter Einstellungen zu diesem Thema (Deal, 2003; Hagler, Vargo & Semple, 1987; Weisel, Kravetz, Florian & Shurka-Zernitsky, 1988) bilden auch in diesem Gebiet klassische, auf Selbstbericht beruhende Einstellungsskalen den derzeitigen Forschungsstandard. Allerdings werden Rufe nach alternativen Erfassungs-

methoden (Strike, Skovholt & Hummel, 2004; Yazbeck, McVilly & Parmenter, 2004), wie z.B. der RRT (Antonak & Livneh, 1995, 2000), immer lauter.

In der vorliegenden Studie wurden mit Hilfe eines Online-Panels 1160 Menschen ohne körperliche oder geistige Behinderungen im WWW zum Thema Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung befragt. Wie in den beiden ersten Experimenten wurden die Befragten zunächst aufgefordert, demographische Fragen zu beantworten, gefolgt von Fragen zu ihren Erfahrungen mit und Einstellungen zu Menschen mit körperlicher und geistiger Behinderung. Das kritische Item stammte aus einer Untersuchung von Yazbeck et al. (2004) und wurde jedem Befragten einmal in einer Version mit körperlicher Behinderung („Fühlen Sie sich unwohl in der Anwesenheit von Menschen mit körperlicher Behinderung?“) und einmal in einer Version mit geistiger Behinderung („Fühlen Sie sich unwohl in der Anwesenheit von Menschen mit geistiger Behinderung?“) präsentiert. Die Befragten wurden in einem Verhältnis von 1:2:2:1 einer der vier Gruppen MPT (n = 200), RRT1 (niedrige Randomisierungswahrscheinlichkeit p_1, n = 383), RRT2 (hohe Randomisierungswahrscheinlichkeit p_2, n = 385) und direkte Befragung (n = 192) randomisiert zugeteilt. Auch in diesem Experiment unterschieden sich die vier Fragebogenvarianten lediglich durch das Format der kritischen Fragen. In der direkten Befragungsbedingung wurden die kritischen Fragen wie gewohnt ohne weitere Erklärungen gestellt. In der MPT-Bedingung wurden dieselben Fragen im projektiven Format gestellt, d.h. „Glauben Sie, dass sich die meisten Menschen unwohl in der Anwesenheit von Menschen mit körperlicher Behinderung fühlen?“ bzw. „Glauben Sie, dass sich die meisten Menschen unwohl in der Anwesenheit von Menschen mit geistiger Behinderung fühlen?“. In der RRT1-Gruppe wurden die Befragten aufgefordert, die kritische Frage zur körperlichen Behinderung inhaltsunabhängig mit „Ja“ zu beantworten, wenn Ihre Mutter im Februar bis April geboren wurde (p_1 = 3/12 = 1/4 = 0.25), und ansonsten ehrlich zu antworten. In der RRT2-Gruppe wurden sie dagegen um eine inhaltsunabhängige „Ja“-Antwort gebeten, wenn ihre Mutter im Januar oder Mai bis Dezember geboren wurde (p_2 = 1 – p_1 = 9/12 = 3/4 = 0.75), und um eine ehrliche Antwort, wenn ihre Mutter im Februar bis April geboren wurde. Um in jeder Gruppe eine weitere kritische Frage stellen zu können, ohne den wahren Status des Befragten zu enthüllen (vgl. Kulka, Weeks & Folsom, 1981), wurden für die kritische Frage zur geistigen Behinderung in jeder

Gruppe sowohl die Geburtsmonate (von Februar bis April zu Januar bis März) als auch die Person, deren Geburtstag für die Verschlüsselung von Relevanz war (von der Mutter der Befragten zu ihrem Vater), verändert. Wie einschlägige Geburtsstatistiken des Statistischen Bundesamts bestätigen, lagen auch hier die Randomisierungswahrscheinlichkeiten bei $p_1 = 0.25$ und $p_2 = 0.75$. Das entsprechende verbundene multinomiale Modell bestand somit in Experiment III aus vier Bäumen, je einem für die MPT-, RRT1-, RRT2- und die direkte Befragungsgruppe.

Die unabhängige Variable bildete in dieser Untersuchung erneut der Befragungsmodus mit den Stufen „RRT“ versus „MPT“ versus „direkte Befragung“. Die beiden abhängigen Variablen stellten die Prävalenz von negativen Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung dar; von besonderem Interesse waren hierbei die obere Schranke der jeweiligen RRT-Prävalenzschätzung sowie das 95%ige Konfidenzintervall der jeweiligen MPT-Schätzung. Für den Fall, dass die MPT die wahre Prävalenz von negativen Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung überschätzt, wurde erwartet, dass die obere Schranke der jeweiligen RRT-Prävalenzschätzung außer- und noch unterhalb des 95%igen Konfidenzintervalls der jeweiligen MPT-Schätzung liegen würde. Falls die MPT realistische obere Schranken für die Prävalenz von negativen Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung liefert, wurde erwartet, dass die obere Schranke der jeweiligen RRT-Prävalenzschätzung innerhalb des 95%igen Konfidenzintervalls der jeweiligen MPT-Schätzung liegen würde.

In Tabelle 5 finden sich die Parameterschätzungen für das saturierte Modell mit $G^2(0) = 0$. Hinsichtlich der kritischen Frage zur körperlichen Behinderung bekannten sich bei der direkten Befragung nur 8% der Befragten zu einer negativen Einstellung sowie $\pi_{\text{körperlich}} = 11\%$ unter RRT-Bedingungen. Unter MPT-Bedingungen fiel die geschätzte Prävalenz für eine negative Einstellung mit 55% deutlich höher aus, und zwar sowohl signifikant höher als die direkte Schätzung, $\Delta G^2(1) = 108.35$, $p < .01$, als auch signifikant höher als die RRT-Schätzung, $\Delta G^2(1) = 70.95$, $p < .01$. Wie schon in den beiden ersten Experimenten verweigerte mit $\gamma_{\text{körperlich}} = 33\%$ ein beträchtlicher Anteil der Befragten unter RRT-Bedingungen die Regeln bei der Frage zur körperlichen Behinderung, $\Delta G^2(1) = 88.83$, $p < .01$. Somit ergab sich als untere Schranke für die Prävalenz von negativen Einstellungen gegenüber Menschen mit körperlicher Behin-

derung $\pi_{\text{körperlich}} = 11\%$ und als obere Schranke $\pi_{\text{körperlich}} + \gamma_{\text{körperlich}} = 11\% + 33\% = 44\%$. Diese obere Schranke lag leicht unterhalb des 95%igen Konfidenzintervalls der MPT-Schätzung (48-61%). In Bezug auf die kritische Frage zur geistigen Behinderung waren die Ergebnisse ähnlich, jedoch noch deutlicher ausgeprägt: Unter direkten Befragungsbedingungen bekannten sich ähnlich viele Befragte (27%) zu einer negativen Einstellung wie unter RRT-Bedingungen ($\pi_{\text{geistig}} = 24\%$), während die mittels MPT geschätzte Prävalenz mit 79% sowohl die Schätzung der direkten Befragung, $\Delta G^2(1) = 111.55$, $p < .01$, als auch die RRT-Schätzung, $\Delta G^2(1) = 109.28$, $p < .01$, signifikant übertraf. Auch bei der Frage zur geistigen Behinderung unterschied sich der Verweigereranteil mit $\gamma_{\text{geistig}} = 22\%$ bedeutsam von Null, $\Delta G^2(1) = 38.28$, $p < .01$. Dementsprechend betrug die untere Schranke der RRT-Schätzung $\pi_{\text{geistig}} = 24\%$ und die obere Schranke $\pi_{\text{geistig}} + \gamma_{\text{geistig}} = 24\% + 22\% = 46\%$. Auch hier lag die obere Schranke – diesmal deutlich – unterhalb des 95%igen Konfidenzintervalls der MPT-Schätzung für negative Einstellungen gegenüber Menschen mit geistiger Behinderung (73-85%).

Die Ergebnisse von Experiment III bestätigen die Vermutung von Bégin und Boivin (1980), die argumentiert haben, dass die MPT zur Reduktion von Antwortverzerrungen wenig geeignet ist. Offenbar wird durch die MPT nicht nur einer möglichen Verzerrung entgegengewirkt, vielmehr kann es zu einer Überschätzung der Prävalenz des sensiblen Merkmals kommen. Anders als von Armacost et al. (1991) erhofft ermöglicht es die MPT nicht, eine obere Schranke für die Prävalenz sensibler Merkmale zu bestimmen. Sowohl bei der kritischen Frage zum Umgang mit Menschen mit körperlicher Behinderung (48%) als auch besonders bei der kritischen Frage zur geistigen Behinderung (73%) lag die untere Grenze des 95%igen Konfidenzintervalls der MPT-Schätzung oberhalb der oberen Schranke der jeweiligen RRT-Schätzung (44% bzw. 46%).

Tabelle 5

Negative Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung in Abhängigkeit vom Befragungsmodus.

„Fühlen Sie sich unwohl in der Anwesenheit von Menschen mit körperlicher (geistiger) Behinderung?“

(N = 1160)

		Körperliche Behinderung	Geistige Behinderung
Direkte Befragung (n = 192)			
	% „Ja“	8%	27%
	% „Nein“	92%	73%
Randomized-Response-Befragung (n = 768)			
	Ehrliches „Ja“ (π)	11%	24%
	Ehrliches „Nein“ (β)	56%	54%
	Verweigerer (γ)	33%	22%
	Obere Schranke ($\pi + \gamma$)	*44%*	*46%*
Most-People-Befragung (n = 200)			
	% „Ja“	55%	79%
	95%iges Konfidenzintervall	*48-61%*	*73-85%*
	% „Nein“	45%	21%
	$\Delta G^2(1)$: $\gamma = 0$[†]	88.83**	38.28**
	$\Delta G^2(1)$: % MPT „Ja“ = % DB „Ja“[‡]	108.35**	111.55**
	$\Delta G^2(1)$: % MPT „Ja“ = π[††]	70.95**	109.28**
	$\pi + \gamma \notin$ 95% Konfidenzintervall$_{MPT}$, wobei $\pi + \gamma < M_{MMPQ} - 1.96 \cdot SE_{MPT}$?[‡‡]	Ja	Ja

Bemerkungen: [†]Hohe Werte bedeuten, dass sich die Modellpassung verschlechtert, wenn man annimmt, dass es keine Verweigerer in der Stichprobe gibt ($\gamma = 0$). [‡]...dass sich der Anteil der behindertenfeindlichen Befragten in der MPT-Bedingung (% MPT „Ja“) nicht vom Anteil der behindertenfeindlichen Befragten in der DB-Bedingung (% DB „Ja“) unterscheidet. [††]…dass sich der Anteil der behindertenfeindlichen Befragten in der MPT-Bedingung (% MPT „Ja“) nicht vom Anteil der behindertenfeindlichen Befragten in den RRT-Bedingungen (π) unterscheidet. [‡‡]Liegt die obere Schranke der RRT-Prävalenzschätzung unterhalb des 95%igen Konfidenzintervalls der MPT-Schätzung? **p < .01.

5.4 *Experiment IV*: Antwortsymmetrie und Verweigererrate

In Experiment IV wurde geprüft, ob sich die multinomial modellierte Verweigererdetektionsvariante der RRT durch eine geeignete Adaptation des Befragungsmodells so verändern lässt, dass die Verweigererrate reduziert wird.

Unter 2.3.4 wurde bereits dargestellt, dass RRT-Modelle in der Vergangenheit wegen ihrer Anfälligkeit für die Nichtbefolgung der Regeln kritisiert wurden. Einen Ansatz, mit diesem Problem umzugehen, haben Clark und Desharnais (1998) mit der Entwicklung des Verweigererdetektionsmodells vorgeschlagen. Dieses ermöglicht, das Ausmaß der Regelverweigerung zu erfassen und bei der Schätzung der Prävalenz des kritischen Merkmals zu berücksichtigen. Wünschenswert wäre natürlich, den Verweigereranteil von vornherein klein zu halten. Um das zu erreichen, muss man die möglichen Gründe für eine Nichtbefolgung der RRT-Instruktionen untersuchen. In der RRT-Literatur wird hierzu üblicherweise zwischen *respondent jeopardy* und *risk of suspicion* unterschieden (Antonak & Livneh, 1995). Unter *respondent jeopardy* versteht man die Befürchtung der Merkmalsträger, bei einer „Ja"-Antwort als Merkmalsträger identifizierbar zu sein, was zur Regelverweigerung führen kann. Diese Befürchtung und damit mutmaßlich auch der Verweigereranteil lassen sich reduzieren, indem man anstatt einer hohen Wahrscheinlichkeit, die kritische Frage ehrlich beantworten zu müssen, eine Randomisierungswahrscheinlichkeit in der Nähe von 0.50 wählt. Bei einer solchen Randomisierungswahrscheinlichkeit herrscht maximale Unsicherheit darüber, ob eine „Ja"-Antwort vom Zufallsgenerator erzeugt wurde oder auf den wahren Merkmalsstatus hinweist, womit der Merkmalsträger optimal geschützt ist. Allerdings führt eine solche Wahl der Randomisierungswahrscheinlichkeit zu einer Verringerung der Effizienz: Je mehr Antworten durch die Zufallsverschlüsselung „verloren" gehen, desto höher wird die Varianz der Parameterschätzung und desto niedriger ihre Effizienz (Antonak & Livneh, 1995). Der Reduktion des Verweigereranteils durch eine Reduktion der respondent jeopardy mit Hilfe einer Manipulation der Randomisierungswahrscheinlichkeit sind also praktische Grenzen gesetzt.

Als *risk of suspicion* bezeichnet man die Befürchtung von Nicht-Merkmalsträgern, bei einer „Ja"-Antwort fälschlich mit dem kritischen Merkmal in Verbindung gebracht zu werden. Dies kann – wie die respondent jeopardy bei den

Merkmalsträgern – ebenfalls zur Regelverweigerung führen. Von Relevanz ist dies besonders bei Forced-Response-Modellen der RRT, da dort Befragte mit einer bestimmten Wahrscheinlichkeit zu inhaltsunabhängigen „Ja"-Antworten aufgefordert werden, was Nicht-Merkmalsträgern nachgewiesenermaßen schwer fallen kann (Lensvelt-Mulders & Boeije, 2007). Zur Reduktion des risk of suspicion – und damit zur Reduktion des Verweigereranteils – hat Bourke (1984) vorgeschlagen, Antwortsymmetrie herzustellen. Ein RRT-Modell ist nach Bourke (1984) antwortsymmetrisch, wenn keine der möglichen Antworten – also weder eine „Ja"- noch eine „Nein"-Antwort – einen eindeutigen Rückschluss auf den wahren Status des Befragten zulässt. Bei Verwendung eines antwortsymmetrischen Modells sollte es Nicht-Merkmalsträgern leichter fallen, der Aufforderung, unabhängig vom Inhalt der kritischen Frage mit „Ja" zu antworten, nachzukommen.

Wendet man die obige Definition von Antwortsymmetrie auf das Forced-Response-Modell von Dawes und Moore (1980) an, das der Verweigererdetektionsvariante von Clark und Desharnais (1998) zugrundeliegt, so wird ersichtlich, dass es sich um ein asymmetrisches Design handelt: Eine „Nein"-Antwort charakterisiert den Befragten zweifelsfrei als Nicht-Merkmalsträger. Dadurch ist der Anreiz, trotz der Aufforderung, inhaltsunabhängig mit „Ja" zu antworten, „Nein" zu sagen, hoch. Morton (beschrieben in Greenberg et al., 1969) hat eine symmetrische Variante des Forced-Response-Modells entwickelt: In seiner Abwandlung werden die Befragten je nach Ausgang des Randomisierungsprozesses aufgefordert, entweder inhaltsunabhängig mit „Ja" (p_{ja}) oder mit „Nein" (p_{nein}) oder ehrlich ($1 - p_{ja} - p_{nein}$) zu antworten. Dieses Modell ist insofern symmetrisch, als dass eine „Nein"-Antwort nicht mehr eindeutig ist, da sie sowohl von einem Merkmals- als auch von einem Nicht-Merkmalsträger stammen kann. Dadurch sollte für Nicht-Merkmalsträger der Anreiz, auf „Nummer sicher" zu gehen und mit „Nein" zu antworten – also die Befolgung der RRT-Regeln zu verweigern, geringer werden.

Im vorliegenden Experiment wurde das Verweigererdetektionsmodell an das symmetrische Morton-Modell adaptiert und mit dem asymmetrischen Dawes & Moore-Modell, welches den ersten drei Einzelarbeiten zugrundelag, im Hinblick auf seine Fähigkeit verglichen, den Verweigereranteil zu reduzieren. Befragt wurden 2254 chinesische Studenten zum Thema Prüfungsbetrug, das in früheren RRT-Unter-

suchungen wiederholt als anfällig für Antwortverzerrungen identifiziert werden konnte (vgl. Dawes & Moore, 1980; Kerkvliet, 1994). Die kurzen Fragebögen bestanden aus wenigen demographischen Fragen sowie einer kritischen Frage zum Thema Prüfungsbetrug. Diese lautete: „Haben Sie während Ihrer Schulzeit oder Ihres Studiums schon einmal in einer Prüfung betrogen?“. Die Studenten wurden randomisiert und gleichanteilig einer der fünf Gruppen RRT1 (asymmetrisches Dawes & Moore-Modell, niedrige Randomisierungswahrscheinlichkeit p_1, n = 449), RRT2 (asymmetrisch, hohe Randomisierungswahrscheinlichkeit p_2, n = 452), RRT3 (symmetrisches Morton-Modell, niedrige Randomisierungswahrscheinlichkeiten p_3 und p_4, n = 451), RRT4 (symmetrisch, hohe Randomisierungswahrscheinlichkeiten p_5 und p_6, n = 439) und direkte Befragung (n = 463) zugewiesen. In der direkten Kontrollbedingung wurden die Befragten wie gewohnt aufgefordert, die kritische Frage mit „Ja“ oder „Nein“ zu beantworten. In der RRT1-Gruppe wurden die Teilnehmer aufgefordert, die kritische Frage inhaltsunabhängig mit „Ja“ zu beantworten, wenn sie im Januar oder Juli geboren wurden, und ansonsten ehrlich zu antworten. Aufgrund der nicht über alle Monate ganz gleich verteilten Geburtshäufigkeiten in der Volksrepublik (VR) China betrug p_1 in dieser Gruppe 0.16, wie einschlägige Daten des Ministeriums für Statistik der VR China nachweisen. In der entsprechenden RRT2-Gruppe sollten die Teilnehmer die kritische Frage zum Prüfungsbetrug inhaltsunabhängig mit „Ja“ beantworten, wenn sie nicht im Januar oder Juli geboren wurden ($p_2 = 1 - p_1 = 0.84$), und ehrlich antworten, wenn sie im Januar oder Juli geboren wurden. In der RRT3-Gruppe lautete die Instruktion, inhaltsunabhängig mit „Ja“ zu antworten, wenn man im Januar geboren wurde (p_3 = 0.09), inhaltsunabhängig mit „Nein“ zu antworten, wenn man im Juli geboren wurde (p_4 = 0.07) und ehrlich zu antworten, wenn man in einem der anderen Monate geboren wurde. Schließlich lautete in der RRT4-Gruppe die Aufforderung, die kritische Frage unabhängig vom Inhalt mit „Ja“ zu beantworten, wenn man im Februar bis Juni geboren wurde (p_5 = 0.37), mit „Nein“ zu beantworten, wenn man im August bis Dezember geboren wurde (p_6 = 0.47), und ansonsten ehrlich zu antworten. Das verbundene multinomiale Modell, das dieses Design repräsentierte, bestand somit aus fünf Bäumen, d.h. je einem für die vier RRT-Gruppen und einem für die direkte Befragungsgruppe.

Die unabhängige Variable stellte in Experiment IV der Befragungsmodus mit den Stufen „asymmetrische RRT“ versus „symmetrische RRT“ versus „direkte Befragung“ dar. Bei dem Verweigereranteil sowie dem Anteil der ehrlichen Merkmalsträger bzw. Nicht-Merkmalsträger handelte es sich um die abhängigen Variablen. Die Hypothesen lauteten, dass die symmetrische RRT-Variante von Morton im Vergleich zu der asymmetrischen Dawes & Moore-RRT-Variante den Verweigereranteil reduzieren sollte (Hypothese 1). Ferner wurde erwartet, dass der reduzierte Verweigereranteil mit einem erhöhten Anteil an ehrlichen Nicht-Merkmalsträgern einhergeht, während der Anteil der ehrlichen Merkmalsträger von der Manipulation unbeeinflusst bleibt (Hypothese 2).

Tabelle 6 zeigt die Ergebnisse von Experiment IV. Direkt befragt gaben 50% der Studenten zu, schon einmal bei einer Prüfung während der Schulzeit oder während des Studiums betrogen zu haben. Unter RRT-Bedingungen lag der Anteil mit $\pi = 54\%$ in einem ähnlichen Bereich. Jedoch hielt sich auch mit $\gamma = 20\%$ ein bedeutsamer Teil der Befragten nicht an die RRT-Regeln, $\Delta G^2(1) = 116.26$, $p < .01$. Die restlichen $\beta = 26\%$ wurden als Nicht-Prüfungsbetrüger identifiziert. Bei getrennter Betrachtung der mittels des asymmetrischen Dawes & Moore-RRT-Modells befragten Studenten wurde deutlich, dass die Verhältnisse dort nicht stark von dem Muster in der Gesamtstichprobe abwichen; $\pi_{\text{Dawes \& Moore}} = 52\%$ bekannten sich zu Prüfungsbetrug, während $\gamma_{\text{Dawes \& Moore}} = 21\%$ keine regelkonforme Antwort gaben. Der Anteil der nicht regelkonform Antwortenden unterschied sich bedeutsam von Null, $\Delta G^2(1) = 105.05$, $p < .01$. Der Anteil der Studenten, der noch nie bei einer Prüfung betrogen hat, wurde auf $\beta_{\text{Dawes \& Moore}} = 27\%$ geschätzt. Von den mit dem symmetrischen Morton-RRT-Modell befragten Studenten gaben sich $\pi_{\text{Morton}} = 54\%$ als Prüfungsbetrüger zu erkennen. Diese Schätzung unterschied sich hypothesenkonform nicht signifikant von der Schätzung des Dawes & Moore-RRT-Modells, $\Delta G^2(1) = 0.21$, *ns*. Mit $\gamma_{\text{Morton}} = 7\%$ verweigerten jedoch wie erwartet im symmetrischen Morton-Modell signifikant weniger Studenten eine regelkonforme Antwort als im asymmetrischen Dawes & Moore-Modell, $\Delta G^2(1) = 4.12$, $p < .05$. Der Verweigereranteil war damit so gering, dass er sich nicht signifikant von Null unterschied, $\Delta G^2(1) = 1.09$, *ns*. Die restlichen $\beta_{\text{Morton}} = 39\%$ wurden im Rahmen des Modells als Nicht-Merkmalsträger klassifiziert, also als Studenten, die noch nie bei einer Prüfung betrogen haben. Diese Schätzung war erwartungsgemäß, allerdings nur

deskriptiv höher als die entsprechende Schätzung mit dem Dawes & Moore- Modell, $\Delta G^2(1) = 1.74$, *ns*.

Tabelle 6

Lebenszeitprävalenz von Prüfungsbetrug in Abhängigkeit vom Befragungsmodus.

„Haben Sie während Ihrer Schulzeit oder Ihres Studiums schon einmal in einer Prüfung betrogen?" (*N* =2254)				
Direkte Befragung (*n* = 463)				
	% „Ja"	50%		
	% „Nein"	50%		
Randomized-Response-Befragung gesamt (*n* = 1791)			*Randomized-Response asymmetrisch Dawes & Moore* (*n* = 901)	*Randomized-Response symmetrisch Morton* (*n* = 890)
	Ehrliches „Ja" (π)	54%	52%	54%
	Ehrliches „Nein" (β)	26%	27%	39%
	Verweigerer (γ)	20%	21%	7%
	$\Delta G^2(1)$: $\gamma = 0^{\dagger}$	111.26**		
	$\Delta G^2(1)$: $\gamma_{\text{Dawes \& Moore}} = 0^{\ddagger}$		105.05**	
	$\Delta G^2(1)$: $\gamma_{\text{Morton}} = 0^{\ddagger}$			1.09
	$\Delta G^2(1)$: $\gamma_{\text{Dawes \& Moore}} = \gamma_{\text{Morton}}^{\dagger\dagger}$		4.12*	
	$\Delta G^2(1)$: $\pi_{\text{Dawes \& Moore}} = \pi_{\text{Morton}}^{\ddagger\ddagger}$		0.21	
	$\Delta G^2(1)$: $\beta_{\text{Dawes \& Moore}} = \beta_{\text{Morton}}^{\dagger\dagger\dagger}$		1.74	

Bemerkungen: †Hohe Werte bedeuten, dass sich die Modellpassung verschlechtert, wenn man annimmt, dass es keine Verweigerer in der Gesamtstichprobe gibt ($\gamma = 0$). ‡…dass sich die Modellpassung verschlechtert, wenn man annimmt dass es keine Verweigerer in der jeweiligen Substichprobe gibt ($\gamma_{\text{Dawes \& Moore}} = 0$ bzw. $\gamma_{\text{Morton}} = 0$). ††…dass sich der Anteil der Verweigerer unter dem Dawes & Moore-RRT-Modell ($\gamma_{\text{Dawes \& Moore}}$) nicht vom Anteil der Verweigerer unter dem Morton-RRT-Modell (γ_{Morton}) unterscheidet. ‡‡…dass sich der Anteil der Prüfungsbetrüger unter dem Dawes & Moore-RRT-Modell ($\pi_{\text{Dawes \& Moore}}$) nicht vom Anteil der Prüfungsbetrüger unter dem Morton-RRT-Modell (π_{Morton}) unterscheidet. †††…dass sich der Anteil der Nicht-Prüfungsbetrüger unter dem Dawes & Moore-RRT-Modell ($\beta_{\text{Dawes \& Moore}}$) nicht vom Anteil der Nicht-Prüfungsbetrüger unter dem Morton-RRT-Modell (β_{Morton}) unterscheidet. *$p < .05$, **$p < .01$.

Das Ergebnismuster von Experiment IV legt nachdrücklich nahe, dass die Verwendung eines antwortsymmetrischen RRT-Modells den Verweigereranteil im Vergleich zu einem asymmetrischen RRT-Modell bedeutsam zu reduzieren vermag. Im vorliegenden Beispiel erreichte der Verweigereranteil bei antwortsymmetrischer Befragung nur noch einen Wert, der zufallskritisch nicht mehr von Null unterscheidbar war. Diese Reduktion ging nicht mit einer Veränderung der Prävalenzrate der ehrlichen Merkmalsträger einher, die von solch einer Manipulation nicht betroffen sein sollten, sondern mit einer – wenn auch nur deskriptiv erkennbaren – Erhöhung der Prävalenzrate der ehrlichen Nicht-Merkmalsträger: Diesen fiel es offensichtlich leichter, die kritische Frage zum Prüfungsbetrug inhaltsunabhängig zu bejahen. Denn durch das Wissen darüber, dass auch inhaltsunabhängige „Nein“-Antworten vorkommen können, war der Anreiz, mit einer „Nein“-Antwort auf Nummer sicher zu gehen, geringer als in dem entsprechenden asymmetrischen RRT-Design.

6 Diskussion

In der Übersicht haben die hier vorgestellten Einzelarbeiten gezeigt, dass die Verweigererdetektionsvariante der RRT nicht nur online, sondern auch in Papier-und-Bleistift-Untersuchungen (Experiment I, II, IV) sowie in einer Vielzahl von unterschiedlichen sensiblen Themenbereichen (Experiment I bis IV) erfolgreich eingesetzt werden kann. Außerdem demonstrierten die Ergebnisse, dass das dank seiner multinomialen Reformulierung flexibel erweiterbare Grundmodell sowohl zur Untersuchung von inhaltlich interessanten Gruppenunterschieden (Experiment II) als auch zum Vergleich der Verweigererdetektionsvariante mit konkurrierenden Methoden zur Reduktion von Antwortverzerrungen (Experiment III) bzw. verbesserten Abwandlungen der Verweigererdetektionsvariante selbst (Experiment IV) gewinnbringend verwendet werden kann.

Wie bei der Herleitung der Fragestellung der vorliegenden Dissertation betont wurde, handelt es sich bei der Verweigererdetektionsvariante der RRT um ein sowohl im Vergleich zu konventionellen direkten Befragungen, aber auch im Vergleich zu konventionellen RRT-Modellen relativ wenig effizientes Verfahren, so dass es überzeugender Vorzüge zur Kompensation dieses Nachteils bedarf. Die Einzelarbeiten haben gezeigt, dass solche Vorzüge existieren, aber durchaus noch ein Verbesserungspotential und weiterer Forschungsbedarf bestehen.

So fällt zunächst auf, dass in allen Studien relativ häufig verweigert wurde (7% bis 47%) und dass die Verweigerer sowohl bei einer direkten Befragung als auch bei einer Befragung mittels eines älteren RRT-Modells unentdeckt geblieben wären. Dies hätte die Prävalenzschätzungen in dem Ausmaß verfälscht, in dem es sich bei den Verweigerern um Merkmalsträger handelte. Diesen Umstand berücksichtigend bietet die multinomiale Verweigererdetektionsvariante die Möglichkeit zur Bestimmung der Verweigererrate und damit einer oberen Schranke für die Prävalenz des sensiblen Merkmals. Eine solche obere Schranke versuchen zwar auch andere, teils effizientere Methoden, wie z.B. die projektive MPT, anzugeben; die Ergebnisse von Experiment III belegen jedoch, dass zumindest die projektive MPT zur Bestimmung einer oberen Schranke nicht geeignet ist, weil es bei ihrer Verwendung zu Überschätzungen der

Prävalenz kommen kann. Experiment IV hat gezeigt, dass sich sogar die Schätzungen des Verweigererdetektionsmodells durch Einführung von Antwortsymmetrie noch weiter verbessern lassen, da dadurch der Verweigereranteil reduziert werden kann.

Als Fazit der vorliegenden Dissertation lässt sich festhalten, dass der Einsatz des symmetrischen Verweigererdetektionsmodells anstelle effizienterer RRT-Modelle ohne Verweigererdetektion immer dann indiziert erscheint, wenn der Verweigereranteil und eine obere Schranke für die Prävalenz des sensiblen Merkmals bestimmt werden sollen. Die Ermittlung eines solchen Worst-Case-Szenarios, welches mit Hilfe konventioneller RRT-Modelle nicht erfasst werden kann, kann beispielsweise bei der Erforschung neuer oder bisher wenig untersuchter Dunkelfelder von großem Interesse sein. Die Ergebnisse von Experiment IV sprechen allerdings dafür, dass es sich in einer asymmetrischen Befragung bei den Verweigerern zu einem erheblichen Teil um Nicht-Merkmalsträger handeln könnte. Wenn diese aus Angst, durch eine inhaltsunabhängige „Ja“-Antwort mit dem kritischen Merkmal fälschlich in Verbindung gebracht zu werden, die Befolgung der RRT-Instruktionen verweigern, liegt die mit Hilfe des Verweigerer-detektionsmodells bestimmte untere Schranke näher am wahren Wert der Prävalenz als die obere Schranke. Wann immer dies der Fall ist, brächte der Einsatz der Verweigererdetektionsvariante keine Vorteile außer der prinzipiellen Möglichkeit zur Verweigererdetektion mit sich; der Rückgriff auf ein konventionelles, aber effizienteres symmetrisches RRT-Modell wäre dann zumindest unter pragmatischen Gesichtspunkten vertretbar. Eine eindeutige Antwort darauf, welches Befragungsmodell im Nachhinein vorzuziehen gewesen wäre, können allenfalls „harte“ Validierungsstudien geben, in denen der wahre Merkmalsstatus aller Befragten mit hohem Aufwand individuell ermittelt wird; dies erübrigt dann allerdings gleichzeitig die Durchführung einer zusätzlichen, auf Selbstauskünften beruhenden Umfrage.

7 Ausblick

Wie im vorangehenden Kapitel angedeutet wurde, hat die vorliegende Dissertation viele Fragen bezüglich der multinomial modellierten Verweigererdetektionsvariante der RRT beantworten können. Dabei wurden aber auch neue Fragen aufgeworfen, die im Rahmen dieser Arbeit nicht beantwortet werden konnten.

Das symmetrische Verweigererdetektionsmodell hat in Experiment IV zu einer Reduktion der Verweigererrate geführt. In Folgeuntersuchungen sollte geprüft werden, wie gut diese symmetrische RRT-Verweigererdetektionsvariante im Vergleich zu alternativen und eventuell noch besseren oder zumindest einfacheren Methoden zur Reduktion von Antwortverzerrungen, beispielsweise der Unmatched-Count-Technik, abschneidet. Die Vorteile der Unmatched-Count- gegenüber der Randomized-Response-Technik liegen in ihrer leichten Kommunizierbarkeit und vereinfachten Durchführung, die Nachteile in der Schwierigkeit, die Parameter effizient zu schätzen, und der fehlenden Möglichkeit zur Schätzung des Verweigereranteils. Diese Nachteile könnten aber für den Fall, dass bei der vermeintlich transparenteren und leichter zu verstehenden Unmatched-Count-Technik der Verweigereranteil entsprechend niedriger ist, weniger stark ins Gewicht fallen.

Die Frage zur relativen Güte der Verweigererdetektionsvariante sowohl im Vergleich zu anderen RRT-Modellen als auch im Vergleich zu nicht auf der RRT beruhenden Verfahren lässt sich am besten in harten Validierungsstudien beantworten, in denen die wahre Prävalenz des kritischen Merkmals bekannt ist. Wie mehrfach angedeutet wurde, sind solche Studien jedoch aus verschiedenen Gründen schwierig durchzuführen. Dennoch könnte sich der zusätzliche Aufwand lohnen. Studien, in denen der Ausgang des Randomisierungsprozesses bekannt ist – als eine zweite Art von harten Validierungsstudien – erlauben zwar, die Effektivität von verschiedenen Interventionsstrategien zur Verweigererreduktion innerhalb eines RRT-Modells zu untersuchen; sie sind jedoch zur Beantwortung der an dieser Stelle interessanten Fragen nicht geeignet.

Sollte sich die symmetrische multinomiale Verweigererdetektionsvariante bei den vorgeschlagenen harten Validierungsstudien als die beste Methode zur Ver-

weigererdetektion erweisen, wären neue Studien zu ihrer weiteren Verbesserung wünschenswert. Hier scheint insbesondere der Ansatz von Interesse, das Versuchsdesign um eine zusätzliche dritte Gruppe zu erweitern. Auf diese Weise wird nämlich eine mögliche Verletzung der Modellvoraussetzungen – insbesondere eine bedingungsabhängig unterschiedliche Verweigererrate – testbar. Morten Moshagen und Kollegen überprüfen derzeit in Computersimulationen die Eigenschaften einer solchen Modellerweiterung.

Der Selbstbericht der Befragten braucht auch im Lichte der vorliegenden Ergebnisse nicht als die von vielen Forschern bevorzugte – und häufig auch einzige zur Verfügung stehende – Datenquelle verworfen zu werden. Die Ergebnisse zeigen jedoch deutlich, dass zur Kontrolle von Antwortverzerrungen geeignete Maßnahmen getroffen werden müssen. Der Rückgriff auf konventionelle, direkte Befragungsmethoden erscheint angesichts der vorliegenden Untersuchungen nicht länger zu rechtfertigen, wenn aufgrund sensibler Inhalte Antwortverzerrungen sicher erwartet werden können. Die hier untersuchte, multinomial modellierte, symmetrische Verweigererdetektionsvariante der RRT bietet eine nützliche Möglichkeit, Antwortverzerrungen modellbasiert zu quantifizieren und auf diese Weise zu kontrollieren.

8 Zusammenfassende Thesen

- Ein direkter Selbstbericht liefert bei Befragungen zu sensiblen Themen Antworten, die aufgrund sozialer Erwünschtheit verzerrt sind. Die Prävalenz sensibler Merkmale wird dadurch unterschätzt.
- Die Randomized-Response-Technik (RRT) ist eine geeignete Methode zur Reduktion solcher Antwortverzerrungen und damit zur Erhöhung der Validität der Prävalenzschätzung bei sensiblen Merkmalen. Sie ist jedoch anfällig für Verweigerer, also Befragte, die sich nicht an die zuweilen wenig intuitiven und nicht ohne weitere Erläuterung verständlichen RRT-Regeln halten. Sofern es sich bei den Regelverweigerern um Träger des sensiblen Merkmals handelt, unterschätzt auch die RRT die wahre Prävalenz des Merkmals.
- Die Verweigererdetektionsvariante der RRT von Clark und Desharnais (1998) ist ein vielversprechender Ansatz zur quantitativen Erfassung des relativen Anteils der Verweigerer. Als multinomiales Modell formuliert kann sie flexibel an unterschiedliche Befragungskontexte angepasst und zur Parameterschätzung sowie Hypothesenprüfung verwendet werden.
- Die Verweigererdetektionsvariante ist nicht nur wie in den bisherigen Vorarbeiten online erfolgreich einsetzbar, sondern auch in Papier-und-Bleistift-Untersuchungen (Experiment I, II und IV) zu ganz unterschiedlichen sensiblen Themen (Experiment I bis IV).
- Mit Hilfe multinomialer Erweiterungen des Verweigererdetektionsmodells lässt sich zeigen, dass Non-Compliance bei der Medikamenteneinnahme ein größeres Problem darstellt als ein direkter Selbstbericht nahe legen würde (Experiment I), dass der Bildungseffekt bei ausländerfeindlichen Einstellungen ein echter Einstellungsunterschied und kein Artefakt ist (Experiment II), dass die projektive Most-People-Technik zu Überschätzungen der Prävalenz negativer Einstellungen gegenüber Menschen mit körperlicher und geistiger Behinderung führt (Experiment III), und dass eine symmetrische Formulierung der multinomialen Verweigererdetektionsvariante in der Lage ist, den Verweigereranteil bis auf ein Minimum zu reduzieren (Experiment IV).

Literaturverzeichnis

Abraham, H.H.L. (1966). Social distance and patterns of prejudice in Germany and Sweden. *Archiv für die Gesamte Psychologie, 118*, 229-252.

Alpert, M. (1971). Identification of determinant attributes: A comparison of methods. *Journal of Marketing Research, 8*, 184-191.

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*, 99-145.

Antonak, R.F. & Livneh, H. (2000). Measurement of attitudes towards persons with disabilities. *Disability and Rehabilitation, 22*, 211-224.

Armacost, R.L., Hosseini, J.C., Morris, S.A. & Rehbein, K.A. (1991). An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information. *Decision Sciences, 22*, 1073-1090.

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

Bégin, G. & Boivin, M. (1980). Comparison of data gathered on sensitive questions via direct questionnaire, randomized response technique, and a projective method. *Psychological Reports, 47*, 743-750.

Bergmann, W. & Erb, R. (1991). *Antisemitismus in der Bundesrepublik Deutschland. Ergebnisse der empirischen Forschung von 1946-1989*. Opladen: Leske + Budrich.

Bogardus, E.S. (1925). Measuring social distance. *Journal of Applied Sociology, 9*, 299-308.

Bogardus, E.S. (1933). A social distance scale. *Sociology and Social Research, 17*, 265-271.

Bourke, P.D. (1984). Estimation of proportions using symmetric randomized response designs. *Psychological Bulletin, 96*, 166-172.

Bradburn, N.M. & Sudman, S. (1979). *Improving interview method and questionnaire design. Response effects to threatening questions in survey research*. San Francisco: Jossey-Bass Publishers.

Campbell, A.A. (1987). Randomized response technique. *Science, 236*, 1049.

Campbell, C. & Joiner, B.L. (1973). How to get the answer without being sure you've asked the question. *American Statistician, 27*, 229-231.

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.

Cobb, M.D. (2002). *Unobtrusively measuring racial attitudes: The consequences of social desirability effects*. Unveröffentlichte Doktorarbeit. Urbana-Champaign: University of Illinois.

Cohen Shabat, M. (1993). Prejuicio etnico en estudiantes universitarios. [Ethnische Vorurteile bei Universitätsstudenten] *Revista Mexicana de Psicologia, 10*, 183-188.

Davoli, M., Perucci, C.A., Sangalli, M., Brancato, G. & Dell'Uomo, G. (1992). Reliability of sexual behavior data among high school students in Rome. *Epidemiology, 3*, 531-535.

Dawes, R.M. & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen. In F. Petermann (Hrsg.), *Einstellungsmessung, Einstellungsforschung* (S. 117-133). Göttingen: Hogrefe.

Dawes, R.M. & Smith, T.L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology, volume 1* (pp. 509-566). New York: Random House.

Deal, M. (2003). Disabled people's attitudes toward other impairment groups: A hierarchy of impairments. *Disability and Society, 18*, 897-910.

DiMatteo, M.R. (2004). Variations in patients' adherence to medical recommendations. A quantitative review of 50 years of research. *Medical Care, 42*, 200-209.

Düsing, R. (2003). Non-Compliance in der Hochdrucktherapie. Die wichtigsten Ursachen, und was dagegen getan werden kann. *Cardiovasc, 4*, 30-32.

Edgell, S.E., Himmelfarb, S. & Duchan, K.L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods & Research, 11*, 89-100.

Edwards, A.L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

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: American Psychological Association.

Farmer, K.C. (1999). Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice. *Clinical Therapeutics, 21*, 1074-1090.

Fend, H. (1994). Ausländerfeindlich-nationalistische Weltbilder und Aggressionsbereitschaft bei Jugendlichen in Deutschland und der Schweiz – kontextuelle und personale Antecedensbedingung. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 14*, 131-162.

Fiedler, K. Schmidt, J. & Stahl, T. (2002). What is the current truth about polygraph lie detection? *Basic and Applied Social Psychology, 24*, 313-324.

Fisher, R.J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*, 303-315.

Fox, J.A. & Tracy, P.E. (1986). *Randomized response: A method for sensitive surveys.* Beverly Hills: Sage.

Freud, S. (1938). Totem and taboo. In A.A. Brill (Ed.), *The basic writings of Sigmund Freud* (pp. 807-930). New York: Random House.

Gagné, C. & Godin, G. (2005). Improving self-report measures of non-adherence to HIV medications. *Psychology and Health, 20*, 803-815.

Garber, M.C., Nau, D.P., Erickson, S.R., Aikens, J.E. & Lawrence, J.B. (2004). The concordance of self-report with other measures of medication adherence. *Medical Care, 42*, 649-652.

Granger, B.B., Swedberg, K., Ekman, I., Ostergren, J., Yusuf, S., Michelson, E.L., Zeneca, A., Granger, C.B. & Pfeffer, M.A. (2004). Adherence, even to placebo, is strongly and independently related to outcome in patients with chronic heart failure: Results from the CHARM program. *Circulation, 110*, 557.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R. & Horvitz, D.G. (1969). The unrelated question randomized response model. Theoretical framework. *Journal of the American Statistical Association, 64*, 520-539.

Hagler, P., Vargo J. & Semple, J. (1987). The potential for faking on the Attitudes Toward Disabled Persons Scale. *Rehabilitation Counseling Bulletin, 31*, 72-76.

Hopf, W. (1999). Ungleichheit der Bildung und Ethnozentrismus. *Zeitschrift für Pädagogik, 45*, 847-865.

Hu, X. (1999). Multinomial processing tree models: An implementation. *Behavior Research Methods, Instruments, and Computers, 31*, 689-695.

Hu, X. & Batchelder, W.H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59*, 21-47.

Hyman, H. (1944). Do they tell the truth? *Public Opinion Quarterly, 8*, 557-559.

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

Jerabek, I. & de Man, A.F. (1994). Social distance among Caucasian-Canadians and Asian, Latin-American and Eastern European Immigrants in Quebec: A two-part study. *Social Behavior and Personality, 22*, 297-304.

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.

Jo, M.-S., Nelson, J.E. & Kiecker, P. (1997). A model for controlling social desirability bias by direct and indirect questioning. *Marketing Letters, 8*, 429-437.

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

Kerkvliet, J. (1994). Cheating by economics students: A comparison of survey results. *Journal of Economic Education, 25*, 121-133.

Kravitz, R.L. & Melnikow, J. (2004). Medical adherence research. Time for a change in direction? *Medical Care, 42*, 197-199.

Krueger, J. & Clement, R.W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*, 596-610.

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, Research Triangle Institute, North Carolina.

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

Lee, R.M. (1993). *Doing research on sensitive topics*. London: Sage.

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.

Lensvelt-Mulders, G.J.L.M., Hox, J.J. & van der Heijden, P.G.M. (2005). How to improve the efficiency of randomised response designs. *Quality & Quantity, 39*, 253-265.

Lensvelt-Mulders, G., Hox, J. van der Heijden, P. & Maas, C. (2005). Meta-analysis of randomized-response research. Thirty-five years of validation. *Sociological Methods & Research, 33*, 319-348.

Lensvelt-Mulders, G.J.L.M., van der Heijden, P.G.M., Laudy, O. & van Gils, G. (2006). A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. *Journal of the Royal Statistical Society A, Part 2, 169*, 305-318.

Lewicki, P. (1983). Self-image bias in person perception. *Journal of Personality and Social Psychology, 45*, 384-393.

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.

Maccoby, E.E. & Maccoby, N. (1954). The Inverview: A tool of social science. In G. Lindzey (Ed.), *Handbook of Social Psychology* (pp. 449-487). Cambridge: Addison-Wesley.

Maddala, G.S. (1983). *Limited dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Mangat, N.S. (1994). An improved randomised response strategy. *Journal of the Royal Statistical Society, 56*, 93-95.

Marks, G. & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin, 107*, 77-90.

McCrae, R.R. & Costa, P.T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*, 882-888.

Mielke, R. & Mummendey, H.D. (1995). Wenn Normen zu sehr wirken – Ausländerfeindlichkeit, Bildungsgrad und soziale Erwünschtheit. *Bielefelder Arbeiten zur Sozialpsychologie, 175*, 1-9.

Miller, J.D. (1984). *A new survey technique for studying deviant behavior*. Unveröffentlichte Doktorarbeit. Washington: George Washington University.

Miller, J.D. (1985). The nominative technique: A new method of estimating heroin prevalence. In B.A. Rouse, N.J. Kozel & L.G. Richards (Eds.), *Self-report methods of estimating drug use. Meeting current challenges to validity* (pp. 104-124). Washington: Government Printing Office.

Moshagen, M., Musch, J., Ostapczuk, M. & Zhao, Z. (in Vorbereitung). *Reducing socially desirable responding in epidemiological surveys using a cheating detection extension of the randomized-response-technique*.

Mummendey, H.D. (1987). *Die Fragebogen-Methode. Grundlagen und Anwendung in persönlichkeits-, Einstellungs- und Selbstkonzeptforschung*. Göttingen: Hogrefe.

Mummendey, H.D., Bolten, H.G. & Isermann-Gerke, M. (1982). Experimentelle Überprüfung des Bogus-Pipeline-Paradigmas: Einstellungen gegenüber Türken, Deutschen und Holländern. *Zeitschrift für Sozialpsychologie, 13*, 300-311.

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

Musch, J. & Bröder, A. (eingereicht). *An experimental investigation of unethical behavior using a cheating detection extension of the randomized response technique*.

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: Pabst.

Musch, J. & Plessner, H. (eingereicht). *Estimating the prevalence of doping using a cheating detection variant of the randomized-response technique*.

Nederhof, A.J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*, 263-280.

Ong, A.D. & Weiss, D.J. (2000). The impact of anonymity on responses to sensitive questions. *Journal of Applied Social Psychology, 30*, 1691-1708.

Ostapczuk, M., Moshagen, M., Zhao, Z. & Musch, J. (eingereicht). *Assessing sensitive attributes using the randomized-response-technique: Evidence for the importance of response symmetry*.

Ostapczuk, M. & Musch, J. (eingereicht). *Projective questioning overestimates the prevalence of negative attitudes towards people with physical and mental disabilities*.

Ostapczuk, M., Musch, J. & Moshagen, M. (eingereicht a). *Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique*.

Ostapczuk, M., Musch, J. & Moshagen, M. (eingereicht b). *A randomized-response investigation of the education effect in attitudes towards foreigners*.

Pass, M.G. (1988). Race relations and the implications of education within prison. *Journal of Offender Counseling, Services & Rehabilitation, 12*, 145-151.

Pauls, C.A. & Crost, N.W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences, 37*, 1137-1151.

Photiadis, J.D. & Biggar, J. (1962). Religiosity, education, and ethnic distance. *American Journal of Sociology, 67*, 666-672.

Ray, J.J. (1990). Racism, conservatism and social class in Australia: With German, Californian and South African comparisons. *Personality and Individual Differences, 11*, 187-189.

Reamer, F.G. (1979). Protecting research subjects and unintended consequences: The effect of guarantees of confidentiality. *Public Opinion Quarterly, 43*, 497-506.

Rittenhouse, B.E. (1996a). A novel compliance assessment technique. The randomized response interview. *International Journal of Technology Assessment in Health Care, 12*, 498-510.

Rittenhouse, B.E. (1996b). Respondent-specific information from the randomized response interview: Compliance assessment. *Journal of Clinical Epidemiology, 49*, 545-549.

Robinson, D. & Rohde, S. (1946). Two experiments with an anti-Semitism poll. *Journal of Abnormal and Social Psychology, 41*, 136-144.

Saenger, G. & Gilbert, E. (1950). Customer reactions to the integration of Negro sales personnel. *International Journal of Opinion and Attitude Research, 4*, 57-76.

Scheers, N.J. (1992). Methods, plainly speaking: A review of randomized response techniques. *Measurement and Evaluation in Counseling and Development, 25*, 27-41.

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

Singer, E., Hippler, H.-J. & Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research, 4*, 256-268.

Smith, G.H. (1954). *Motivation research in advertising and marketing*. New York: McGraw-Hill.

Snir, R. & Harpaz, I. (2002). To work or not to work: Non-financial employment commitment and the social desirability bias. *Journal of Social Psychology, 142*, 635-644.

Soeken, K.L. (1987). Randomized response methodology in health research. *Evaluation & Health Professions, 10*, 68-66.

SPSS 13.0 (2004). *Users' guide*. New York: McGraw-Hill.

Stahl, C. & Klauer, K.-C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods, 39*, 267-273.

Stöber, J. (1999). Die Soziale-Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität. *Diagnostica, 45*, 173-177.

Stöber, J., Dette, D.E. & Musch, J. (2002). Comparing continuous and dichotomous scoring of the balanced inventory of desirable responding. *Journal of Personality Assessment, 78*, 370-389.

Strike, D.L., Skovholt, T.M. & Hummel, T.J. (2004). Mental health professionals' disability competence: Measuring self-awareness, perceived knowledge, and perceived skills. *Rehabilitation Psychology, 49*, 321-327.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47*, 143-148.

Thieda, P., Beard, S., Richter, A. & Kane, J. (2003). An economic review of compliance with medication therapy in the treatment of schizophrenia. *Psychiatric Services, 54*, 508-516.

Umesh, U.N. & Peterson, R.A. (1991). A critical evaluation of the randomized response method. *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, CASI and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research, 28*, 505-537.

Volicer, B.J. & Volicer, L. (1982). Randomized response technique for estimating alcohol use and noncompliance in hypertensives. *Journal of Studies on Alcohol, 43*, 739-750.

Wagner, U. & Zick, A. (1995). The relationship of formal education to ethnic prejudice: Its reliability, validity and explanation. *European Journal of Social Psychology, 24*, 41-56.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63-69.

Weiner, N.L. (1974). The effect of education on police attitudes. *Journal of Criminal Justice, 2*, 317-328.

Weisel, A., Kravetz, S., Florian, V. & Shurka-Zernitsky, E. (1988). The structure of attitudes toward persons with disabilities: An Israeli validation of Siller's Disability Factor Scales-General (DFS-G). *Rehabilitation Psychology, 33*, 227-238.

Weitz, J. & Nuckols, R.C. (1953). The validity of direct and indirect questions in measuring job satisfaction. *Personnel Psychology, 6*, 387-494.

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

Yazbeck, M., McVilly, K. & Parmenter, T.R. (2004). Attitudes toward people with intellectual disabilities: An Australian perspective. *Journal of Disability Policy Studies, 15*, 97-111.

Yesalis, C.E. & Courson, S.P. (1991). Anabolic steroid use among self-selected sample of NFL players. In S. Courson & L.R. Schreiber (Eds.), *False Glory: Steelers and steroids. The Steve Courson Story* (pp. 205-215). Stamford: Longmeadow Press.

Zerbe, W. & Paulhus, D.L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review, 12*, 250-264.

Einzelarbeiten

Nachfolgend sind die Arbeiten aufgeführt, auf denen diese Dissertation basiert. Die darin zitierte Literatur ist im Anhang der jeweiligen Arbeit aufgeführt.

- Ostapczuk, M., Musch, J. & Moshagen, M. (eingereicht a). *Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique*.

- Ostapczuk, M., Musch, J. & Moshagen, M. (eingereicht b). *A randomized-response investigation of the education effect in attitudes towards foreigners*.

- Ostapczuk, M. & Musch, J. (eingereicht). *Projective questioning overestimates the prevalence of negative attitudes towards people with physical and mental disabilities*.

- Ostapczuk, M., Moshagen, M., Zhao, Z. & Musch, J. (eingereicht). *Assessing sensitive attributes using the randomized-response-technique: Evidence for the importance of response symmetry*.

Running head: Medication Non-Adherence and RRT

Word count: 3679

Improving self-report measures of medication non-adherence using a cheating detection extension of the Randomized-Response-Technique

Martin Ostapczuk, Dipl.-Psych.

Jochen Musch, Prof. Dr.

Morten Moshagen, Dipl.-Psych.

Heinrich-Heine-Universitaet Duesseldorf

Address correspondence to:

Martin Ostapczuk

University of Duesseldorf

Institute of Experimental Psychology

Universitaetsstr. 1

D-40225 Duesseldorf

Germany

Phone: +49 211 81 10524

Fax: +49 211 81 11753

E-Mail: martin.ostapczuk@uni-duesseldorf.de

Word count: 250

Background: Medication non-adherence is a serious problem for medical research and clinical practice. Self-reports are only moderately valid, and objective methods are cumbersome and expensive to administer.

Objectives: We sought to improve self-reports of medication non-adherence using a cheating detection extension of the randomized-response-technique (RRT), which encourages more honest responses by offering interviewees a higher degree of anonymity while simultaneously allowing us to estimate the proportion of respondents disobeying the RRT instructions.

Methods: 597 patients in 3 different German medical institutions were asked to report their lifetime prevalence of medication non-adherence under 1 of 2 different questioning procedures, direct questioning (n=124) or randomized-response (n=473).

Results: When questioned directly, only 20.9% of patients admitted to intentional medication non-adherent behavior, as opposed to 32.7% of patients under RRT conditions (P<0.05). Additionally, a cheating detection extension of the RRT revealed a significant proportion of patients (47.1%, P<0.001) disobeying the instructions in the RRT condition. Assuming that either none or all of them were non-adherent, a lower and upper bound of 32.7% and 79.8%, respectively, could be estimated for the lifetime prevalence of non-adherent behavior.

Conclusions: The results demonstrate that self-report measures as well as traditional variants of the RRT, which do not take cheating into account may provide considerably distorted estimates of the prevalence of medication non-adherence. A cheating detection extension of the RRT was found to be a more suitable instrument to reduce and quantify the pervasive bias accompanying self-reports of sensitive behavior.

Key words: medication non-adherence, medication non-compliance, self-report measures, randomized-response-technique

Introduction

Hippocrates already called attention to the fact that patients do not always take their medicine as prescribed by the physician.[1] More than 2400 years later, patient compliance is still a significant topic in medical research.[2] Whereas compliance traditionally referred to a one-sided obedience of the patient to the physician's authority, the new term adherence emphasizes the patient's and physician's cooperative effort to improve the patient's health.[3-6] Irrespective of terminology, the core definition of adherence and compliance remains the extent to which a patient's behavior coincides with medical advice.[3] A further distinction can be made according to the intentionality of the behavior, however. Patients can be non-adherent either deliberately, e.g., by deciding to no longer take their medication, or undeliberately, e.g., by simply forgetting to take it.[5,6]

For two reasons, non-adherence constitutes a fundamental problem for both medical research and clinical practice. First, outcomes cannot be interpreted as being the result of medication if there is doubt as to whether all patients adhered to the prescriptions.[3] Recently, Granger et al.[7] demonstrated that for patients with heart failure, adherence was an even more important determinant of the outcome than the prescribed agent. Second, improper medication resulting from non-adherence is likely to cause immense economic costs,[8] which were estimated as high as 10.5 billion Deutschmark (i.e., approximately 7 billion US dollar) in Germany in 1999.[9] As Farmer put it, "suboptimal adherence to (…) medication regimens is a well-recognized but poorly understood problem."[3,(p.1074)] One of the major impediments to research on adherence is the lack of assessment methods that are both viable and valid. Previous estimates of the prevalence of non-adherence accordingly vary

across the entire range from 0% to 95%, with a median of 24%. Figures depend to a great extent on the exact definition of adherence, the patient population studied, the treatment regimens under consideration, and the method of measurement.[2,3,10]

It is useful to distinguish between subjective and objective adherence measures. Subjective measures include patient self-reports (e.g., diaries, interviews, questionnaires) and collateral reports (e.g., by relatives or nursing staff). Objective measures include pill counts, the detection of the drug, metabolite or biological marker in physiological fluids (e.g., blood or urine), and medication event monitoring systems (MEMS) as the alleged gold standard.[1-3,11] The main advantage of subjective methods is their practicality, time-efficiency, and low cost. On the other hand, subjective methods are quite susceptible to social desirability bias. Due to deliberate and undeliberate denial (e.g., when keeping a diary) and feigning (e.g., when being watched by a nurse), subjective methods may underestimate the true prevalence of non-adherence.[3,12] Validation studies have shown that collateral reports are usually no more valid than patient's self-reports; among the latter, questionnaires and diaries tend to yield more valid data than interviews.[2,13]

While less susceptible to social desirability distortions, objective measures suffer from different problems. First, they are generally much more cumbersome and expensive to administer than subjective measures. Moreover, while pill counts and MEMS may be immune to undeliberate denial, they are not immune to deliberate feigning. Finally, monitoring the drug level or a biological marker may help to obtain a credible qualitative measure of adherence, but falls short of providing exact information on drug dosage. Additionally, monitoring drug levels cannot detect "white-

coat compliance" where the patient is non-adherent until shortly before a clinic visit, and returns to non-adherent behavior just after the appointment.[1,3,4]

In sum, existing adherence measures tend to be impractical or of questionable validity, particularly with regard to the assessment of intentional non-adherence. Accordingly, DiMatteo concluded that "the question of how best to measure (adherence) is still an open one."[2,(p.207)] This methodological problem severely restricts current knowledge about non-adherence and how best to reduce it.[10]

Even improved objective measures are unlikely to become less complex and time-consuming than subjective measures. Therefore it seems worthwhile to attempt to increase the validity of subjective measures. Several authors have criticized a lack of research on this topic.[4] Garber et al. suggested that self-reports could be improved by offering patients a higher degree of anonymity,[13] an approach that has already proven fruitful in research on sensitive issues in the social sciences.[14,15] Soeken[16] and Rittenhouse[17,18] explicitly suggested to employ the randomized-response-technique (RRT)[19] in health research. In the present study, we heeded this advice in order to gain more insight into the lifetime prevalence of medication non-adherence in a German patient sample. The RRT is a sophisticated self-report measure supposed to elicit more honest answers by assuring interviewees the confidentiality of their responses by appropriate randomization.

For example, in the so-called "forced-response" variant of RRT, all respondents are confronted with the critical question (e.g., "Have you ever not taken medication prescribed by a physician as directed?"); but before answering, a randomization device is used to determine whether respondents are requested to answer the question honestly, or whether they are requested to provide a prespecified response

(e.g., “yes”). The outcome of the randomization procedure is unknown to the experimenter, who thus never knows whether a “yes”-response resulted from truthful answering or from the randomization process. The technique makes it much easier for the respondent to answer a critical question in the affirmative, because a “yes”-response is no longer stigmatizing. Arguably, this encourages more honest responding. Knowing the probability distribution of the randomization device, the researcher may estimate the proportion of “yes”-answers that have not been prompted by the randomization procedure. These are considered truthful avowals.[20,21]

When surveying sensitive behavior, the RRT often yields more valid estimates than traditional surveys.[22] In the as yet only study using the RRT to assess non-adherence, hypertensives reported a somewhat higher non-compliance in taking prescribed medication when being asked with the RRT, as compared with a direct questioning procedure.[12] The RRT has also been successfully used to obtain information on the prevalence of sensitive behavior as diverse as academic cheating, doping, illegal drug use, abortion, software piracy, tax evasion, shoplifting, and rape.[22-25] In all cases, the randomized-response model tried to estimate the population proportions of 2 disjoint and exhaustive groups: respondents who engaged in the critical behavior, and respondents who did not. The respective proportions of these 2 groups in the population were represented by parameters π and β; because they add up to 1, only 1 parameter had to be estimated. This could be easily done on the basis of the 1 proportion of “yes”-responses available in the typical RRT model.

Despite their many successful applications, traditional RRT models have been criticized for being susceptible to cheaters, that is, respondents who are not answering as directed by the randomization device.[26] Indeed, there is evidence that such cheating occurs.[27,28] To the extent that participants fail to follow the instructions by denying the critical behavior although they are asked by the randomization device simply to say "yes" regardless of the question content, the prevalence of the critical behavior is underestimated. To improve the procedure used in [12] and to address the problem of cheating in RRT surveys, we employed an experimental cheating detection extension of the RRT.[23-25,29]

In what we refer to as the cheating detection model (CDM), Clark and Desharnais[29] assume that some participants may deny the sensitive behavior in a randomized-response survey despite being directed by the randomization device to affirm it. Nothing can be and is assumed regarding the true status of such cheating respondents. They may deny a behavior in which they have actually engaged; however, it is also possible that they are innocent respondents attempting to rule out even the slightest suspicion of them committing an undesirable act. Because there is no way to distinguish between these two cases within the model, the true status of a cheater who is not following the instructions remains unknown.

Figure 1 illustrates how the CDM can be graphically depicted as a multinomial model aimed at dividing the population into 3 disjoint and exhaustive classes: π (the proportion of non-cheating and honest "yes"-respondents, i.e., patients who truthfully admit to medication non-adherent behavior), β (the proportion of non-cheating and honest "no"-respondents, i.e., medication adherent patients), and γ (=1-π-β, the

proportion of cheating patients who disobey the RRT rules by replying “no” to the critical question regardless of the outcome of the randomization process).

Insert Figure 1 about here

Obviously, there are now 2 independent parameters in this model, for the 3 proportions π, β, and γ are constrained to add up to 1. The parameters can therefore no longer be estimated on the basis of the single proportion of “yes”-responses that traditional RRT models provide. Instead, to obtain a sufficient data base, it is necessary to pursue an experimental approach. In particular, 2 independent samples of respondents have to be questioned with different probabilities $\underline{p_1}$ and $\underline{p_2}$ of being forced by the randomization device to answer in the affirmative.[23-25,29] Figure 1 depicts only one of these groups, in which probability $\underline{p_1}$ applies; the second group could be represented by an identical figure with the sole exception that probability $\underline{p_1}$ would have to be replaced with probability $\underline{p_2}$. Assuming that the same proportions π, β, and γ apply in both groups when participants are randomly assigned to conditions, the CDM allows us to observe 2 independent proportions of “yes”-responses. These 2 proportions suffice to estimate the 2 independent parameters π and β; γ can then be computed as $1-\pi-\beta$. For this particular model, Clark and Desharnais provide closed-form solutions for maximum likelihood estimates of the parameters π, β, and γ as well as a statistical test of the null hypothesis that no cheating occurs ($\gamma=0$).[29] However, their model can be subsumed under the more general family of multinomial models for which Riefer et al. have developed statistical procedures.[24,30,31] Using a

multinomial modeling framework, it is easily possible to test parameter restrictions, such as the assumption that no cheating occurs (γ=0), and also to formulate more complex models incorporating additional parameters.[23-25] These may represent, for example, subgroups for which parameters have to be estimated separately. In the present study, we took advantage of this possibility by breaking down the sample by variables that are potentially relevant for adherence (e.g., sex, age, medical records). We were thus able to investigate the possible influence of these variables in suitably expanded multinomial models.

We also made use of a unique theoretical advantage the CDM offers over both traditional surveys and previous RRT models. If no cheating occurs (i.e., if the proportion γ of cheating respondents can be set equal to 0 without a significant loss in the goodness of fit of the model), the parameter π provides an asymptotically unbiased estimate of the population proportion engaged in the sensitive behavior. If, however, there is a significant proportion of cheating respondents, the CDM still allows us to compute a lower and upper bound for the sensitive attribute by assuming that cheating respondents, whose real behavior cannot be ultimately determined, either all did not or did engage in the critical behavior.[23-25]

To the best of our knowledge, the as yet only study that used an RRT procedure to measure medication adherence was conducted by Volicer and Volicer.[12] In this study, hypertensives reported a higher incidence of daily use of alcohol and higher non-adherence to their prescribed blood pressure medication by a dichotomous response RRT as compared with direct questioning (DQ), and more drinks per week, but similar medication non-adherence by a quantitative response RRT as compared with DQ.[12] Regarding the desirable reduction of self-report bias, these results may be

considered promising, but the RRT variants used were not capable of detecting cheaters and thus, an upper bound for the prevalence of non-adherence could not be determined.

In the present study, we explored the feasibility of investigating the lifetime prevalence of medication non-adherence using the cheating detection extension of the RRT outlined above. To this end, we conducted a multi-center survey in 3 different German medical institutions. We expected patients to report a higher – and presumably more valid – lifetime prevalence of medication non-adherence under randomized-response conditions. To obtain an estimate of how much response bias may be reduced using the CDM, and to make results comparable with traditional self-report studies of medication non-adherence, we also included a DQ control condition.

Methods

Patients and Setting

To obtain results that are generalizable across different settings, we recruited 617 patients in 2 medical practices (a gastroenterologist and a general practice) and 1 clinic for cardio-thoracic surgery. The 3 different sites were located in different German cities. Exclusion criteria were: 1) age below 15 or above 90; 2) impaired vision or reading capability; 3) psychiatric illness requiring treatment; 4) dementia; 5) post-operative stroke or other complications. Patients were asked to fill out a questionnaire on an anonymous and voluntary basis either before or after an appointment or treatment. Questionnaires were returned in a sealed envelope.

Study Design and Critical Question

Patients answered a variety of demographic questions. Additionally, they answered questions regarding their health and their personal attitudes on a number of issues as well as several questions dealing with their medication taking habits. Only one of these questions was sensitive in nature and relevant for the present research. It read, "Have you ever intentionally and for a considerable time not taken medication prescribed by a physician as directed (e.g., by taking it for considerably too short or too long a period, too frequently or too infrequently, too early or too late in the daytime)?" The wording of the item thus referred to the various patterns[3] of deliberate non-adherence.[5,6]

In the DQ baseline condition, respondents were simply asked to reply "yes" or "no" to the critical question. In the other conditions, the sensitive question was asked in RRT format. Two RRT conditions were needed because employing the CDM requires 2 groups with different randomization probabilities, $\underline{p_1}$ and $\underline{p_2}$. In the low probability group, instructions read: "If your father was born in January or February, then please reply 'yes' to the following question independently of its content. If your father was born in another month, please reply truthfully." The probability of being forced to say "yes" thus approximated $\underline{p_1}$=2/12=1/6=0.17, as shown by birth statistics made available to us by the German Federal Agency for Statistics. In the high probability condition, the following instruction was given: "If your father was born in January or February, then please reply to the following question truthfully. If your father was born in March, April, May, June, July, August, September, October, November or December, then please reply 'yes' to the following question independently of its content." The probability $\underline{p_2}$ of being forced to say "yes" thus approximated 1.00-

p_1=10/12=5/6=0.83 in the high probability condition. Detailed instructions explained that as a result of this randomization procedure, the confidentiality of the participants' responses was guaranteed, and individual answers could no longer be linked to the respondents' true status with regard to the critical behavior.

Statistical Methods

The program G*Power 2[32] was used to compute the sample size required to achieve a statistical power of 95% to detect a small to medium effect (w=0.15) of questioning mode on self-reported non-adherence. Results indicated that at the conventional alpha level of .05, a sample size of n=578 was needed.

Based on the number of "yes"- and "no"-responses to the critical question, we computed maximum likelihood estimates for the multinomial model parameters π, β, and γ using the EM-algorithm[31] implemented in the program HMMTree.[33] Model fit was assessed by the asymptotically chi-square distributed log-likelihood statistic G^2. The multinomial model for the total sample was saturated (df=0); the 2 proportions of observable "yes"-responses in the 2 RRT conditions just sufficed to estimate the 2 independent parameters π and β (and γ=1-π-β). Parameter restrictions imposed on the model were tested by the ΔG^2 statistic.

Results

Patient Characteristics

Participants who had not answered the sensitive question were excluded from further analyses. This resulted in a final data set consisting of 597 patients with a mean age

of 51.9 (SD=17.1) years. Of the respondents, 53.8% were female, and 91.3% of German nationality. Educational level was rather low on average; only 24% of the sample had acquired the “Abitur” (an advanced high school diploma in Germany allowing the student to commence studies at the university). To compensate for the loss of efficiency of parameter estimates associated with the use of the randomization procecure,[19] patients were randomly assigned to 1 of 3 conditions by a ratio of 2:2:1 (RRT with p_1 : RRT with p_2 : DQ) This resulted in 241 and 232 participants, respectively, in each of the 2 RRT conditions, and 124 respondents in the DQ condition.

Lifetime Non-Adherence: RRT vs. Direct Questioning

Table 1 shows the parameter estimates for the whole sample, computed for the saturated model with df=0 and G^2=0. When questioned directly, only 20.9% [95% confidence interval (CI)=13.8-28.1%] of patients admitted to medication non-adherent behavior, suggesting a rather low lifetime prevalence of medication non-adherence. The results of the RRT conditions drew quite a different picture, however. Using the multinomial model, the proportion of patients honestly admitting to non-adherent behaviors was estimated at π=32.7% [95% CI=24.9-40.4%]. As expected, this proportion was significantly higher than the corresponding proportion in the DQ condition as shown by fitting a restricted model which assumed that the proportion of “yes”-answers to the direct question did not differ from the estimated proportion of honest “yes”-answers in the RRT condition. Imposing this restriction (% yes=π) significantly deteriorated the fit of the model, ΔG^2(df=1)=4.59, P<0.05. Moreover, results also indicate a considerable proportion of cheaters disobeying the RRT instructions, γ=47.1% [95% CI=38.9-55.2%]. Assuming that no cheating

occurred (γ=0) resulted in a significantly worse fit of the model, ΔG^2(df=1)=174.19, P<0.001. To compute a lower and upper bound for the true proportion of lifetime medication non-adherence, we assumed that the cheating patients, whose true behavior cannot be ultimately determined, either did not or did engage in non-adherent behaviors. This resulted in a lower bound of π=32.7% and an upper bound of $\pi+\gamma$=32.7%+47.1%=79.8% for the true proportion of lifetime medication non-adherent patients.

Insert Table 1 about here

Post-Hoc Analyses

Previous studies have shown that demographic and medical variables may exert an influence on the prevalence of medication adherence.[2,10] Within the multinomial model framework, we therefore conducted a series of additional analyses to identify potential moderators. However, none of the variables we investigated influenced non-adherence estimates. The fit of an extended model did not suffer significantly from imposing the restriction of identical prevalence parameters across groups differing in age, sex, education, and medical records (all P's>0.05).

Discussion

Determining the prevalence of medication non-adherence – and of intentional non-adherence in particular – is difficult, but of fundamental importance for understanding

and tackling the problem of non-adherent patient behavior. We therefore explored whether a multinomial model of an experimental cheating detection extension of the forced-response variant of the RRT[23-25,29] would allow us to improve the prevalence estimates of lifetime non-adherence as compared with a traditional self-report measure.

The results are in accordance with a number of previous studies comparing DQ procedures with the RRT format. As suggested by Garber et al.,[13] Soeken,[16] and Rittenhouse,[17,18] the anonymity guaranteed by the randomization of responses led to a higher and presumably more valid estimate of intentional medication non-adherence. When a randomizing procedure protected the privacy of respondents, the percentage of patients admitting to non-adherent behaviors was estimated at 32.7%. A significantly lower proportion of respondents (20.9%) admitted to the same behavior when asked directly. Additionally, employing a cheating detection approach, we were able to show that almost one half of patients failed to conform to the RRT rules and decided to play safe by denying the critical behavior even though the randomization device told them otherwise. Even though nothing can and should be assumed about the true status of these cheating respondents, determining their number allowed us to compute an upper bound for the prevalence estimate by assuming, in a worst-case-scenario, that every single patient disobeying the RRT instructions was actually non-adherent. The resulting upper bound of 79.8% was much closer to the maximum (95%), rather than to the median (24%) of the prevalence rates previously reported in the literature. This result suggests that intentional non-adherence may constitute a much more serious problem than has often been assumed.

Taken together, our results indicate that both traditional self-report and traditional randomized-response formats not considering cheating, such as the one used by Volicer and Volicer,[12] may provide strongly distorted estimates of medication non-adherence. On the other hand, RRT models which do take cheating into account, may help to gain valuable additional insight into the prevalence of medication non-adherence. In particular, a cheating detection approach to the RRT seems to be the only method allowing us to compute an upper limit for the prevalence of non-adherent behavior. While yielding more valid results, this method is almost as easy to administer as other subjective measures of non-adherence, and much less cumbersome than objective methods.

Some limitations of our study, however, should be acknowledged. First, we do not claim that the estimates we obtained are representative for all German patients and medical settings. Second, our estimates may be somewhat higher than in at least some of the previous studies,[2,3,10] because we assessed lifetime non-adherence, rather than non-adherence in a specified time frame. On the other hand, unlike many other studies, we restricted our analysis to intentional non-adherence, which probably resulted in a lower estimate. Third, due to the randomization procedure, all RRTs introduce random error and therefore have lower efficiency, i.e., greater sampling variation, than a direct question. Even though the forced-response variant of the RRT has been advocated as the most efficient RRT design available,[34] employing the RRT still requires larger samples than traditional surveying. The loss of efficiency associated with the use of the RRT is compensated by a reduction of response bias only when questions of a sufficiently sensitive nature are being asked, as in the present study. However, we argue that even though these drawbacks require a careful decision prior to each survey, adopting the CDM may often be helpful and

sometimes the only way to obtain meaningful results. In sum, we therefore strongly recommend the use of cheating detection extensions of randomized-response models in future studies on medication non-adherence.

Acknowledgments

This work was supported by grant MU 2674/1-1 from the Deutsche Forschungsgemeinschaft (German Research Foundation). We wish to thank Arno Krian, Barbara and Thilo Moshagen, Anna-Maria and Stefan Ostapczuk, Birgit and Klaus Scholz, and Günther and Karin Szibor for their help in conducting the study and collecting the data.

References

1. Düsing R. Non-Compliance in der Hochdrucktherapie. Die wichtigsten Ursachen, und was dagegen getan werden kann. [Non-compliance in the treatment of hypertension. Major reasons and solutions] Cardiovasc 2003;4:30-32.

2. DiMatteo MR. Variations in patients' adherence to medical recommendations. A quantitative review of 50 years of research. Med Care 2004;42:200-209.

3. Farmer KC. Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice. Clin Ther 1999; 21:1074-1090.

4. Gagné C, Godin G. Improving self-report measures of non-adherence to HIV medications. Psychol Health 2005;20:803-815.

5. Morisky DE, Green LW, Levine DM. Concurrent and predictive validity of a self-reported measure of medication adherence. Med Care 1986;42:67-74.

6. Sewitch MJ, Dobkin PL, Bernatsky S, et al. Medication non-adherence in women with fibromyalgia. Rheumatology 2004;43:648-654.

7. Granger BB, Swedberg K, Ekman I, et al. Adherence, even to placebo, is strongly and independently related to outcome in patients with chronic heart failure: Results from the CHARM program. Circulation 2004;110:557.

8. Thieda P, Beard S, Richter A, et al. An economic review of compliance with medication therapy in the treatment of schizophrenia. Psychiatr Serv 2003;54:508-516.

9. Volmer T, Kielhorn, A. Kosten der Non-Compliance [Costs of non-compliance]. Gesundheitsökonomisches Qualitätsmanagement 1999;4:55-61.

10. Kravitz RL, Melnikow J. Medical adherence research. Time for a change in direction? Med Care 2004;42:197-199.

11. Lahdenperä TS, Kyngäs HA. Compliance and its evaluation in patients with hypertension. J Clin Nurs 2000;9:826-833.

12. Volicer BJ, Volicer L. Randomized response technique for estimating alcohol use and noncompliance in hypertensives. J Stud Alcohol 1982; 43:739-750.

13. Garber MC, Nau DP, Erickson SR, et al. The concordance of self-report with other measures of medication adherence. Med Care 2004;42:649-652.

14. Joinson A. Social desirability, anonymity, and internet-based questionnaires. Behav Res Methods Instrum Comput 1999;31:433-438.

15. Ong AD, Weiss, DJ. The impact of anonymity on responses to sensitive questions. J Appl Soc Psychol 2000;30,1691-1708.

16. Soeken, KL. Randomized response methodology in health research. Eval Health Prof 1987;10:58-66.

17. Rittenhouse, BE. A novel compliance assessment technique – The randomized response interview. Int J Technol Assess Health Care 1996;12:498-510.

18. Rittenhouse, BE. Respondent-specific information from the randomized response interview: Compliance assessment. J Clin Epidemiol 1996;49:545-549.

19. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. J Am Stat Assoc 1965;60:63-69.

20. Greenberg BG, Abul-Ela A-LA, Simmons WR, et al. The unrelated question randomized response model. Theoretical framework. J Am Stat Assoc 1969;64:520-539.

21. Dawes RM, Moore M. Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen. [Guttman scaling of orthodox and randomized reactions] In: Petermann F, ed. Einstellungsmessung, Einstellungsforschung. [Attitude measurement, attitude research] Göttingen: Hogrefe;1980:117-133.

22. Lensvelt-Mulders G, Hox J, van der Heijden P, et al. Meta-analysis of randomized response research. Thirty-five years of validation. Sociol Methods Res 2005;33:319-348.

23. Musch J, Bröder A. An experimental investigation of unethical behavior using a cheating detection extension of the randomized response technique. Submitted.

24. Musch J, Bröder A, Klauer KC. Improving survey research on the World-Wide-Web using the randomized response technique. In: Reips UD, Bosnjak M, eds. Dimensions of Internet Science. Lengerich: Pabst;2001:179-192.

25. Musch J, Plessner H. Estimating the prevalence of doping using a cheating detection variant of the randomized-response technique. Submitted.

26. Campbell AA. Randomized response technique. Science 1987;236:1049.

27. Stem DE, Steinhorst RK. Telephone interview and mail questionnaire applications of the randomized response model. J Am Stat Assoc 1984;74:555-564.

28. Lensvelt-Mulders GJLM, Boeije HR. Evaluating compliance with a computer assisted randomized response technique: A qualitative study into the origins of lying and cheating. Comput Hum Behav 2007;23:591-608.

29. Clark SJ, Desharnais RA. Honest answers to embarrassing questions: Detecting cheating in the randomized response model. Psychol Methods 1998;3:160-168.

30. Riefer DM, Batchelder WH. Multinomial modeling and the measurement of cognitive processes. Psychol Rev 1998;95:318-339.

31. Hu X, Batchelder WH. The statistical analysis of general processing tree models with the EM algorithm. Psychometrika 1994;59:21-47.

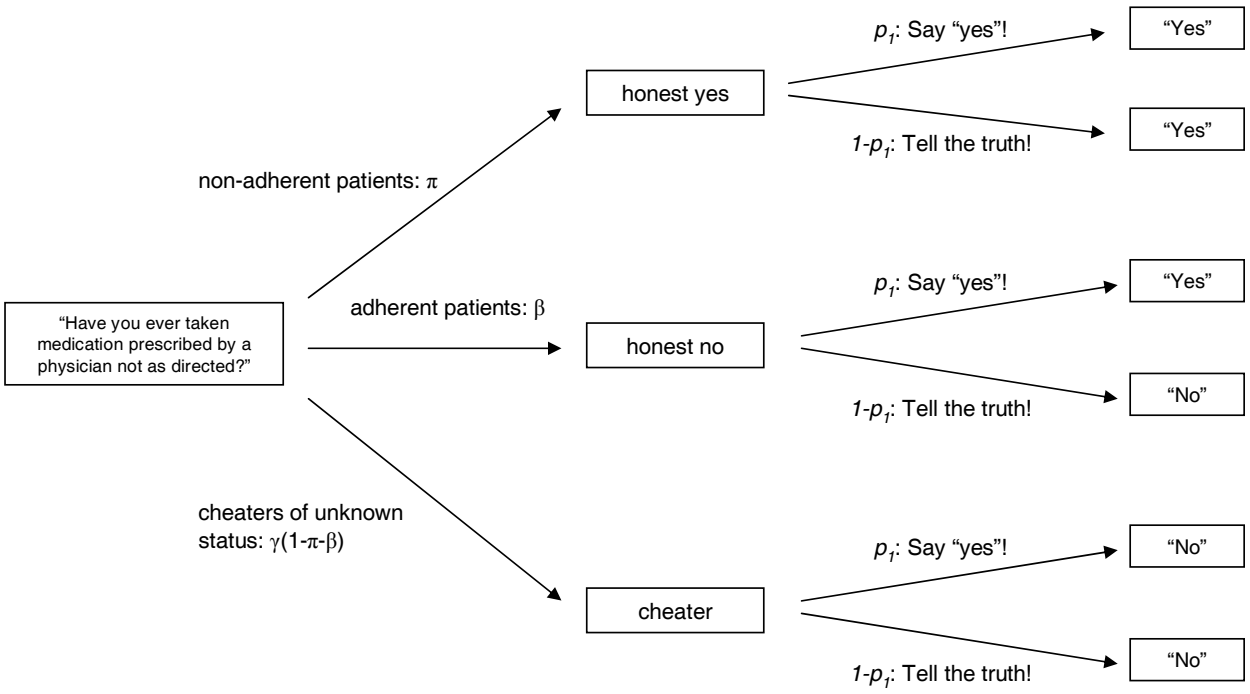
32. G*Power [computer program]. Version 2. Bonn: Erdfelder E, Faul F, Buchner A;1996.

33. HMMTree [computer program]. Freiburg: Stahl C, Klauer, KC; 2007.

34. Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM. How to improve the efficiency of randomised response designs. Quality & Quantity 2005;39:253-265.

Figures

Figure 1. A multinomial model of the cheating detection extension of the randomized-response-technique



Tables

Table 1. Lifetime Prevalence of Medication Non-Adherence

		Total
		n=597
Direct Questioning	% yes	20.9%
	[95% CI]	[13.8-28.1]
	% no	79.1%
	[95% CI]	[71.9-86.2]
Randomized-Response-Technique	Honest yes (π)	32.7%
	[95% CI]	[24.9-40.4]
	Honest no (β)	20.2%
	[95% CI]	[6.9-33.4]
	Cheaters (γ)	47.1%
	[95% CI]	[38.9-55.2]
	ΔG^2 (df=1): $\gamma=0$[1]	174.19***
	ΔG^2 (df=1): % yes=π [2]	4.59*

CI = confidence interval.

[1]High values indicate that the fit of the model worsens under the assumption that no cheating occurs ($\gamma = 0$).

[2]High values indicate that the fit of the model worsens under the assumption that the proportion of admitting patients in the DQ question (% yes) does not differ from the estimated proportion of honest "yes"-answers (π) in the RRT condition.

*P<0.05, ***P<0.001.

A randomized-response investigation of the education effect in attitudes towards foreigners

Martin Ostapczuk

Jochen Musch

Morten Moshagen

University of Duesseldorf, Germany

Correspondence address:

Martin Ostapczuk

Heinrich-Heine-University of Duesseldorf

Institute of Experimental Psychology

Universitaetsstr. 1

D-40225 Düsseldorf

Germany

Phone: +49 211 81 10524

Fax: +49 211 81 11753

E-Mail: martin.ostapczuk@uni-duesseldorf.de

Abstract

While negative correlations have often been found between a respondent's education and his attitudes towards foreigners, the reasons for this education effect are still under debate. We examined the hypothesis that the highly educated may not be genuinely less xenophobic, but simply more prone to give socially desirable, xenophile answers in attitude questionnaires. We therefore compared the attitudes of respondents who were either questioned directly or using a cheating detection extension of the randomized-response-technique. The latter is supposed to yield more honest answers to sensitive questions by experimentally offering the interviewee a higher degree of confidentiality. Under direct questioning conditions, we replicated the education effect; 75% of the highly educated expressed xenophile attitudes, as opposed to only 55% of the less educated. Under randomized-response conditions, we obtained significantly reduced estimates of 53% for the proportion of xenophiles among the highly educated, and 24% among the less educated, indicating a strong distortion of self-reported attitudes towards foreigners in both groups. However, a significant proportion of participants disobeyed the randomized-response-technique instructions regardless of education. Because the education effect was found even after controlling for social desirability, it seems to be a genuine effect, rather than an artefact of a differential response bias.

For decades, psychologists and sociologists have tried to pinpoint the correlates and determinants of xenophobia. One of the most robust findings that has been observed across different cultures, age groups, and attitude measures is a negative correlation between a respondent's formal education, and his or her prejudice against foreigners. More educated people have often been found to be less xenophobic, and more xenophile (in Germany: Abraham, 1966; Bergmann & Erb, 1991; Fend, 1994; Mielke & Mummendey, 1995; Ray, 1990; Silbermann & Hüsers, 1995, and Wagner & Zick, 1995; in the United States: Pass, 1988; Photiadis & Biggar, 1962; Robinson & Rohde, 1946; Weiner, 1974; in Canada: Jerabek & de Man, 1994; in Mexico: Cohen Shabat, 1993; in Australia and South Africa: Ray, 1990; in the Netherlands, France, and Great Britain: Wagner & Zick, 1995; in Austria: Jimenez, 1999; in Switzerland: Fend, 1994, and in Sweden: Abraham, 1966). Potential reasons underlying the education effect include a different number of positive contacts with foreign people (Hewstone & Brown, 1986; Wagner, van Dick, Pettigrew & Christ, 2003), a perception of individual or group deprivation on the part of the less educated (Vanneman & Pettigrew, 1982), and an increased commitment to democratic norms of equality possibly associated with a higher formal education (Lipset, 1983). An important alternative explanation, however, is that the education effect does not in fact reflect a genuine difference, but rather is the result of a methodological artefact. More educated people may not actually be less xenophobic, but simply more receptive to the sensitive nature of an inquiry regarding their attitudes towards foreigners. They might therefore be more inclined to bias their responses in a socially desirable, xenophile direction (Hopf, 1999; Mielke & Mummendey, 1995; Wagner & Zick, 1995).

A few studies have tried to address this issue directly. In an early attempt, Robinson and Rohde (1946) questioned a large sample of New Yorkers about their attitudes towards Jews. The interviewers were either of Jewish or non-Jewish appearance. Assuming that the more highly educated participants are more sensitive for social desirability concerns, they expected they would be more influenced by the presence of a Jewish-looking interviewer. However, Robinson and Rohde (1946) found no such effect. Besides being expectedly more anti-Semitic, the less educated seemed to be even slightly more sensitive to the appearance of the interviewer than the highly educated. Unfortunately, the authors did not assess the statistical significance of these effects.

Wagner and Zick (1995) compared a direct question condition with the bogus pipeline procedure (Jones & Sigall, 1971) to reduce the tendency to give socially desirable responses. In a German sample, they found a significant effect of education on prejudice against Turks under both conditions. They concluded that the more highly educated do not seem to give more biased responses when questioned directly. In a second study, Wagner and Zick (1995) employed the Blatant and Subtle Prejudice Scales developed by Pettigrew and Meertens (1995). The differential bias explanation would predict a negative correlation between education and scores on the blatant prejudice scale, but a nonsignificant – or maybe even positive – correlation between education and the subtle prejudice scale comprised of items which are more socially acceptable (Manganelli Rattazzi & Volpato, 2003). Contrary to the differential bias explanation, Wagner and Zick (1995) observed negative correlations for both scales. However, another application of the Blatant and Subtle Prejudice Scales led to inconsistent findings. In a study of Mielke and Mummendey (1995), the instruction to "fake good" unexpectedly

reduced subtle prejudice in both educational groups, but increased blatant prejudice regardless of education. This result is inconsistent with both the genuine difference and the differential bias hypothesis.

In sum, previous studies of the origin of the education effect have not provided conclusive results. Collecting additional, independent evidence using alternative methodological approaches seems therefore desirable to accurately assess the validity of the differential bias explanation for the education effect (Hopf, 1999; Mielke & Mummendey, 1995; Wagner & Zick, 1995). Accordingly, the use of reaction time analyses (Ling, 2002) and anonymizing survey procedures (Cobb, 2002) have been proposed to address the problem. We decided to follow the latter recommendation and used the randomized-response-technique (RRT; Warner, 1965) to better address the question of the validity of the education effect.

The Randomized-Response-Technique (RRT)

The RRT is a sophisticated self-report measure supposed to elicit more honest answers by assuring interviewees the confidentiality of their responses by appropriate randomization. It has already proven helpful in researching a number of sensitive topics in the social sciences (Lensvelt-Mulders, Hox, van der Heijden & Maas, 2005). The rationale of the RRT is that interviewees are more honest when the confidentiality of their responses is guaranteed.

In the so-called "forced-response" variant of RRT, all respondents are therefore confronted with the critical question. Before answering however, a randomization device is used to determine whether respondents are asked to answer the question honestly or whether they are asked to provide a prespecified response (e.g., "yes"). The outcome of the randomization procedure is unknown to the experimenter, who thus never knows whether a "yes"-response resulted from truthful answering or from the randomization process. The technique makes it easier for the respondent to answer a sensitive question in the affirmative, because such an answer is now no longer stigmatizing. Knowing the probability distribution of the randomization device, the researcher may estimate the number of "yes"-answers that have not been prompted by the randomization procedure. These are considered truthful avowals (Dawes & Moore, 1980; Greenberg, Abul-Ela, Simmons & Horvitz, 1969).

Surveys of sensitive behaviour have repeatedly shown that the RRT can help to yield more valid estimates than traditional surveys. The technique has been successfully used to obtain information on the prevalence of sensitive behaviours as diverse as academic cheating, illegal drug use, abortions, shoplifting, and rape (Lensvelt et al., 2005). In all cases, RRT models tried to divide the population into two disjoint and exhaustive groups: respondents who admit to a socially undesirable attitude or behaviour, and respondents who do not. The respective proportions of these two groups in the population were represented by the parameters π and β, respectively; because they add up to 1, only one parameter had to be estimated. This was easily possible using the proportion of responses affirming the sensitive attitude.

Despite their many successful applications, traditional RRT models have been criticized for being susceptible to cheaters, that is, respondents who do not answer as directed by the randomization device (Campbell, 1987). Indeed, there is evidence that such cheating occurs (Lensvelt-Mulders & Boeije, 2007; Stem & Steinhorst, 1984). The prevalence of socially undesirable attitudes is underestimated to the extent that participants fail to follow the instructions by denying the sensitive attitude in spite of being asked by the randomization device to admit to it.

To go beyond the methodology used in previous RRT studies, we therefore decided to employ an experimental cheating detection extension of the RRT (Clark & Desharnais, 1998; Musch, Bröder & Klauer, 2001). In what we will refer to as the cheating detection model (CDM), Clark and Desharnais (1998) assume that some participants may deny a sensitive attitude in a randomized-response survey despite being directed by the randomization device to confess to it. Nothing can be and is assumed regarding the true status of such cheating respondents. They may deny a sensitive attitude in spite of actually sharing it; however, it is also possible that they are not sharing the sensitive attitude but want to rule out even the slightest suspicion that it could be otherwise. Because there is no way to distinguish between these two cases within the model, the true attitude of a cheater who is not following the instructions necessarily remains unknown and cannot be ultimately determined.

Figure 1 illustrates how the CDM can be graphically depicted as a multinomial model aimed at dividing the population into three disjoint and exhaustive classes: π (the proportion of non-cheating and honest “yes”-respondents, i.e., participants who

truthfully admit to their xenophobic attitude), β (the proportion of non-cheating and honest "no"-respondents, i.e., xenophile participants truthfully denying that they have a xenophobic attitude) and γ (= 1-π-β, the proportion of cheating participants who disobey the RRT rules by denying to adopt a socially undesirable xenophobic attitude regardless of the outcome of the randomization process). Henceforth, we will refer to these three groups as the "xenophobes" (π), the "xenophiles" (β), and the "cheaters" (γ).

Insert Figure 1 about here

There are now two independent parameters in this model, because the three proportions π, β, and γ are constrained to add up to 1. The parameters can therefore no longer be estimated on the basis of the single proportion of "yes"-responses that traditional RRT models provide. Instead, to obtain a sufficient data base, it is necessary to pursue an experimental approach. In particular, two independent samples of respondents have to be questioned with different probabilities $\underline{p}_1$ and $\underline{p}_2$ of being prompted by the randomization device to answer the critical question in the affirmative (Clark & Desharnais, 1998; Musch et al., 2001). Figure 1 depicts only one of these groups, in which probability $\underline{p}_1$ applies; the second group, however, could be represented by an identical figure with the only exception that probability $\underline{p}_1$ would be replaced by probability $\underline{p}_2$. Assuming that the same proportions π, β, and γ apply in both groups when participants are randomly assigned to conditions, the CDM allows us to observe two independent proportions of "yes"-responses. These two proportions suffice to estimate the two independent parameters π and β (γ can then be computed as 1-π-β). For

this particular model, Clark and Desharnais (1998) provided closed-form solutions for maximum likelihood estimates of the parameters π, β and γ as well as a statistical test of the null hypothesis that no cheating occurs ($\gamma = 0$). The CDM thereby provides a unique theoretical advantage over both traditional surveys and more simple RRT models: if no cheating occurs, the parameter π provides an estimate of the proportion of respondents adopting an undesirable attitude. If there is a significant proportion of cheating respondents, the CDM still allows us to compute both a lower and upper bound for the prevalence of the sensitive attitude by assuming that cheating respondents, whose real attitude cannot be ultimately determined, either all do not or do adopt the undesirable attitude (Musch et al., 2001).

The CDM can be subsumed under the more general family of multinomial models for which Hu and Batchelder (1994) and Riefer and Batchelder (1998) have developed statistical procedures. Using the multinomial modelling framework, it is easily possible to test parameter restrictions, such as the assumption that no cheating occurs. This is done by determining whether the proportion γ of cheating respondents can be set equal to zero, without a significant loss in the goodness of fit of the model. More complex models may also be formulated by incorporating additional parameters (Musch et al., 2001). These additional parameters may represent, for example, subgroups for which parameters have to be estimated separately. In the present study, we took advantage of this possibility. To investigate the influence of educational level on attitudes towards foreigners, we tested whether the same prevalence parameters can be used in both the highly and the less educated group without significantly worsening the fit of the model.

To make results comparable with traditional self-report surveys on xenophobia, we also included a direct questioning (DQ) control condition and incorporated it into the multinomial model. In total, the suitably expanded multinomial model thus consisted of six different groups because both within the highly educated and the less educated group, there was an RRT group with randomization probability $\underline{p_1}$ (RRT1), an RRT group with randomization probability $\underline{p_2}$ (RRT2), and a DQ control group. The resulting model offered the possibility to test the validity of the education effect by testing the parameter restrictions $\pi_{\text{less educated}} = \pi_{\text{highly educated}}$ (to explore whether the proportion of xenophobes, π, among the less educated respondents can be set equal to the proportion of xenophobes among the highly educated respondents without significant loss in the goodness of fit of the model), and $\beta_{\text{less educated}} = \beta_{\text{highly educated}}$ (to explore whether the proportion of xenophiles, β, among the less educated respondents can be set equal to the proportion of xenophiles among the highly educated respondents without significant loss in the goodness of fit of the model).

If the education effect reflects a true attitudinal difference, we would expect the highly educated to be less xenophobic and more xenophile, respectively, than the less educated, both under DQ conditions ($\%\ \text{yes}_{\text{less educated}} > \%\ \text{yes}_{\text{highly educated}}$) and under RRT conditions ($\pi_{\text{less educated}} > \pi_{\text{highly educated}}$ and $\beta_{\text{less educated}} < \beta_{\text{highly educated}}$). If however the education effect is due to a differential social desirability bias, we would expect the highly educated to describe themselves as less xenophobic and more xenophile, respectively, only under DQ conditions ($\%\ \text{yes}_{\text{less educated}} > \%\ \text{yes}_{\text{highly educated}}$). Protected by an anonymizing RRT procedure, we would expect them to have levels of xenophobia

or xenophilia, respectively, that are comparable to those of the less educated ($\pi_{\text{less educated}} \approx \pi_{\text{highly educated}}$ and $\beta_{\text{less educated}} \approx \beta_{\text{highly educated}}$, respectively).

Method

Participants and Setting

In order to collect an educationally heterogeneous sample, participants (N = 606; 61% women; mean age = 40.7, SD = 19.3) were recruited in two universities (n = 193 psychology and business students), three medical practices (n = 314), a hospital (n = 76), and a vocational school (n = 23 students). Participants were considered eligible if they were over 15 or below 90 years old. Additional exclusion criteria for patients were: 1) impaired vision or reading capability; 2) psychiatric illness requiring treatment; 3) dementia; 4) post-operative stroke or other complications. Participants were asked to fill out a questionnaire on a voluntary basis. The questionnaire was to be returned in a sealed envelope. Respondents received no financial incentives.

Pre-Test, Materials and Study Design

Participants were asked to respond to a variety of demographic questions and questions related to their personal experience with foreigners in general and with dark-skinned Africans in particular. Only one of these questions was sensitive in nature. The sensitive item was selected based on the following criteria: first, it had to originate from a validated attitude scale. Second, it had to be of moderate social (un-)desirability to avoid floor or ceiling effects in the number of "yes"- and "no"-responses. To satisfy these criteria, we conducted an internet-based pre-test in which we asked 63

respondents to indicate how comfortable they would feel admitting 76 different attitudes and behaviours. Participants were to respond on a scale from 1 (= "neither comfortable nor uncomfortable") to 5 (= "very uncomfortable"). The item we finally selected for the present study was adapted from the Bogardus Social Distance Scale (Bogardus, 1925, 1933) and had previously been used in studies by Silbermann and Hüsers (1995) and Jimenez (1999). It achieved a mean undesirability rating of $\underline{M} = 3.79$ ($\underline{SD} = 1.23$) in our pre-test and read, "Assuming that you have a 20-year-old daughter. Would you mind her having a relationship with a dark-skinned Nigerian?" For the pre-test, we had decided to ask for the specific ethnic minority of dark-skinned Nigerians rather than the abstract target group "foreigners", because attitudes towards ethnic and religious minorities in Germany differ. Dark-skinned Africans were found to be somewhat more liked than Turks and Arabs, but less liked than Jews and Israelis (Bergmann & Erb, 1991).

To compensate for the loss of efficiency in parameter estimation associated with the use of the randomization procedure, participants were randomly assigned to one of the three conditions by a ratio of 2:2:1 (RRT with randomization probability $\underline{p_1}$: RRT with randomization probability $\underline{p_2}$: DQ). In the DQ baseline condition, respondents were simply asked to reply "yes" or "no" to the critical question. In the two other conditions, the sensitive question was asked in RRT format. Two RRT conditions were needed, since the CDM requires questioning two groups with different randomization probabilities $\underline{p_1}$ and $\underline{p_2}$. In the low probability group, instructions read: "If your mother was born in January or February, then please reply 'yes' to the following question independently of its content. If your mother was born in another month, please answer

truthfully." The probability of being forced to say "yes" thus approximated $\underline{p_1} = 2/12 = 1/6 = 0.17$, as shown by birth statistics made available to us by the German Federal Agency for Statistics. In the high probability condition, the following instruction was given: "If your mother was born in January or February, please answer truthfully. If your mother was born in March, April, May, June, July, August, September, October, November or December, then please reply 'yes' to the following question independently of its content." The probability $\underline{p_2}$ of being forced to say "yes" thus approximated $1.00-\underline{p_1} = 10/12 = 5/6 = 0.83$ in the high probability condition. Detailed instructions explained that as a result of this randomization procedure, the confidentiality of the participant's responses was guaranteed, and individual answers could no longer be linked to the respondents' true status with regard to the critical behaviour.

Education was operationalized post-hoc as a dichotomous variable (cf. Mielke & Mummendey, 1995; Wagner & Zick, 1995). A participant was considered as less or highly educated depending on whether he had acquired the German "Abitur" (an advanced high school diploma allowing the student to commence studies at the university).

Statistical Methods

Based on the number of "yes"- and "no"-responses to the sensitive question in the DQ and the RRT condition, we computed maximum likelihood estimates of the multinomial model parameters separately for the two educational groups. Parameter estimates were performed using the EM-algorithm (Hu & Batchelder, 1994) implemented in the program HMMTree (Stahl & Klauer, 2007). The fit of the model was assessed via the

asymptotically chi-square distributed log-likelihood statistic $\underline{G^2}$. The multinomial model representing the two educational groups under two different questioning procedures (DQ vs. RRT) was saturated with no degree of freedom ($\underline{df} = 0$, $\underline{G^2} = 0$), as the two proportions of observable "yes"-responses in the two RRT conditions just sufficed to estimate the two independent parameters π and β (with $\gamma = 1-\pi-\beta$) for each educational level.

Results

The final data set consisted of 282 highly educated and 324 less educated participants. In the highly educated group, there were 113 and 104 participants, respectively, in each of the two RRT conditions, and 65 participants in the DQ condition. In the less educated group, there were 133 and 126 participants, respectively, in each of the two RRT conditions, and 65 in the DQ condition.

Reliability of the Education Effect

First, we assessed the reliability of the education effect by analyzing the data in the DQ conditions. Table 1 shows the parameter estimates for the saturated model with $\underline{df} = 0$ and $\underline{G^2} = 0$.

Insert Table 1 about here

As expected, we replicated the education effect under DQ conditions. When questioned directly, only 24.6% of the highly educated admitted to a xenophobic attitude by objecting to a relationship of their daughter with a dark-skinned Nigerian; 75.4% of the highly educated had to be classified as xenophile according to their non-objecting response to this direct question. Xenophobic responses were more prevalent among the less educated respondents, of whom 44.6% admitted to a xenophobic attitude; only 55.4% of the less educated were classified as xenophiles in the DQ condition. The assumption that the proportion of xenophobic answers to the direct question did not differ in the two different educational groups led to a significantly deteriorated fit of the model; it thus had to be rejected, $\underline{\Delta G^2}$ ($\underline{df} = 1$) = 5.81, $\underline{p} < 0.05$. The previously observed effect of education on attitudes towards foreigners was thereby replicated in the DQ condition of the present study.

Validity of the Education Effect

Next, we investigated the validity of the education effect by comparing the results under DQ conditions with the results under RRT conditions. Using the multinomial model, the proportion of xenophobes (π) among the highly educated respondents was estimated at $\pi_{\text{highly educated}} = 30.1\%$ as compared to $\pi_{\text{less educated}} = 38.1\%$ among the less educated. Restricting the proportion of xenophobes to equality across educational groups did not deteriorate the fit of the model, $\underline{\Delta G^2}$ ($\underline{df} = 1$) = 1.00, $\underline{ns}$. Thus, the proportion of xenophobes did not differ between groups with different educational background. Assuming that the estimated proportion of xenophiles (β) among the highly educated does not differ from the estimated proportion of xenophiles among the less educated did however result in a significantly worsened fit of the model, $\beta_{\text{highly educated}} = 52.7\%$; β_{less}

$_{\text{educated}}$ = 23.8%; $\underline{\Delta G^2}$ ($\underline{df}$ = 1) = 4.74, $\underline{p}$ < 0.05. Thus, the proportion of xenophiles was estimated to be higher in the more educated group in the RRT condition.

Additionally, a sizable proportion of cheaters disregarding the RRT rules was detected both among the highly educated ($\gamma_{\text{highly educated}}$ = 17.2%) and among the less educated ($\gamma_{\text{less educated}}$ = 38.1%). Assuming that no cheating occurred ($\gamma_{\text{highly educated}}$ = 0 and $\gamma_{\text{less educated}}$ = 0, respectively) significantly worsened the fit of the model in both subgroups, $\underline{\Delta G^2}$ ($\underline{df}$ = 1) = 13.23, $\underline{p}$ < 0.001 and $\underline{\Delta G^2}$ ($\underline{df}$ = 1) = 67.29, $\underline{p}$ < 0.001, respectively. The estimated prevalence of cheaters was also significantly higher in the less educated than in the highly educated group, $\underline{\Delta G^2}$ ($\underline{df}$ = 1) = 6.86, $\underline{p}$ < 0.01.

The cheating participants whose true attitude cannot be ultimately determined may either do or do not adopt a xenophobic attitude. Alternately assuming that they all do not or all do, we computed a lower bound of $\pi_{\text{highly educated}}$ = 30.1% and an upper bound of $\pi_{\text{highly educated}}+\gamma_{\text{highly educated}}$ = 30.1%+17.2% = 47.3% for the true proportion of xenophobes among the highly educated as well as a lower bound of $\pi_{\text{less educated}}$ = 38.1% and upper bound of $\pi_{\text{less educated}}+\gamma_{\text{less educated}}$ = 38.1%+38.1% = 76.2% for the true proportion of xenophobes among the less educated. This upper bound for the prevalence of xenophobes among the less educated was thus markedly higher than the respective figure in the highly educated group. Similarly, we also computed a lower bound of $\beta_{\text{highly educated}}$ = 52.7% and an upper bound of $\beta_{\text{highly educated}}+\gamma_{\text{highly educated}}$ = 52.7%+17.2% = 69.7% for the true proportion of xenophiles among the highly educated, as well as a lower bound of $\beta_{\text{less educated}}$ = 23.8% and an upper bound of $\beta_{\text{less educated}}+\gamma_{\text{less educated}}$ = 23.8%+38.1% = 61.9% for the true proportion of xenophiles among the less educated.

Estimates of the prevalence of xenophiles were thus lower for the group of less educated respondents. Taken together, these results suggest that the education effect obtained under DQ conditions still remains valid under RRT conditions; controlling for social desirability bias does not make it vanish.

Effect of the Questioning Mode

Finally, we tested the assumption that the RRT provides less biased estimates by directly comparing the DQ results with the RRT estimates. While the results of the DQ conditions suggest that – despite educational differences – a notable proportion of xenophile participants exists both among the highly educated (% $no_{highly\ educated}$ = 75.4%) and among the less educated (% $no_{less\ educated}$ = 55.4%), the results of the RRT suggest a slightly different picture. Assuming that the proportions of xenophiles under DQ conditions did not differ from the corresponding proportions estimated by RRT ($\beta_{highly\ educated}$ = 52.7% and $\beta_{less\ educated}$ = 23.8%, respectively) significantly worsened the fit of the model in both subgroups, ΔG^2 (df = 1) = 4.51, $p < 0.05$ and ΔG^2 (df = 1) = 8.37, $p < 0.01$, respectively. However, due to the considerable proportion of cheaters in both subgroups, the estimates of the prevalence of xenophobic attitudes in the DQ condition (% $yes_{highly\ educated}$ and %$yes_{less\ educated}$, respectively) did not significantly differ from the respective RRT estimates ($\pi_{highly\ educated}$ and $\pi_{less\ educated}$, respectively).

On the whole, this pattern of results exemplifies the ability of the RRT to provide less biased estimates of xenophobic and xenophile attitudes, respectively, as compared to a DQ procedure.

Discussion

The education effect in attitudes towards foreigners is a basic finding of socio-psychological research on xenophobia and prejudice. Yet, its well-established stability contrasts with the sparse number of studies directly addressing the issue whether the effect reflects a true attitudinal difference or just an artefact of more socially desirable responding among the more educated respondents. We therefore scrutinized the validity of the education effect by comparing the results of a traditional DQ survey on prejudice against dark-skinned Nigerians with the results obtained using a cheating detection variant of the RRT (Clark & Desharnais, 1998; Musch et al., 2001).

As expected, the results under DQ conditions replicated the education effect. Only 24.6% of the highly educated participants were prejudiced against Nigerians as opposed to 44.6% from the less education group. More interestingly, this education effect was also found under RRT conditions. Among the highly educated, 52.7% were classified as unprejudiced, a significantly higher number than the 23.8% in the low education group. Moreover, only 30.1% of the highly educated gave a xenophobic response, as opposed to 38.1% of the less educated. While this latter difference failed to attain statistical significance due to a significant proportion of cheaters, the results still emphasize the utility of employing a cheating detection variant of the RRT because the noncompliant behaviour of some of the RRT respondents would have gone unnoticed if more simple RRT models or a direct question only would have been used.

The additional results of the RRT analysis confirmed the validity of the education effect. Assuming in a worst case scenario that all participants disobeying the RRT instructions were in fact xenophobic, resulted in an upper bound for the prevalence of xenophobes among the highly educated of 47.3%. The corresponding figure for the less educated was as high as 76.2%. Conversely, assuming that every single participant disobeying the RRT instructions was in fact xenophile, resulting in an upper bound of 69.7% for the prevalence of xenophiles for the highly educated group, which was still higher than the corresponding upper bound for the less educated group (61.9%). Finally, the direct comparison of the results obtained by DQ and RRT demonstrated the utility of the RRT in providing less distorted estimates: when anonymity was guaranteed by appropriate randomization, both the highly and the less educated participants proved to be considerably less xenophile than suggested by their responses to a direct question.

Similar to previous research by Robinson and Rohde (1946) and Wagner and Zick (1995), the present results are more in line with a true difference than a differential bias hypothesis. The highly educated do not just seem to pretend to be less prejudiced than the less educated, they really seem to be less prejudiced. However, judging from the results of the RRT, both educational groups are less unprejudiced than their answers to a direct question had suggested.

The present result of a true education effect on attitudes towards foreigners can easily be reconciled with the results of studies investigating potential mediators of the relationship between educational level and xenophobia, which include cognitive abilities, openness to experience, self-esteem- or identity-related variables, group

deprivation, incongruency between the cultural values of the national ingroup and those of the minority, and the amount of (positive) contact (Ackerman & Heggestad, 1997; Wagner & Zick, 1995). A highly educated person who is – on average – more intelligent, more open to experience, of higher self-esteem, less economically deprived, sensing a smaller incongruency between his or her own culture and the one of the outgroup, and who also has had more (positive) contact with members of the ethnic minority, is not unlikely to really be, and not just pretending to be, more xenophile.

Finally, some limitations of the present survey should also be considered. First, it remains to be investigated whether the education effect also holds for real-life behaviour, rather than for answers in a questionnaire. Wagner and Zick (1995, p. 54) remarked that "it is questionable whether the educational difference in outgroup rejection would still be observed in situations in which a much higher degree of personal involvement is entailed, such as when people are made aware of plans to establish a home for refugees in their (middle class) neighbourhood". Second, our results are based on a single sensitive item. The generalizability of the results to other measures and question contents has yet to be shown. Third, all RRT models introduce random error and thus have lower efficiency, i.e., greater sampling variation, than DQ surveys. In spite of being the most efficient RRT variant available (Lensvelt-Mulders, Hox & van der Heijden, 2005), the forced-response variant of the RRT still requires larger samples than traditional surveys. The loss of efficiency is compensated for by a reduction of response bias only when questions of a sufficiently sensitive nature are being asked, as in the present study.

Even though a careful decision must therefore be made prior to adopting the present methodology, the CDM may often be a helpful way to obtain meaningful results when surveying sensitive attitudes. In the present study, the cheating detection variant of the RRT proved its utility by demonstrating the validity of the education effect, which does not seem to be just an artefact of group differences in social desirability bias.

Author note

This work was supported by grant MU 2674/1-1 from the Deutsche Forschungsgemeinschaft (German Research Foundation). We wish to thank Arno Krian, Barbara and Thilo Moshagen, Anna-Maria and Stefan Ostapczuk, Birgit and Klaus Scholz, Günther and Karin Szibor, Monika Undorf, and Michael Wolf for their help in conducting the study and collecting the data.

References

Abraham, H.H.L. (1966). Social distance and patterns of prejudice in Germany and Sweden. Archiv für die Gesamte Psychologie, 118, 229-252.

Ackerman, P.L., & Heggestad, E.D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. Psychological Bulletin, 121, 219-245.

Bergmann, W., & Erb, R. (1991). Antisemitismus in der Bundesrepublik Deutschland. Ergebnisse der empirischen Forschung von 1946-1989. [Antisemitism in the Federal Republic of Germany. Results of empirical research, 1949-1989] Opladen: Leske + Budrich.

Bogardus, E.S. (1925). Measuring social distance. Journal of Applied Sociology, 9, 299-308.

Bogardus, E.S. (1933). A social distance scale. Sociology and Social Research, 17, 265-271.

Campbell, A.A. (1987). Randomized response technique. Science, 236, 1049.

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

Cobb, M.D. (2002). Unobtrusively measuring racial attitudes: The consequences of social desirability effects. Doctoral dissertation. Urbana-Champaign: University of Illinois (UMI No. 3023034).

Cohen Shabat, M. (1993). Prejuicio etnico en estudiantes universitarios. [Ethnic prejudice in university students] Revista Mexicana de Psicologia, 10, 183-188.

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

Fend, H. (1994). Ausländerfeindlich-nationalistische Weltbilder und Aggressionsbereitschaft bei Jugendlichen in Deutschland und der Schweiz – kontextuelle und personale Antecedensbedingung. [Hostile political attitudes toward foreigners, nationalistic thinking, and aggression in Germany and Switzerland – contextual and personality antecedents] Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 14, 131-162.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R., & Horvitz, D.G. (1969). The unrelated question randomized response model. Theoretical framework. Journal of the American Statistical Association, 64, 520-539.

Hewstone, M., & Brown, R. (1986). Contact is not enough: An intergroup perspective on the "contact hypothesis". In M. Hewstone & R. Brown (Eds.), Contact and conflict in intergroup encounters (p. 1-44). Oxford: Blackwell.

Hopf, W. (1999). Ungleichheit der Bildung und Ethnozentrismus. [Inequality of education and ethnocentrism] Zeitschrift für Pädagogik, 45, 847-865.

Hu, X., & Batchelder, W.H. (1994). The statistical analysis of general processing tree models with the EM algorithm. Psychometrika, 59, 21-47.

Jerabek, I., & de Man, A.F. (1994). Social distance among Caucasian-Canadians and Asian, Latin-American and Eastern European Immigrants in Quebec: A two-part study. Social Behavior and Personality, 22, 297-304.

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 [Political psychology of xenophobia. Victims – offenders – complices] (pp. 293-306). Weinheim: Juventa.

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

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.

Lensvelt-Mulders, G.J.L.M., Hox, J.J., & van der Heijden, P.G.M. (2005). How to improve the efficiency of randomised response designs. Quality & Quantity, 39, 253-265.

Lensvelt-Mulders, G., Hox, J., van der Heijden, P., & Maas, C. (2005). Meta-analysis of randomized-response research. Thirty-five years of validation. Sociological Methods & Research, 33, 319-348.

Ling, C. (2002). The political implications of motivation to control prejudice: Public opinion towards Blacks, Hispanics and Asians. Doctoral dissertation. Stony Brook: State University of New York (UMI No. 3023772).

Lipset, S.M. (1983). Political Man. Heinemann: London.

Manganelli Rattazzi, A.M., & Volpato, C. (2003). Social desirability of subtle and blatant prejudice scales. Psychological Reports, 92, 241-250.

Mielke, R., & Mummendey, H.D. (1995). Wenn Normen zu sehr wirken – Ausländerfeindlichkeit, Bildungsgrad und soziale Erwünschtheit [When norms work too well – xenophobia, educational level and social desirability]. Bielefelder Arbeiten zur Sozialpsychologie, 175, 1-9.

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: Pabst.

Pass, M.G. (1988). Race relations and the implications of education within prison. Journal of Offender Counseling, Services & Rehabilitation, 12, 145-151.

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

Photiadis, J.D., & Biggar, J. (1962). Religiostiy, education, and ethnic distance. American Journal of Sociology, 67, 666-672.

Ray, J.J. (1990). Racism, conservatism and social class in Australia: With German, Californian and South African comparisons. Personality and Individual Differences, 11, 187-189.

Riefer, D.M., & Batchelder, W.H. (1998). Multinomial modeling and the measurement of cognitive processes. Psychological Review, 95, 318-339.

Robinson, D., & Rohde, S. (1946). Two experiments with an anti-Semitism poll. Journal of Abnormal and Social Psychology, 41, 136-144.

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.

Stahl, C., & Klauer, K.C. (2007). HMMTree: A computer program for hierarchical multinomial processing tree models. Behavior Research Methods, 39, 267-273.

Stem, D.E., & Steinhorst, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. Journal of the American Statistical Association, 74, 555-564.

Vanneman, R.D., & Pettigrew, T. (1972). Race and relative deprivation in the urban United States. Race, 13, 461-486.

Wagner, U., van Dick, R., Pettigrew, T.F., & Christ, O. (2003). Ethnic prejudice in East and West Germany: The explanatory power of intergroup contact. Group Processes and Intergroup Relations, 6, 22-36.

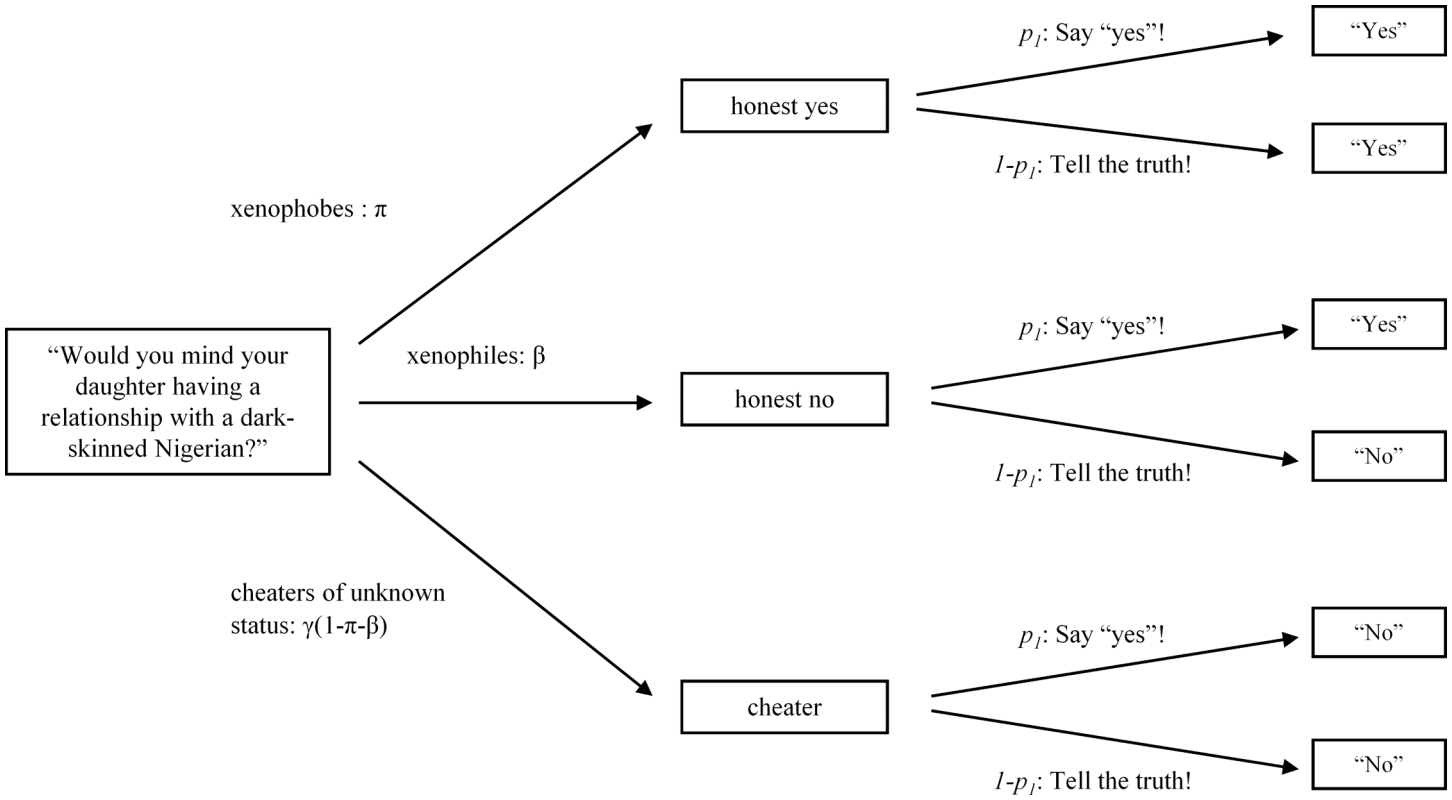
Wagner, U., & Zick, A. (1995). The relationship of formal education to ethnic prejudice: Its reliability, validity and explanation. European Journal of Social Psychology, 24, 41-56.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63-69.

Weiner, N.L. (1974). The effect of education on police attitudes. Journal of Criminal Justice, 2, 317-328.

Figures

Figure 1. A multinomial formulation of the cheating detection extension of the randomized-response-technique



Tables

Table 1. Xenophobic and xenophile attitudes by questioning mode and education

		Low education	High education
Total	N	324	282
Direct Questioning (DQ)	N	65	65
	% yes	44.6%	24.6%
	(SE)	(6.2)	(5.3)
	% no	55.4%	75.4%
	(SE)	(6.2)	(5.3)
Randomized-Response-Technique (RRT)	N	259	217
	Honest yes (π)	38.1%	30.1%
	(SE)	(5.5)	(5.8)
	Honest no (β)	23.8%	52.7%
	(SE)	(9.2)	(9.4)
	Cheaters (γ)	38.1%	17.2%
	(SE)	(5.6)	(5.5)
	ΔG^2 (df=1): $\gamma_{\text{less educated}} = 0$[1]	67.29***	
	ΔG^2 (df=1): $\gamma_{\text{highly educated}} = 0$[1]		13.23***
	ΔG^2 (df=1): $\gamma_{\text{less educated}} = \gamma_{\text{highly educated}}$[2]	6.86**	
	ΔG^2 (df=1): % $\text{yes}_{\text{less educated}}$ = % $\text{yes}_{\text{highly educated}}$ (% $\text{no}_{\text{less educated}}$ = % $\text{no}_{\text{highly educated}}$, respectively)[3]	5.81*	
	ΔG^2 (df=1): $\pi_{\text{less educated}} = \pi_{\text{highly educated}}$[4]	1.00	
	ΔG^2 (df=1): $\beta_{\text{less educated}} = \beta_{\text{highly educated}}$[4]	4.74*	

[1]High values indicate that the fit of the model worsens under the assumption that no cheating occurs ($\gamma_{\text{less educated}} = 0$ and $\gamma_{\text{highly educated}} = 0$, respectively) in this subgroup of the RRT condition.

[2]High values indicate that the fit of the model worsens when assuming that the estimated proportion of cheating participants among the less educated ($\gamma_{\text{less educated}}$) in the RRT condition does not differ from the estimated proportion of cheating participants among the highly educated ($\gamma_{\text{highly educated}}$) in the RRT condition.

[3]High values indicate that the fit of the model worsens when assuming that the proportion of xenophobic (% $\text{yes}_{\text{less educated}}$) and xenophile participants (% $\text{no}_{\text{less educated}}$), respectively, among the less educated in the DQ condition does not differ from the proportion of xenophobic (% $\text{yes}_{\text{highly educated}}$) and xenophile participants (% $\text{no}_{\text{highly educated}}$), respectively, among the highly educated in the DQ condition.

[4]High values indicate that the fit of the model worsens when assuming that the estimated proportion of xenophobic ($\pi_{\text{less educated}}$) and xenophile participants ($\beta_{\text{less educated}}$), respectively, among the less educated in the RRT condition does not differ from the estimated proportion of xenophobic ($\pi_{\text{highly educated}}$) and xenophile participants ($\beta_{\text{highly educated}}$), respectively, among the highly educated in the RRT condition.

*$p<0.05$, **$p<0.01$, ***$p<0.001$

Running head: *Attitudes towards people with disabilities*

Projective questioning overestimates the prevalence of negative attitudes towards people with physical and mental disabilities

Martin Ostapczuk[1]* and Jochen Musch[1]*

[1]Heinrich-Heine-Universitaet Duesseldorf, Germany

Word count (exc. figures/tables): 6121

*Requests for reprints should be addressed to Martin Ostapczuk or Jochen Musch, Institute of Experimental Psychology, Universitaetsstr. 1, D-40225 Düsseldorf, Germany (e-mail: martin.ostapczuk@uni-duesseldorf.de, jochen.musch@uni-duesseldorf.de).

Despite being susceptible to social desirability bias, attitudes towards people with disabilities are traditionally assessed via self-report. We investigated two methods presumably providing more valid prevalence estimates of sensitive attitudes than a direct self-report. Most People Projective Questioning (MPPQ) attempts to reduce bias by asking interviewees to estimate the number of other people holding a sensitive attribute, rather than confirming or denying the attribute for themselves. The Randomised-Response-Technique (RRT) tries to reduce bias by assuring confidentiality through a random scrambling of the respondent's answers. We validated MPPQ and RRT estimates by comparing them with a direct questioning (DQ) control condition. Estimates obtained by MPPQ exceeded the DQ estimates. Employing a cheating detection extension of the RRT, we were able to determine the proportion of respondents disregarding the RRT instructions and thus, to compute an upper bound for the prevalence of negative attitudes. MPPQ estimates exceeded this upper bound and were thus shown to overestimate the prevalence of sensitive attitudes. This result casts doubt on the validity of MPPQ estimates, and leads us to recommend the use of the more conservative RRT, which proved to be successful in reducing underreporting bias while simultaneously controlling for overestimation and non-adherence to instructions.

Attitudes towards people with disabilities have been measured for more than seven decades. Assessment methods, however, have rarely changed in all these years; self-report is still the standard data collection technique (Antonak & Livneh, 1995a, 1995b, 2000; Deal 2003; Strike, Skovholt & Hummel, 2004; Yazbeck, McVilly & Parmenter, 2004). Self-report data, however, are prone to social desirability bias when the topic of inquiry is sensitive in nature (Edwards, 1957; Hyman, 1944). Accordingly, standard inventories assessing attitudes towards people with disabilities, as for example the Attitudes Toward Disabled Persons Scale (ATDP; Yuker, Block & Young, 1970), are also susceptible to bias (Hagler, Vargo & Semple, 1987; Weisel, Kravetz, Florian & Shurka-Zernitsky, 1988). As a possible means to controlling underreporting bias, projective questioning has been proposed (Antonak & Livneh, 1995a, 1995b, 2000).

Most people projective questioning (MPPQ)

Psychoanalytic theory originally conceived projection as an ego defence mechanism. When faced with his or her own undesirable impulses, feelings, or attitudes, the individual is assumed to solve the resulting conflict by unconsciously attributing them to the outer world in order to reduce tension and anxiety (Freud, 1938). Projective techniques attempt to make use of this mechanism by allowing respondents to project their inner feelings on neutral, external stimuli. According to psychoanalytic theory, the individual need not be aware of the conflict or the projection process; rather, validity is expected to increase by disguising the real purpose of measurement (Kassarjian, 1974). Most projective questioning formats are unstructured and associative in nature.

Third person or most people projective questioning (Alpert, 1971; Smith, 1954), however, is a more structured variant of the technique. Instead of stating their own opinion when faced with a sensitive question ("Do you feel uneasy in the presence of people with disabilities?"), interviewees are being asked to estimate the agreement of other people ("Do you think most people feel uneasy…?"). Provided with such an opportunity to project their own attitudes onto others, respondents may more readily express socially undesirable attitudes while hiding behind a façade of impersonality. Possibly without even knowing it, interviewees are supposed to be more honest when questioned projectively (Simon & Simon, 1974).

Unlike other projective techniques, most people projective questioning (MPPQ) has been used extensively even beyond the realm of clinical psychology since the 1950's, especially by marketing researchers attempting to predict (unconscious) consumer motivations and sensitive behaviours (e.g. Alpert, 1971; Fisher, 1993; see Kassarjian, 1974, for a review). Additionally, researchers applied projective methods to assess sensitive or unconscious attitudes and behaviours such as discrimination against black people (Saenger & Gilbert, 1950), job satisfaction (Weitz & Nuckols, 1953; Jo, Nelson & Kiecker, 1997), motives for studying medicine (Aron, Thouvenot, Martin, Barus & Tajan, 1968), money incentives for controlling family size (Simon & Simon, 1974), doping among professional American football players (Yesalis & Courson, 1991), sexual experiences (Davoli, Perucci, Sangalli, Brancato & Dell'Uomo, 1992), and non-financial employment commitment (Snir & Harpaz, 2002). However, in spite of the urgent need for alternative attitude measures in disability research, there has been not a single study using MPPQ in this area,

and only five studies made use of any kind of projective technique in disability research so far (Antonak & Livneh, 2000).

Of course, psychoanalytic theory is controversial, and already Smith (1954) criticised that it would be naïve to presume that every time a respondent speaks about a third person, he or she is revealing something significant about him- or herself. The two psychoanalysts Murray and Morgan (1945) also acknowledged that the impersonal approach might have its drawbacks. Yet, one does not need to reason psychoanalytically in order to consider MPPQ a suitable means to reducing response bias. Alternatively, one may also assume that when being asked about most people's opinion on a sensitive topic, respondents may simply be trying to count the number of people holding the socially undesirable attitude in their vicinity, and extrapolate from this to the population at large. As in the process, the interviewees do not need to reveal anything about themselves, the resulting estimate might still be less biased by social desirability than direct self-reports (Miller, 1985). However, this latter reasoning can be criticised as being problematic as well (Marks & Miller, 1987). The accuracy of projective estimates is still necessarily limited by the respondents' actual knowledge about the behaviours and attitudes of people in their vicinity, and also dependent on their ability and willingness to extrapolate from this knowledge – even if it exists – to other people's attitudes. Due to self-serving bias (Lewicki, 1983) and overconfidence (Svenson, 1981), individuals have been argued to overestimate how prevalent negative attitudes and behaviours might be among other people. Irrespective of the underlying theoretical framework, it is therefore problematic to assume that higher prevalence

estimates of socially undesirable attitudes resulting from MPPQ are necessarily a sign of increased validity (Fisher, 1993; Snir & Harpaz, 2002). Obviously, studies comparing MPPQ estimates with an (external) criterion are needed for a convincing test of this assumption. Such studies, however, are extremely rare, and their results are only moderately compelling. Weitz and Nuckols (1953) found job-related attitudes as assessed by MPPQ clearly inferior to direct questioning (DQ) in predicting actual job survival. Their study, however, can be criticised on methodological grounds, as questioning mode was manipulated within-subjects thus making respondents acquainted with both question formats. In a second study, Alpert (1971) found DQ to be superior to MPPQ in identifying attributes determining consumer preferences. However, preferences in this study were not validated against actual consumer behaviour, and therefore remained hypothetical. In sum, in spite of its proven potential to provide higher estimates when surveying sensitive topics (Judd, Smith & Kidder, 1986; Petty & Cacioppo, 1981), the suspicion that MPPQ estimates may actually overestimate the true prevalence of socially undesirable attitudes has not been dispelled convincingly, mainly because external validation criteria that would have helped to decide on this issue have been lacking. In the present study, we tried to clarify this issue by comparing MPPQ estimates with estimates obtained using a cheating detection variant of the randomised-response-technique.

The randomised-response-technique (RRT)

While having repeatedly been suggested as an alternative to DQ in disability research (Antonak & Livneh, 1995a, 2000; Deal, 2003), the randomised-response-technique (RRT; Warner, 1965) has never been used for this

purpose. The RRT is a sophisticated self-report measure that proved helpful in yielding more valid data on a vast number of sensitive topics. The rationale of the RRT is that respondents are more honest when the confidentiality of their responses is guaranteed. In the so-called “forced-response” variant of RRT, all interviewees are confronted with the sensitive question. But before answering, a randomisation device is used to determine whether respondents are asked to answer the question honestly, or whether they are asked to provide a prespecified response (e.g. “yes”). The outcome of the randomisation procedure is unknown to the experimenter, who therefore never knows whether a “yes”-response resulted from truthful answering or from the randomisation process. The technique thus guarantees the confidentiality of responses, and arguably encourages more honest responding. Knowing the probability distribution of the randomisation device, at an aggregate level the researcher can nevertheless estimate the proportion of “yes”-answers that have not been prompted by the randomisation device. These are considered truthful avowals (Dawes & Moore, 1980; Greenberg, Abul-Ela, Simmons & Horvitz, 1969).

As surveys of sensitive behaviours and attitudes have repeatedly shown, the use of the RRT often yields more valid estimates than traditional surveys, i.e. higher estimates than estimates obtained by DQ and, more importantly, estimates diverging less from the true prevalence in cases where the true prevalence is known. The technique has been successfully used to obtain information on sensitive topics such as academic cheating, illegal drug use, abortions, shoplifting, and rape (Lensvelt-Mulders, Hox, van der Heijden & Maas, 2005). In all cases, RRT models tried to divide the population into two disjoint and exhaustive groups: respondents who engaged in the critical

behaviour, and respondents who did not. The respective proportions of these groups in the population were represented by parameters π and β, respectively. Because the two parameters are constrained to add up to 1, only one parameter had to be estimated. This could easily be done based on the one proportion of “yes”-responses that is provided in the typical RRT application.

In spite of their many successful applications, traditional RRT models have been criticised for being susceptible to cheaters, i.e. respondents who are not answering as directed by the randomisation device (Campbell, 1987). Indeed, there is evidence that such cheating occurs (Lensvelt-Mulders & Boeije, 2007; Stem & Steinhorst, 1984). To the extent that participants fail to follow the instructions by denying the sensitive behaviour or attitude in spite of being directed by the randomisation device to affirm it, the prevalence of the critical behaviour or attitude is underestimated. Clark and Desharnais (1998) therefore proposed to no longer assume that interviewees are always obeying the rules of the RRT. Instead, in what we will refer to as the cheating detection model (CDM), they presumed that some participants may deny the critical attribute in spite of being directed by the randomisation device to attest to it. Nothing can be, and is assumed in the model about whether these cheating respondents are actually holding the sensitive attitude. They may deny a socially undesirable attitude they are actually subscribing to; however, it is also possible that they just want to rule out even the slightest suspicion by denying an attitude they do not hold in spite of being told otherwise by the randomisation device. Because there is no way to distinguish between these two cases within the model, the

true status of a cheater who is not following the instructions necessarily remains unknown, and cannot be ultimately determined.

Figure 1 illustrates how the CDM can be graphically depicted as a multinomial model dividing the population into three disjoint groups: π (the proportion of non-cheating and honest “yes”-respondents, i.e. participants who truthfully admit to their negative attitudes towards people with disabilities), β (the proportion of non-cheating and honest “no”-respondents, i.e. participants with positive attitudes towards people with disabilities), and γ (= 1-π-β, the proportion of cheating participants who disobey the RRT rules by replying “no” to the critical question regardless of the outcome of the randomisation procedure).

Insert Figure 1 about here

In the CDM, there are now two independent parameters, as the proportions π, β, and γ are constrained to add up to 1. Thus, the parameters can no longer be estimated on the basis of the one proportion of “yes”-responses that is provided in traditional RRT models. Instead, to obtain a sufficient data base, it is necessary to pursue an experimental approach. Specifically, two independent samples of respondents have to be questioned with different probabilities $\underline{p}_1$ and $\underline{p}_2$ of being forced by the randomisation device to say “yes” (Clark & Desharnais, 1998; Musch, Bröder & Klauer, 2001). Figure 1 depicts only one of

these groups, in which probability $\underline{p_1}$ applies. The second group can be represented by an identical figure, with the only exception that probability $\underline{p_1}$ would have to be replaced by probability $\underline{p_2}$. Assuming that the same proportions π, β, and γ apply in both groups when participants are randomly assigned to conditions, the CDM allows us to observe two independent proportions of "yes"-responses. These two proportions suffice to estimate the two independent parameters π and β; the third parameter γ can then be computed as 1-π-β. For this particular variant of the CDM, Clark and Desharnais (1998) derived closed-form solutions for maximum likelihood estimates of the parameters π, β, and γ, as well as a statistical test of the null hypothesis that no cheating occurs (γ = 0). Their CDM provides a unique theoretical advantage over both DQ and previous RRT models: if no cheating occurs, the parameter π provides an asymptotically unbiased estimate of the population proportion holding a sensitive attitude. If, however, there is a significant proportion of cheating respondents, the CDM still allows us to compute both a lower and upper bound by assuming that cheating respondents – whose real attitude cannot be ultimately determined – either all do not or do hold the sensitive attitude (Musch et al., 2001). The CDM can also be subsumed under the more general family of multinomial models, for which Hu and Batchelder (1994) and Riefer and Batchelder (1988) have developed statistical procedures. Using a multinomial modeling framework, it is easily possible to test substantively motivated parameter restrictions, such as the assumption that no cheating occurs (γ = 0). This is done by testing whether the proportion γ of cheating respondents can be set equal to zero without significant loss in the goodness of fit of the model. It is also possible to formulate more

complex models incorporating additional parameters (Musch et al., 2001). These may represent, for instance, subgroups for which parameters have to be estimated separately. We took advantage of this possibility by breaking down the sample by social desirability scores; comparing respondents scoring high vs. low on two dimensions of a social desirability scale, we were able to investigate the possible influence of this variable on attitudes towards people with disabilities. In line with previous research, we also tested for possible influences of sex, age, and education in suitably expanded multinomial models.

Integration of MPPQ and RRT

Several authors have suggested to combine direct attitudes measures, such as DQ and RRT, with indirect ones, such as MPPQ, in order to increase the validity of the measurement of sensitive attitudes (Jo et al., 1997; Umesh & Peterson, 1991). With regard to attitudes towards people with disabilities in particular, Antonak and Livneh (1995) argued that “RRT might be effectively combined with an indirect response method and a direct response method to measure disability attitudes. Such an investigation could afford the researcher with a unique opportunity to investigate the transcontextual stability of these attitudes in addition to providing a better understanding of the convergent and discriminant validities of these three attitude measurement approaches and their susceptibility to respondent biases” (p. 140). Unfortunately, studies pursuing such an integrative approach have been rare; we found only two studies directly comparing RRT and MPPQ. For a number of different sensitive behaviours, Bégin and Boivin (1980) found MPPQ estimates to be consistently higher than DQ and RRT estimates. They concluded that MPPQ overestimates

the true prevalence of sensitive attributes, and is therefore unsuited for reducing social desirability bias. It is important to note, however, that using a traditional variant of the RRT, they were not able to determine whether their RRT estimates were lowered by cheating respondents who disobeyed the instructions. If this was the case, the MPPQ estimates they obtained might actually have been valid. In a second study, Armacost, Hosseini, Morris and Rehbein (1991) surveyed a sample of chief executive officers about legal, but ethically questionable activities in their companies. In spite of observing similar results as Bégin and Boivin (1980), these authors favoured a different interpretation and suggested that the large gap between DQ / RRT estimates on the one hand, and MPPQ estimates on the other hand, might be due to serious underreporting in the DQ and RRT conditions. They therefore argued that estimates obtained by MPPQ might actually represent a reasonable upper bound for the prevalence of a sensitive attribute (Armacost et al., 1991).

Both of the above interpretations are defensible, and the empirical results do not allow us to distinguish between them, since both studies used an RRT variant not capable of determining the amount of cheating, i.e. the proportion of participants not following the instructions. As outlined above, however, a cheating detection extension of the RRT would have provided a unique advantage over the RRT models that have actually been employed; namely, it would have allowed us to compute an upper bound for the sensitive attribute ($\pi+\gamma$). In the present study, we took advantage of this possibility in order to test the two conflicting interpretations by comparing estimates of the prevalence of negative attitudes towards people with disabilities obtained by RRT with

estimates obtained by MPPQ. If MPPQ overshoots when used to control for response bias in surveys of socially undesirable attitudes, and leads to overestimates as was suggested by Bégin and Boivin (1980), we would expect MPPQ estimates to significantly exceed even an upper bound determined by the cheating detection variant of the RRT. If, however, MPPQ just effectively reduces social desirability bias, it should always provide estimates that are falling below this upper bound.

In order to make results comparable with traditional self-report surveys on attitudes towards people with disabilities, we included a DQ control condition in our survey and incorporated it into the multinomial model. In total, our suitably expanded multinomial model thus consisted of four different groups: an RRT group with randomisation probability p_1 (RRT1), an RRT group with randomisation probability p_2 (RRT2), an MPPQ group, and a DQ control group.

Generality vs. specificity of attitudes towards people with disabilities

Apart from the methodological discussion of how best to avoid social desirability bias in disability attitude research, another substantive area of discussion in the field has been whether people tend to hold attitudes that vary from one disabling condition to another, or whether they hold generalised attitudes across different disabilities (Gething, 1991). Although there is a vast amount of criteria for classifying diverse disabling conditions (cf. Feldman & Crandall, 2007; Gething, 1991), the classification of physical vs. mental disabilities is perhaps the most often used and accepted one (cf. Deal, 2003). While some survey results seem to imply that attitudes towards people with physical vs. mental

disabilities differ, with persons with mental disability facing more negative attitudes (e.g. Tringo, 1970), other studies found no differences between these two groups (e.g. Gething, 1991). In his summarising review, Deal (2003) therefore concluded that empirical evidence is not conclusive, but slightly favoured the specificity assumption. In the present study, we tested the generality vs. specificity assumption by presenting our respondents the same sensitive attitude item twice, once in a physical and once in a mental disability version.

Method

Participants

Participants were 1300 registered users of an online panel who had previously agreed to participate in exchange for a financial incentive of about 0.75 Euro. There were 71 participants who did not complete the study; however, a Chi-square test confirmed that drop-out rates did not differ between experimental conditions (MPPQ: 2.8%, RRT1: 5.8%, RRT2: 7.0%, DQ: 4.2%; $X^2(3) = 5.91$, ns). Another 69 participants identified as being physically or mentally handicapped beyond the degree of severity of visual impairment were also excluded from further analyses. Thus, the final data set consisted of 1160 respondents; 200 participated in the MPPQ condition, 383 in the RRT1 condition, 385 in the RRT2 condition and 192 in the DQ condition. Of the participants, 581 (50%) were female, 617 (53%) were highly educated (a participant was considered highly educated, if he or she had acquired the German “Abitur”, an advanced high school diploma allowing the student to commence studies at the university), 670 (58%) had at least some experience

with people with physical disability (i.e. they knew at least one person with a physically disabling condition), and 508 (44%) had at least some experience with people with mental disability. Age of the respondents ranged from 18 to 81 years, with a mean of M = 36.0 (SD = 11.0) years.

Materials and design

Participants responded to a variety of demographic questions and questions related to their personal experiences with people with physical and mental disabilities, only two of which, however, were sensitive in nature and central for the present study. Participants also completed the German adaptation of the Balanced Inventory of Desirable Responding (BIDR-G; Musch, Brockhaus & Bröder, 2002) in a slightly modified version by Eichstaedt and Musch (2007). The BIDR-G measures the two dimensions of social desirability identified by Paulhus (1998), self-deceptive enhancement (10 items), and impression management (10 items). High scores on both scales indicate a greater tendency to select socially desirable responses. The psychometric indices of the German adaptation of the BIDR have been found to be satisfactory (Musch et al., 2002; Eichstaedt & Musch, 2007). The BIDR-G was included to test the effects of social desirability on attitudes towards people with disabilities.

The critical sensitive item measuring negative attitudes towards individuals with disability was adapted from a study of Yazbeck et al. (2004), and asked whether respondents felt uneasy in the presence of a disabled person. To test the generality vs. specificity assumption, we created physical and mental disability versions of this item which read, “Do you feel uneasy in the presence of people

with physical disabilities?" and "Do you feel uneasy in the presence of people with mental disabilities?", respectively.

To compensate for the loss of efficiency in parameter estimation associated with the use of the RRT, participants were randomly assigned to one of the four conditions by a ratio of 1:2:2:1 (MPPQ : RRT1 : RRT2 : DQ). In the DQ baseline condition, respondents were simply requested to reply "yes" or "no" to the two critical questions. In the MPPQ condition, respondents were requested to answer the same questions asked in a projective format. The relevant questions read, "Do you think most people feel uneasy in the presence of people with physical disabilities?" and "Do you think most people feel uneasy in the presence of people with mental disabilities?", respectively. In the two remaining conditions, the same sensitive questions were asked in a randomised-response format. Two RRT conditions were needed, as the CDM requires two groups with different randomisation probabilities $\underline{p}_1$ and $\underline{p}_2$. In the low probability group, instructions for the physical disability item read: "If your mother was born in February, March or April, then please reply 'yes' to the following question independently of its content. If your mother, however, was born in another month, then please answer truthfully." The probability of being forced to say "yes" thus approximated $\underline{p}_1 = 3/12 = 1/4 = 0.25$, as shown by birth statistics made available to us by the German Federal Agency for Statistics. In the high probability condition, the following instruction was given for the physical disability item: "If your mother was born in February, March or April, then please answer the following question truthfully. If, however, your mother was born in another month, then please answer 'yes' irrespective of the content of the

question." The probability p_2 of being forced to say "yes" thus approximated 1.00-p_1 = 9/12 = 3/4 = 0.75 in the high probability condition. Detailed instructions explained that owing to the randomisation procedure, the confidentiality of responses was guaranteed, because individual answers could no longer be linked to the respondents' true status with regard to the sensitive attitude.

For the second question concerning attitudes towards the mentally disabled, we changed both the person whose birthday was used for the randomisation procedure (now asking for the month of birth of the participants' father, rather than that of their mother) and the relevant months of birth (now referring to January, February, and March, rather than to February, March, and April). As we explained them in detail, respondents were thus able to answer the second question without revealing anything about their true status with regard to the first question (cf. Kulka, Weeks & Folsom, 1981). The resulting randomisation probabilities, however, remained the same, as confirmed by the birth statistics mentioned above, i.e. p_1 = 0.25 in the low probability and p_2 = 0.75 in the high probability group, respectively.

Statistical analysis

Based on the frequency of "yes"- and "no"-responses to the sensitive questions, we computed maximum likelihood estimates of the multinomial model parameters in the different conditions using the programme HMMTree (Stahl & Klauer, 2007) which implements the EM-algorithm (Hu & Batchelder, 1994). Model fit and parameter restrictions were assessed by the asymptotically chi-square distributed log-likelihood ratio statistic G^2. The basic multinomial model was saturated, as the two proportions of observable "yes"-responses in the two

RRT conditions just suffice to estimate the two independent parameters π and β (with γ = 1-π β). To assess the significance of parameter restrictions, such as γ = 0 (no cheating in the RRT condition) or $\pi_{male} = \pi_{female}$ (no influence of sex on disability attitudes reported by RRT), we used the difference between the log-likelihood ratio statistic for restricted and unrestricted versions of the model (ΔG^2; Riefer & Batchelder, 1988).

Results

Main analyses

Table 1 shows the parameter estimates in the different experimental conditions for both the physical and the mental disability version of the sensitive attitude item. When questioned directly, only 7.8% of the participants admitted to feeling uncomfortable in the presence of people with physical disabilities. The estimate obtained by RRT ($\pi_{physical}$ = 10.8%) was slightly, but not significantly higher; restricting the model by assuming that the proportion of "yes"-answers to the direct question was the same as the estimated proportion of honest "yes"-answers in the RRT condition (% DQ $yes_{physical} = \pi_{physical}$) did not deteriorate the fit of the model, $\Delta G^2(1) = 0.56$, ns. However, using the cheating detection variant of the RRT allowed us to discover that there was a considerable proportion of cheating respondents who disobeyed the rules of the RRT by denying the sensitive attribute regardless of the outcome of the randomisation procedure, $\gamma_{physical}$ = 33.3%. The proportion of cheaters was significantly higher than zero as shown by assuming that no cheating occurred ($\gamma_{physical} = 0$), which significantly worsened the fit of the model, $\Delta G^2(1) = 88.83$, $p<.01$. As explained

above, the CDM does not make an assumption, and therefore cannot decide whether cheating respondents actually do or do not feel uneasy in the presence of people with physical disabilities. However, by alternately assuming – in a best and a worst case scenario – that either none or all of the cheating respondents actually do hold the sensitive attitude, the model allowed us to compute a lower bound of $\pi_{physical}$ = 10.8% and an upper bound of $\pi_{physical}+\gamma_{physical}$ = 10.8%+33.3% = 44.1% for the true prevalence of negative attitudes towards people with physical disabilities. The MPPQ estimate (%MPPQ $yes_{physical}$ = 54.5%) significantly exceeded both the DQ estimate of 7.8% ($\underline{\Delta G^2}(1) = 108.35$, $\underline{p}<.01$) and the RRT estimate of 10.8% ($\underline{\Delta G^2}(1) = 70.95$, $\underline{p}<.01$). Importantly, the upper bound of 44.1% determined by the cheating detection variant of the RRT also fell below the MPPQ physical disability estimate of 54.5% and in fact, even below the lower bound of the confidence interval of the MPPQ estimate (95% $CI_{MPPQ\ physical}$ = 47.6-61.4%).

Turning to the mental disability variant of the sensitive item, 27.1% of the respondents admitted feeling uncomfortable in the presence of people with mental disabilities when questioned directly. RRT provided a similar, not significantly different estimate (π_{mental} = 24.2%, $\underline{\Delta G^2}(1) = 0.33$, $\underline{ns}$). However, again a considerable proportion of respondents was found to disregard the instructions of the RRT; the proportion of cheating respondents (γ_{mental} = 22.2%) was again significantly higher than zero, $\underline{\Delta G^2}(1) = 38.28$, $\underline{p}<.01$. Depending on whether these cheating respondents actually do not or do feel uneasy in the presence of people with mental disabilities, we computed a lower bound of π_{mental} = 24.2% and an upper bound of $\pi_{mental}+\gamma_{mental}$ = 24.2%+22.2% = 46.4%

for the true prevalence of negative attitudes towards people with mental disabilities. Like with physical disability, the MPPQ estimate (% MPPQ yes_{mental} = 79.0%) significantly exceeded both the DQ estimate of 27.1% ($\Delta G^2(1)$ = 111.55, $p<.01$) and the RRT estimate of 24.2% ($\Delta G^2(1)$ = 109.28, $p<.01$). Furthermore, the RRT upper bound of 46.4% again fell well below even the lower bound of the confidence interval of the MPPQ mental disability estimate (95% $CI_{MPPQ\ mental}$ = 73.4-84.6%).

Insert Table 1 about here

As table 1 shows, comparing the prevalence of negative attitudes towards people with physical vs. mental disabilities showed more negative attitudes towards people with mental disabilities in the DQ and MPPQ condition, and also marginally more negative attitudes towards the mentally handicapped in the RRT condition (for which the 95% confidence intervals still overlap, however).

Post-hoc analyses

Previous studies have shown that demographic variables and personal experience with individuals with handicap may influence attitudes towards people with disabilities (Beckwith & Matthews, 1994; Loo, 2004; Yazbeck et al., 2004). We therefore conducted a series of post-hoc analyses investigating the potential moderating role of respondents' sex, age, education, and their

experience with people with physical and mental disabilities. However, we did not find any systematic effects of these variables.

Finally, we tested the influence of different facets of social desirability (self-deceptive enhancement, SDE, vs. impression management, IM) on attitudes towards people with disabilities. Interestingly, under DQ conditions, respondents scoring below the median on SDE were significantly more likely to admit negative attitudes towards people with physical disabilities (% DQ $yes_{physical,\ SDE\ low}$ = 12.5% vs. % DQ $yes_{physical,\ SDE\ high}$ = 4.5%; $\underline{\Delta G^2}(1) = 4.13$, $\underline{p}<.05$). The same pattern emerged for attitudes towards people with mental disabilities (% DQ $yes_{mental,\ SDE\ low}$ = 37.5% vs. DQ $yes_{mental,\ SDE\ high}$ = 19.6%; $\underline{\Delta G^2}(1) = 7.47$, $\underline{p}<.01$). None of the remaining estimates differed significantly between respondents scoring low vs. high on the SDE scale (see table 2).

Insert Table 2 about here

Concerning the IM scale, participants scoring low on IM were significantly more likely to obey the RRT rules ($\gamma_{physical,\ IM\ low}$ = 24.5%) than participants scoring high on IM ($\gamma_{physical,\ IM\ high}$ = 43.3%; $\underline{\Delta G^2}(1) = 5.62$, $\underline{p}<.05$) when questioned about people with physical disabilities. The same pattern also emerged for self-reported attitudes towards people with mental disabilities ($\gamma_{mental,\ IM\ low}$ = 13.7% vs. $\gamma_{mental,\ IM\ high}$ = 31.6%; $\underline{\Delta G^2}(1) = 5.35$, $\underline{p}<.05$). Additionally and paralleling the pattern found for SDE, respondents scoring low in IM admitted negative

attitudes more freely in the DQ condition than respondents scoring high in IM (% DQ $yes_{mental,\ IM\ low}$ = 38.0% vs. % DQ $yes_{mental,\ IM\ high}$ = 17.0%, $\Delta G^2(1) = 10.88$, $p<.01$). The corresponding difference in attitudes towards people with physical disabilities failed however to reach conventional levels of statistical significance (% DQ $yes_{physical,\ IM\ low}$=10.9% vs. % DQ $yes_{physical,\ IM\ high}$ = 5.0%; $\Delta G^2(1) = 2.32$, ns). All remaining estimates did not significantly differ either (see table 2).

Discussion

Obtaining unbiased measures of attitudes towards people with disabilities is difficult, but of considerable importance for intervention programmes designed to remove material and immaterial barriers people with disabilities are confronted with in their daily lives. For this reason, advanced measures such as MPPQ and RRT deserve to be evaluated with regard to their utility to reduce social desirability bias in disability research. As existing evidence on the validity of MPPQ (Alpert, 1971; Smith, 1954) is rare and inconclusive, we explored the utility of the technique by validating it against the RRT, which has repeatedly been shown to be capable of reducing bias in surveys on sensitive topics. In order to determine an upper limit to the prevalence of negative attitudes against which MPPQ estimates could be validated, we employed the cheating detection variant of the RRT (Clark & Desharnais, 1998; Musch et al., 2001). We were thus able to test whether MPPQ overestimates the true prevalence of negative attitudes towards people with disabilities.

In a large survey conducted via an online panel, we found MPPQ estimates to consistently exceed not only DQ, but also RRT estimates, in particular with regard to self-reported attitudes towards people with mental disabilities. This was true even in a worst case scenario, in which an upper limit to the prevalence of negative attitudes was established in the RRT condition by assuming that all participants disobeying the instructions were actually holding the sensitive attribute. We therefore have to conclude that MPPQ provided inflated and unjustified estimates in the current study. This troubling result casts serious doubt on the validity of MPPQ estimates, and leads us to caution other researchers that by employing MPPQ, they are running a serious risk of wrongfully overestimating the prevalence of sensitive attributes. In sum, we conclude that MPPQ does not seem to be a suitable means to reducing social desirability bias in disability attitude research.

In contrast to MPPQ, estimates of the prevalence of negative attitudes obtained by employing the RRT were lower, and well within the range of those reported by other researchers using similar items (cf. Yazbeck et al., 2004). Additionally, using a cheating detection extension to the RRT, we were able to estimate the proportion of respondents who disobeyed the instructions, and could thereby compute both a lower and an upper bound for the prevalence of negative attitudes. The cheating detection extension of the RRT thus proved to be successful in controlling for both overestimation on the part of the researcher and non-adherence to instructions on the part of the participants.

Investigating personality correlates of attitudes towards people with disabilities, we found that participants scoring high on SDE reported themselves to be less prejudiced. This effect was however limited to the DQ condition, and the respective difference did not emerge under RRT conditions. On the other hand, participants scoring high on IM not only reported themselves to be less prejudiced against people with mental disabilities, they also purposely disregarded the RRT rules more often, possibly in an attempt to better mask their true opinion. To the best of our knowledge, this finding provides the first and admittedly preliminary evidence for a potential role of personality in dealing with an RRT questioning procedure. Future research should investigate the possible influence of individual differences on RRT estimates more closely.

Results regarding the generality vs. specificity assumption (Deal, 2003) showed that negative attitudes were more prevalent towards people with mental as compared to physical disabilities. This was true for both the DQ and MPPQ condition. Even though the same descriptive difference marginally failed to reach significance under RRT conditions, the overall pattern of results is clearly easier to reconcile with the specificity assumption. We therefore argue that effective intervention programmes aiming at the reduction of negative attitudes towards people with disabilities should best take such differences into account, rather than treating the target population of “people with disabilities” as a homogeneous group.

Finally, some limitations of the present study should also be acknowledged. First, we do not claim that the estimates we obtained in our online panel are

representative for the German population at large, even though we surveyed a sample that was much more heterogeneous in education and age than the typical student sample. Second, the MPPQ we employed is only one out of several variants of projective questioning. Demonstrating its inaptness to adequately control social desirability bias does not allow us to discard projective techniques in general. Third, our use of the RRT can be criticised for several reasons. First, RRT cannot be used to directly determine the status of an individual. But of course, it is exactly this feature that makes the confidential nature of the technique credible, and encourages interviewees to respond honestly. Moreover, in spite of this limitation, Rittenhouse (1996a, 1996b) has demonstrated that – even though only in a probabilistic way – respondent-specific information is obtainable via the RRT. Sophisticated logistic regression techniques accounting for the random error introduced by the randomisation procedure have been successfully developed and used to correlate individual answers with background explanatory variables (Maddala 1983; van der Heijden, van Gils, Bout, & Hox, 2000).

Second, all RRT models introduce random error and therefore have lower efficiency, i.e. greater sampling variation than a direct question. Although we used the most efficient RRT variant currently available (Lensvelt-Mulders, Hox & van der Heijden, 2005), using the RRT still requires larger samples than either a direct or MPPQ survey. Obviously, the low efficiency of the RRT is compensated by a reduction of response bias only when questions of a sufficiently sensitive nature are being asked, as was the case in the present study.

As another drawback, unlike traditional disability attitude scales, the RRT variant we employed is capable of assessing dichotomous attributes only. Moreover, RRT surveys are somewhat more time-consuming and slightly more complicated than MPPQ or DQ because a randomisation device has to be used and prior to answering, instructions have to be read by the respondents.

In spite of these drawbacks, however, the results of the present study have clearly shown that unlike MPPQ, the cheating detection variant of the RRT shows considerable promise as a means to reducing social desirability bias. On the other hand, naïvely taking the results of a direct question at face value seems to be highly questionable in view of the present results. We therefore recommend that direct questions should no longer be used, nor trusted, as the sole basis for estimating the prevalence of sensitive attitudes. Rather, we strongly recommend to employ a cheating detection approach to the RRT to gain additional insight in future studies on attitudes towards people with disabilities.

Acknowledgements

We would like to thank Dennis Winter, Martin Grupe and Sven Keiner for their help in collecting the data as well as Morten Moshagen for helpful comments on an earlier version of the manuscript. This work was supported by grant MU 2674/1-1 from the Deutsche Forschungsgemeinschaft (German Research Foundation).

References

Alpert, M. (1971). Identification of determinant attributes: A comparison of methods. Journal of Marketing Research, 8, 184-191.

Antonak, R. F., & Livneh, H. (1995a). Randomized-Response Technique: A review and proposed extension to disability attitude research. Genetic, Social, and General Psychology Monographs, 121, 99-145.

Antonak, R., & Livneh, H. (1995b). Direct and indirect methods to measure attitudes toward persons with disabilities, with an exegesis of the error choice test method. Rehabilitation Psychology, 40, 3-24.

Antonak, R. F., & Livneh, H. (2000). Measurement of attitudes towards persons with disabilities. Disability and Rehabilitation, 22, 211-224.

Armacost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information. Decision Sciences, 22, 1073-1090.

Aron, E., Thouvenot, J., Martin, A., Barus, J., & Tajan, A. (1968). La vocation médicale: Enquête auprès de 239 étudiants de 1er année. [The medical profession: A study of 239 1st year students.] Annales Médico-Psychologiques, 2, 493-504.

Beckwith, J. B., & Matthews, J. M. (1994). Measuring comfort in interacting with people with intellectual disabilities. Australian Journal of Psychology, 46, 53-57.

Bégin, G., & Boivin, M. (1980). Comparison of data gathered on sensitive questions via direct questionnaire, randomized response technique, and a projective method. Psychological Reports, 47, 743-750.

Campbell, A. A. (1987). Randomized response technique. Science, 236, 1049.

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

Davoli, M., Perucci, C. A., Sangalli, M., Brancato, G., & Dell'Uomo, G. (1992). Reliability of sexual behavior data among high school students in Rome. Epidemiology, 3, 531-535.

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

Deal, M. (2003). Disabled people's attitudes toward other impairment groups: A hierarchy of impairments. Disability and Society, 18, 897-910.

Eichstaedt, J., & Musch, J. (2007). Verbesserung der Akzeptanzwerte einer Skala zur Erfassung von Selbst- und Fremdtäuschung als Teilaspekte sozialer Erwünschtheit. [Improving acceptance of a scale assessing self-deceptive enhancement and impression management as dimensions of social desirability]. Talk at the 9th conference of the Differential Psychology section of the German Psychological Society, Vienna, Austria.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research. New York: Dryden.

Feldman, D. B., & Crandall, C. S. (2007). Dimensions of mental illness stigma: What about mental illness causes social rejection? Journal of Social and Clinical Psychology, 26, 137-154.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. Journal of Consumer Research, 20, 303-315.

Freud, S. (1938). Totem and taboo. In A. A. Brill (Ed.), The basic writings of Sigmund Freud (pp. 807-930). New York: Random House.

Gething, L. (1991). Generality vs. specificity of attitudes towards people with disabilities. British Journal of Medical Psychology, 64, 55-64.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model. Theoretical framework. Journal of the American Statistical Association, 64, 520-539.

Hagler, P., Vargo J., & Semple, J. (1987). The potential for faking on the Attitudes Toward Disabled Persons Scale. Rehabilitation Counseling Bulletin, 31, 72-76.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. Psychometrika, 59, 21-47.

Hyman, H. (1944). Do they tell the truth? Public Opinion Quarterly, 8, 557-559.

Jo, M.-S., Nelson, J. E., & Kiecker, P. (1997). A model for controlling social desirability bias by direct and indirect questioning. Marketing Letters, 8, 429-437.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1986). Research methods in social relations, 6th edition. Fort Worth: Harcourt Brace Jovanovich College Publishers.

Kassarjian, H. H. (1974). Projective Methods. In R. Ferber (Ed.), Handbook of Marketing research (pp. 3-85 to 3-100). New York: McGraw-Hill.

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, Research Triangle Institute, North Carolina.

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.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., & van der Heijden, P. G. M. (2005). How to improve the efficiency of randomised response designs. Quality & Quantity, 39, 253-265.

Lensvelt-Mulders, G., Hox, J., van der Heijden, P., & Maas, C. (2005). Meta-analysis of randomized-response research. Thirty-five years of validation. Sociological Methods & Research, 33, 319-348.

Lewicki, P. (1983). Self-image bias in person perception. Journal of Personality and Social Psychology, 45, 384-393.

Loo, R. (2004). Attitudes toward employing persons with disabilities: A test of the sympathy-discomfort categories. Journal of Applied Social Psychology, 34, 2200-2214.

Maddala, G. S. (1983). Limited dependent and qualitative variables in econometrics. New York: Cambridge University Press.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. Psychological Bulletin, 107, 77-90.

Miller, J. D. (1985). The nominative technique: A new method of estimating heroin prevalence. In B. A. Rouse, N. J. Kozel & L.G. Richards (Eds.), Self-report methods of estimating drug use. Meeting current challenges to validity (pp. 104-124). Washington: Government Printing Office.

Murray, H. A., & Morgan, C. D. (1945). A clinical study of sentiments. Genetic Psychology Monographs, 32, 3-311.

Musch, J., Brockhaus, R., & Bröder, A. (2002). Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit. [An inventory for the assessment of two factors of social desirability] 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 (Eds.), Dimensions of Internet Science (pp. 179-192). Lengerich: Pabst.

Paulhus, D. L. (1998). Paulhus Deception Scales (PDS): The Balanced Inventory of Desirable Responding-7 User's Manual. North Tonawanda / Toronto: Multi-Health Systems.

Petty, R. E., & Cacioppo, J. T. (1981). Attitudes and persuasion: Classic and contemporary approaches. Dubuque: William C. Brown.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. Psychological Review, 95, 318-339.

Rittenhouse, B. E. (1996a). A novel compliance assessment technique. The randomized response interview. International Journal of Technology Assessment in Health Care, 12, 498-510.

Rittenhouse, B. E. (1996b). Respondent-specific information from the randomized response interview: Compliance assessment. Journal of Clinical Epidemiology, 49, 545-549.

Saenger, G., & Gilbert, E. (1950). Customer reactions to the integration of Negro sales personnel. International Journal of Opinion and Attitude Research, 4, 57-76.

Simon, R. J., & Simon, J. L. (1974). The effect of money incentives on family size: A hypothetical-question study. Public Opinion Quarterly, 38, 585-595.

Smith, G. H. (1954). Motivation research in advertising and marketing. New York: McGraw-Hill.

Snir, R., & Harpaz, I. (2002). To work or not to work: Non-financial employment commitment and the social desirability bias. Journal of Social Psychology, 142, 635-644.

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for hierarchical multinomial processing tree models. Behavior Research Methods, 39, 267-273.

Stem, D. E., & Steinhorst, R. K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. Journal of the American Statistical Association, 74, 555-564.

Strike, D. L., Skovholt, T. M., & Hummel, T. J. (2004). Mental health professionals' disability competence: Measuring self-awareness, perceived knowledge, and perceived skills. Rehabilitation Psychology, 49, 321-327.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? Acta Psychologica, 47, 143-148.

Tringo, J. L. (1970). The hierarchy of preference toward disability groups. Journal of Special Education, 4, 295-306.

Umesh, U. N., & Peterson, R. A. (1991). A critical evaluation of the randomized response method. 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.

Weisel, A., Kravetz, S., Florian, V., & Shurka-Zernitsky, E. (1988). The structure of attitudes toward persons with disabilities: An Israeli validation of Siller's Disability Factor Scales-General (DFS-G). Rehabilitation Psychology, 33, 227-238.

Weitz, J., & Nuckols, R. C. (1953). The validity of direct and indirect questions in measuring job satisfaction. Personnel Psychology, 6, 387-494.

Yazbeck, M., McVilly, K., & Parmenter, T. R. (2004). Attitudes toward people with intellectual disabilities: An Australian perspective. Journal of Disability Policy Studies, 15, 97-111.

Yesalis, C. E., & Courson, S. P. (1991). Anabolic steroid use among self-selected sample of NFL players. In S. Courson & L. R. Schreiber (Eds.), False Glory: Steelers and steroids. The Steve Courson Story (pp. 205-215). Stamford: Longmeadow.

Yuker, H. E., Block. J. R., & Young, J. H. (1970). The measurement of attitudes toward disabled persons (Rehabilitation Series No. 3). Albertson: Human Resources Center.

Tables

Table 1. Negative attitudes towards people with physical and mental disabilities by questioning mode

		Physical disability	Mental disability	95% $CI_{physical}$ < 95% CI_{mental}?[5]
N		1160		
DQ	% yes	7.8%	27.1%	Yes
	[95% CI]	[4.0-11.6]	[20.8-33.4]	(physical < mental)
RRT	Honest yes (π)	10.8%	24.2%	No
	[95% CI]	[3.9-17.8]	[16.5-31.8]	(physical = mental)
	Honest no (β)	55.9%	53.6%	
	[95% CI]	[42.6-69.0]	[40.0-67.3]	
	Cheaters (γ)	33.3%	22.2%	
	[95% CI]	[25.5-41.1]	[14.6-29.7]	
	Upper bound (π+γ)	44.1%	46.4%	
MPPQ	% yes	54.5%	79.0%	Yes
	[95% CI]	[47.6-61.4]	[73.4-84.6]	(physical < mental)
	$\Delta G^2(1)$: γ = 0[1]	88.83**	38.28**	
	$\Delta G^2(1)$: % DQ yes = π[2]	0.56	0.33	
	$\Delta G^2(1)$: % MPPQ yes = π[3]	70.95**	109.28**	
	$\Delta G^2(1)$: % MPPQ yes = % DQ yes[4]	108.35**	111.55**	

CI = confidence interval; DQ = direct questioning; RRT = randomised-response-technique; MPPQ = most people projective questioning.

[1]High values indicate that the fit of the model worsens when assuming that no cheating occurs (γ = 0) in the respective subgroup.

[2]High values indicate that the fit of the model worsens when assuming that the DQ estimate of participants holding a negative attitude (% DQ yes) does not differ from the estimated proportion of honest “yes”-answers (π) in the RRT condition.

[3]High values indicate that the fit of the model worsens when assuming that the MPPQ estimate of participants holding a negative attitude (% MPPQ yes) does not differ from the estimated proportion of honest “yes”-answers (π) in the RRT condition.

[4]High values indicate that the fit of the model worsens when assuming that the MPPQ estimate of participants holding a negative attitude (% MPPQ yes) does not differ from the corresponding estimate obtained by DQ (% DQ yes).

[5]Reflecting whether the 95% confidence intervals for the physical and mental disability version overlap in the respective questioning mode.

Table 2. Negative attitudes towards people with physical and mental disabilities by social desirability

		Physical disability		Mental disability		Physical disability		Mental disability	
		SDE low	SDE high	SDE low	SDE high	IM low	IM high	IM low	IM high
	N	584	576	584	576	603	557	603	557
DQ	% yes	12.5%	4.5%	37.5%	19.6%	10.9%	5.0%	38.0%	17.0%
	[95% CI]	[5.3-19.7]	[0.6-8.2]	[26.9-48.1]	[12.3-27.0]	[4.5-17.2]	[0.7-9.3]	[28.1-48.0]	[9.6-24.4]
RRT	Honest yes (π)	9.3%	12.5%	28.5%	18.9%	6.7%	15.6%	23.6%	24.7%
	[95% CI]	[-0.1-18.7]	[2.2-22.9]	[18.0-39.1]	[7.8-30.0]	[-2.6-16.0]	[5.2-26.0]	[13.1-34.0]	[13.5-36.0]
	Honest no (β)	60.5%	51.0%	48.6%	60.0%	68.8%	41.1%	62.7%	43.7%
	[95% CI]	[42.3-78.7]	[31.8-70.2]	[29.6-67.5]	[40.3-79.5]	[51.0-86.6]	[21.6-60.6]	[44.5-80.9]	[23.4-63.9]
	Cheaters (γ)	30.2%	36.5%	22.9%	21.1%	24.5%	43.3%	13.7%	31.6%
	[95% CI]	[19.3-41.1]	[25.4-47.6]	[12.3-33.5]	[10.3-31.9]	[13.9-35.0]	[31.9-54.7]	[3.8-23.7]	[20.3-43.0]
	Upper bound (π+γ)	39.5%	49.0%	51.4%	40.0%	31.2%	58.9%	37.3%	56.3%
MPPQ	% yes	55.4%	53.5%	79.2%	78.8%	51.0%	57.8%	83.7%	74.5%
	[95% CI]	[45.8-65.1]	[43.7-63.4]	[71.3-87.1]	[70.7-86.8]	[41.1-60.9]	[48.3-67.4]	[76.4-91.0]	[66.1-83.0]
	$\Delta G^2(1)$: $\gamma_{low} = \gamma_{high}$ [1]	0.61		0.05		5.62*		5.35*	
	$\Delta G^2(1)$: $\pi_{low} = \pi_{high}$ [2]	0.21		1.50		1.56		0.02	
	$\Delta G^2(1)$: % MPPQ yes_{low} = % MPPQ yes_{high} [3]	0.07		0.00		0.94		2.55	
	$\Delta G^2(1)$: % DQ yes_{low} = % DQ yes_{high} [4]	4.13*		7.47**		2.32		10.88**	

CI = confidence interval; DQ = direct questioning; RRT = randomised-response-technique; MPPQ = most people projective questioning; SDE = self-deceptive enhancement; IM = impression management.

[1]High values indicate that the fit of the model worsens when assuming that the estimated proportion of cheaters in the RRT condition differs between respondents scoring low vs. high on SDE ($\gamma_{low} = \gamma_{high}$).

[2]High values indicate that the fit of the model worsens when assuming that the estimated proportion of participants holding a negative attitude towards the disabled in the RRT condition differs between respondents scoring low vs. high on SDE ($\pi_{low} = \pi_{high}$).

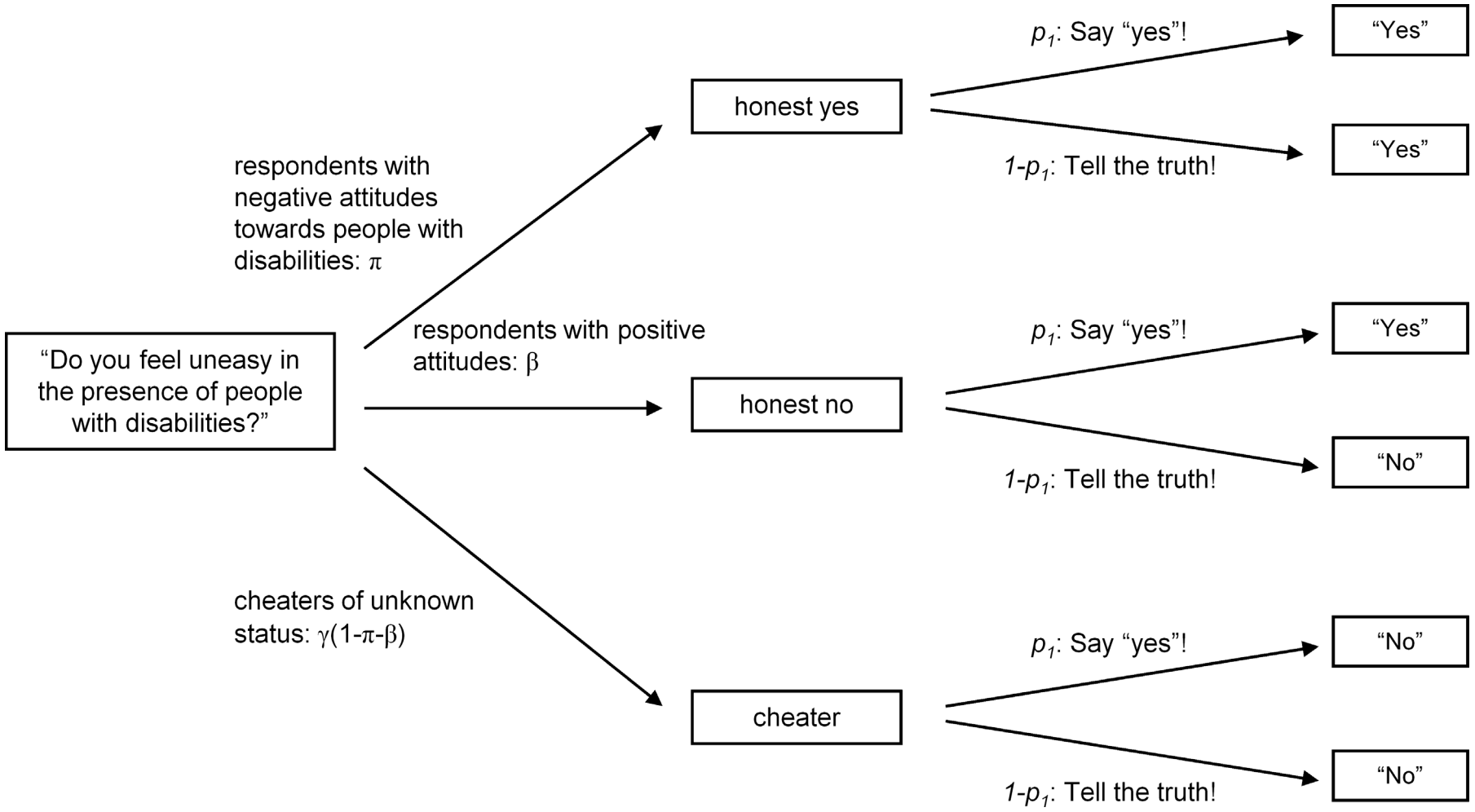
[3]High values indicate that the fit of the model worsens when assuming that the estimated proportion of participants holding a negative attitude towards the disabled in the MPPQ condition differs between respondents scoring low vs. high in SDE or IM (% MPPQ yes_{low} = % MPPQ yes_{high}).

[4]High values indicate that the fit of the model worsens when assuming that the estimated proportion of participants holding a negative attitude towards the disabled in the DQ condition differs between respondents scoring low vs. high in SDE or IM (% DQ yes_{low}) = % DQ yes_{high}).

$^{*}p<.05$, $^{**}p<.01$

Figures

Figure 1. A multinomial formulation of the cheating detection extension of the randomised-response-technique



Assessing sensitive attributes using the randomized-response-technique:

Evidence for the importance of response symmetry

Martin Ostapczuk

Morten Moshagen

Zengmei Zhao

Jochen Musch

Heinrich-Heine-Universitaet Duesseldorf, Germany

Summary

Randomized-response-techniques (RRT) aim to reduce the pervasive response bias accompanying the assessment of socially undesirable attributes. RRT designs differ, however, regarding the degree of privacy protection they offer to the respondent. The less protection a design offers, the more likely respondents are going to cheat by disobeying the instructions and denying the sensitive attribute regardless of the outcome of the randomization device. In asymmetric RRT designs, respondents have the option to play safe by giving the one response that is never associated with the sensitive attribute. Symmetric RRT designs avoid such an incentive to cheat by not allowing responses that dissociate respondents from the socially undesirable attribute. In the present study, we extended Clark and Desharnais' (1998) asymmetric cheating detection model to test whether a symmetric variant of the model increases compliance with the randomization instructions. In a survey of academic dishonesty among 2254 Chinese students, we observed more non-compliance in the asymmetric (21.6%) than in the symmetric variant of the RRT (7.1%). We therefore recommend the use of symmetric cheating detection models, which can easily be estimated and tested within a multinomial modeling framework.

Keywords: randomized-response-technique, privacy protection, cheating detection, response symmetry, multinomial modeling.

Introduction

Self-report is one of the most frequently used data collection techniques in psychology. However, people do not always tell the truth when being asked to answer sensitive questions (Hyman, 1944). Warner (1965) introduced the randomized-response-technique (RRT) to address this problem. The rationale of the RRT is that interviewees are more honest when the confidentiality of their responses is guaranteed by adding random noise to their responses. In the forced response variant of the RRT (Dawes & Moore, 1980), all interviewees are therefore confronted with the sensitive question. A randomization device is used, however, to determine whether respondents are asked to simply provide a prespecified answer ("yes") with probability $\underline{p}_{yes}$, or whether they are asked to answer the sensitive question honestly (with probability $1-\underline{p}_{yes}$). Because the interviewer is unaware of the outcome of the randomization device, the randomization ensures that no individual interviewee can be identified as holding the sensitive attribute on the basis of his or her answer. This is because a "yes"-answer is now no longer the unambiguous result of truthful answering; it may simply be the outcome of the randomization procedure. Because the probability distribution of the randomization device is known, straightforward probability calculations allow researchers to estimate the proportion of "yes"-answers that have not been prompted by the randomization device. The prevalence of the sensitive attribute may thus be estimated at group level, while simultaneously protecting the confidentiality of individual answers. The technique therefore encourages more honest responding.

Since the RRT was first introduced by Warner (1965), many variants of the technique have been proposed and used to estimate the prevalence of a large variety of sensitive behaviors including, for example, tax evasion, illegal drug use, shop lifting, and rape. As a recent meta-

analysis has shown, RRT surveys generally yield more valid estimates than direct questions (Lensvelt-Mulders, Hox, van der Heijden & Maas, 2005).

Efficiency, Privacy, and Cheating in RRT Designs

A major drawback of RRT models is their low efficiency, that is, their greater sampling variance as compared to traditional surveys. Considerable effort has therefore been put into improving the efficiency of RRT models by optimizing design parameters. However, as noted by Lanke (1976) and Guerriero and Sandri (2007), the original goal of increasing privacy protection has not always been taken into account in these efforts. Unfortunately, attempts to increase statistical efficiency usually conflict with the goal to increase the protection of privacy: Randomized responses always contain some information about the interviewee, and they provide more information with an increasing randomization probability of having to answer the sensitive question truthfully. More efficient RRT designs thus necessarily pose an additional threat to the respondents, minimize their willingness to cooperate, and maximize their incentive to disobey the instructions (Antonak & Livneh, 1995; Bourke, 1984; Ljungqvist, 1993). Accordingly, RRT models have not only been criticized for their relative inefficiency, but also for being susceptible to cheating, i.e., to not answering as directed by the randomization device (Campbell, 1987). Indeed, there is evidence that such cheating occurs (Lensvelt-Mulders & Boeije, 2007; Locander, Sudman & Bradburn, 1976). To the extent that participants disobey the instructions by denying the critical behavior in spite of being asked by the randomization device to attest to it, the prevalence of the critical behavior is underestimated.

To address this problem, Clark and Desharnais (1998) proposed to no longer presume that respondents are always obeying the rules of the RRT, and developed a cheating detection extension of the forced response variant of the RRT. Their model takes into account that some

participants might cheat by denying the sensitive attribute despite being directed by the randomization device to attest to it. Nothing can be and is assumed in the model regarding the true status of such cheating respondents. It is conceivable that they deny a behavior in which they have actually been engaged. However, it is also possible that innocent respondents opt to disobey the instructions to rule out even the slightest suspicion of being associated with an undesirable behavior. Consequently, the true status of a cheater can never be ultimately determined.

Using academic dishonesty as an example for the sensitive attribute, Figure 1a illustrates how Clark and Desharnais' (1998) cheating detection extension to the RRT can be graphically depicted as a multinomial model aimed at dividing the population into three disjoint and exhaustive groups: π (the proportion of compliant and honest "yes"-respondents, i.e., participants who truthfully admit to their dishonest exam taking behavior), β (the proportion of compliant and honest "no"-respondents, i.e., honest examinees) and γ (=1-π-β, the proportion of non-compliant cheaters who disobey the rules of the RRT by denying the sensitive attribute regardless of the outcome of the randomization device).

--- Insert Figure 1 about here ---

There are two independent parameters in this model, since the proportions π, β, and γ are constrained to add up to 1. The parameters can therefore no longer be estimated on the basis of the one proportion of "yes"-responses provided in traditional RRT models. Instead, in order to obtain a sufficient data base, it is necessary to pursue an experimental approach. Specifically, two independent samples of respondents have to be questioned with different probabilities $\underline{p}_1$ and $\underline{p}_2$ of being forced by the randomization device to reply "yes" to the critical question (Clark & Desharnais, 1998). Figure 1a depicts only one of these groups, in

which probability p_1 applies; the second group might be represented by an identical figure with the only exception that probability p_1 would have to be replaced by probability p_2. Assuming that the same proportions π, β, and γ apply in both groups when participants are randomly assigned to conditions, the cheating detection model (CDM) now allows us to observe two independent proportions of "yes"-responses. These two proportions suffice to estimate the two independent parameters π and β (with $\gamma=1-\pi-\beta$). For this particular CDM, Clark and Desharnais (1998) derived closed-form solutions for maximum likelihood estimates of the parameters π, β, and γ, as well as a statistical test of the null hypothesis that no cheating occurs. The CDM proposed by Clark and Desharnais (1998) offers a unique theoretical advantage over both traditional surveys and previous RRT models: If no cheating occurs, the parameter π provides an asymptotically unbiased estimate of the population proportion engaged in the sensitive behavior. If, however, there is a significant proportion of cheating respondents, the CDM still allows us to compute both a lower and an upper bound for the sensitive attribute by alternately assuming in a worst and a best case scenario that the cheating respondents either all do or do not carry the critical attribute (Musch, Bröder & Klauer, 2001). As can be shown, the CDM of Clark and Desharnais (1998) can be subsumed under the more general family of multinomial models for which Hu and Batchelder (1994) have developed statistical procedures (Ostapczuk, Musch & Moshagen, submitted). Using a multinomial modeling framework, it is easily possible to estimate the parameters of a large variety of RRT designs, including variants of the original model, and also to test parameter restrictions, such as the assumption that no cheating occurs ($\gamma=0$).

Reasons for Cheating, and Possible Solutions

Two constructs have been proposed to describe potential response hazards, i.e., characteristics of RRT designs that can make both guilty and innocent respondents cheat: respondent

jeopardy, and risk of suspicion (Antonak & Livneh, 1995). Respondent jeopardy refers to the risk of guilty respondents to be identified as such when truthfully admitting the sensitive attribute. Respondent jeopardy may be reduced by choosing randomization probabilities close to .50, which however reduces efficiency by enlarging the variance of parameter estimates (Antonak & Livneh, 1995). Innocent interviewees run a risk of suspicion when being prompted by the randomization device to answer sensitive questions in the affirmative. Innocent respondents tend to feel uncomfortable under such circumstances, because their affirmative answer now seemingly associates them with an undesirable attribute. They may therefore be tempted to play safe by denying the critical attribute in spite of being told otherwise by the randomization procedure (Lensvelt-Mulders & Boeije, 2007).

Bourke (1984) proposed response symmetry as a means of reducing the risk of suspicion. An RRT design is said to be symmetric, if none of the possible responses ("yes" or "no") unequivocally conveys information on the respondent's true status. This can be achieved by forcing some respondents to deny the critical attribute regardless of whether they actually hold it. Because guilty participants are now also sometimes forced to deny, observable responses are no longer linked in any straightforward way to the respondent's true status, and there is no longer an incentive to play safe by denying the critical attribute. In such a symmetric design, interviewees not holding the sensitive attribute should arguably feel less uneasy when being forced to say "yes", and be more likely to follow the RRT rules than in an asymmetric design.

Applying the above definition of response symmetry, forced response variants of the RRT such as Dawes and Moore's (1980) model are asymmetric, because a "no"-response unequivocally identifies an interviewee as not holding the sensitive characteristic. Asymmetric models, however, encourage respondents to cheat by saying "no" despite being asked by the randomization device to answer in the affirmative. Morton (as described in

Greenberg, Abul-Ela, Simmons & Horvitz, 1969) therefore developed a symmetric variant of the forced response model. In his design, depending on the outcome of the randomization process, respondents are either asked to provide the prespecified answers "yes" (with probability p_{yes}) or "no" (with probability p_{no}), or to answer the sensitive question honestly (with probability $1-p_{yes}-p_{no}$). The Morton model is symmetric because both "yes" and "no"-responses may stem from both guilty and innocent respondents. Thus, there is no possibility of playing safe by answering "no", and consequently no incentive to disregard the instructions.

In spite of having been widely discussed in RRT research, it has never been tested whether the increased privacy protection offered by symmetric RRT designs does in fact help to reduce cheating as compared to asymmetric designs. To experimentally investigate the effect of response symmetry on cheating, we therefore compared a symmetric and an asymmetric variant of the CDM of Clark and Desharnais (1998). While Figure 1a depicts the asymmetric model originally proposed by Clark and Desharnais (1998), Figure 1b illustrates the symmetric variant of the CDM which we developed for the purpose of the present study. Note that Figure 1b again depicts only one of the two independent groups required by the model to estimate the number of non-compliant respondents. In this group, probabilities p_3 and p_4 of being forced to say "yes" or say "no", respectively, are being applied. Again, the second group could be represented by an identical Figure in which probability p_3 would be replaced by a different probability p_5, and probability p_4 by a different probability p_6, respectively.

We chose academic dishonesty as the sensitive topic of our investigation. To make our results comparable with traditional self-reports, we included a direct questioning (DQ) baseline condition, and incorporated it into the multinomial model. The full multinomial model for our investigation thus considered five different groups: a DQ control group, an asymmetric RRT group with a low randomization probability p_1 (RRT1), an asymmetric RRT group with a high

randomization probability p_2 (RRT2), a symmetric RRT group with low randomization probabilities p_3 and p_4 (RRT3), and a symmetric RRT group with high randomization probabilities p_5 and p_6 (RRT4).

We formulated the following hypotheses regarding the parameters of the CDM. First, owing to its increased privacy protection, we expected the symmetric design to reduce cheating as compared to the asymmetric design, $\gamma_{symmetric}<\gamma_{asymmetric}$. Second, we expected the response symmetry afforded by forced "no"-answers to decrease the risk of suspicion, which should reduce the incentive to cheat for the innocent interviewees not carrying the sensitive attribute (β). We therefore expected a higher estimate for the proportion of truthfully innocent respondents in the symmetric design ($\beta_{symmetric}>\beta_{asymmetric}$). However, we had no reason to expect an influence of design symmetry on the estimated proportion of respondents carrying the sensitive attribute ($\pi_{symmetric}=\pi_{asymmetric}$).

Materials and Method

Participants and Setting

Data were collected at Beijing Normal University in Beijing, China. Participants were 2254 Chinese first and second year students of various majors, of whom 45.4% were female. Students completed the questionnaires in groups and during lectures on an anonymous and voluntary basis. The participant's month of birth (unknown to the experimenter) was used as a randomization device. Even though the distribution of birthdates is not perfectly equal across months, the distribution is known from official statistics, and the birthdates therefore provide a readily available randomization device that cannot be manipulated by the experimenter. Among some basic demographic questions and questions unrelated to the present study, the questionnaire included the sensitive question concerning academic dishonesty. It read: "Have

you ever been dishonest when taking a school or a university exam?" Participants were randomly assigned to one of the five conditions resulting in 463, 449, 452, 451, and 439 participants in the DQ, RRT1, RRT2, RRT3 and RRT4 conditions, respectively.

In the DQ baseline condition, respondents were simply asked to reply "yes" or "no" to the sensitive question. In the four other conditions, the sensitive question was asked in RRT format. In the asymmetric low probability group (RRT1), instructions read: "If you were born in January or July, then please reply 'yes' to the following question independently of its content. If, however, you were born in another month, then please answer truthfully." The probability of being forced to say "yes" thus approximated $\underline{p_1}$=0.16, as confirmed by birth statistics collected in the 1990 census which were made available by the Chinese National Bureau of Statistics. In the asymmetric high probability group (RRT2), respondents were asked to answer truthfully if they were born in January or July, and to say "yes" if they were born in another month. The probability $\underline{p_2}$ of being forced to say "yes" thus approximated 1.00-$\underline{p_1}$=0.84 in this condition. In the symmetric low probability group (RRT3), instructions read: "If you were born in January, then please reply 'yes' to the following question independently of its content. If you were born in July, then please reply 'no' to the following question independently of its content. If, however, you were born in another month, then please answer truthfully." According to official birth statistics, this resulted in a probability of being forced to say "yes" of $\underline{p_3}$=0.09, and a probability of being forced to say "no" of $\underline{p_4}$=0.07. Finally, in the symmetric high probability group (RRT4), respondents were asked to say "yes" if they were born in February to June (resulting in $\underline{p_5}$ = 0.37), to say "no" if they were born in August to December (resulting in $\underline{p_6}$ = 0.47), and to answer truthfully if they were born in January or July. Detailed instructions explained how this randomization procedure protected the confidentiality of responses. Even though the somewhat unequal distribution of birthdates did not allow us to achieve completely identical probabilities in the two forced response

groups of the symmetric design, the design still fully realized the principled contrast to the asymmetric design which did not offer an opportunity to play safe at all (p_{no}=0).

Ljungqvist (1993) derived a utilitarian measure of the privacy protection offered by different RRT designs. This measure takes into account both the conditional probability of belonging to the sensitive group given a "yes"-answer and the corresponding probability given a "no"-answer.

According to Ljungqvist (1993), privacy protection is maximized if the conditional probability of belonging to the group holding a sensitive attribute (A) given a "yes"-answer, P(A|"yes"), approaches the conditional probability of belonging to group A given a "no"-answer, P(A|"no"). Under such circumstances, the participants' privacy is protected best, and they should therefore be most likely to follow the instructions of the RRT.

More generally, when comparing two RRT designs, the design with the lowest discrepancy between P(A|"yes") and P(A|"no") offers the highest degree of privacy protection. According to Bayes's rule,

$$P(A\,|\,\text{"yes"}) = \frac{\pi_A P(\text{"yes"}\,|\,A)}{\pi_A P(\text{"yes"}\,|\,A) + (1-\pi_A)\,P(\text{"yes"}\,|\,A')} \qquad (1)$$

and

$$P(A\,|\,\text{"no"}) = \frac{\pi_A P(\text{"no"}\,|\,A)}{\pi_A P(\text{"no"}\,|\,A) + (1-\pi_A)\,P(\text{"no"}\,|\,A')} \qquad (2)$$

with P("yes"|A) = 1-P("no"|A), and P("yes"|A')=1-P("no"|A'). While the exact values of these probabilities cannot be determined beforehand because they depend on the population proportion of respondents holding the sensitive attribute, applying these formulas to the forced response variant of the RRT provides theoretical proof of the superiority of symmetric over asymmetric designs; the former offer more privacy to the respondents (Bourke, 1984).

Results

Based on the number of "yes"- and "no"-responses in the different conditions, we computed maximum likelihood estimates for π, β, and γ using the program HMMTree (Stahl & Klauer, 2007) which is implementing the EM-algorithm (Hu & Batchelder, 1994). The fit of the multinomial model was assessed by the asymptotically chi-square distributed log-likelihood ratio statistics G^2 and ΔG^2, respectively.

Table 1 shows the parameter estimates for the different models. When questioned directly, 49.9% of the participants admitted to having been dishonest in an exam at least once; conversely, 50.1% of the participants claimed never to have been dishonest.

--- Insert Table 1 about here ---

In a first step, we compared these results of a direct question with the estimates obtained by a multinomial model for the total sample assuming that π, β, and γ were the same in both the asymmetric and the symmetric RRT. This model estimated the proportion of participants truthfully admitting to academic dishonesty at π=53.4%, and the proportion of honest exam takers at β=26.4%. The estimated proportion of dishonest exam takers was thus descriptively higher than the corresponding estimate in the DQ condition (53.4% vs. 49.9%). The difference, however, was not significant, as restricting the model by assuming the proportions to be equal across groups did not significantly worsen the fit of the model, ΔG^2 (df=1)=1.31, ns. The estimated proportion of honest exam takers β was however significantly lower under RRT (26.4%) as compared to DQ conditions (50.1%), ΔG^2 (df=1)=32.11, p<0.001. This latter finding was a result of the fact that according to the RRT model, a sizable proportion of the sample (γ=20.2%) cheated and disobeyed the rules of the RRT by denying the sensitive

attribute in spite of being told otherwise by the randomization procedure. The statistical significance of the proportion of these non-compliant respondents was indicated by a substantial loss of fit in the model under the assumption of $\gamma=0$ ($\underline{\Delta G^2}$ ($\underline{df}=1$)=116.26, $\underline{p}<0.001$).

As outlined above, the cheating detection variant of the RRT (Clark & Desharnais, 1998) cannot and does not make an assumption regarding the true status of a cheating respondent. However, depending on whether cheating respondents actually did not or did engage in the critical behavior, a lower bound of $\pi=53.4\%$ and an upper bound of $\pi+\gamma=53.4\%+20.2\%=73.6\%$ could be determined for the true proportion of dishonest exam takers. It is important to note, however, that in spite of these plausible parameter estimates, the overall fit of this model to the total sample data was rather poor, $\underline{G^2}$ (df=2)=5.78, $\underline{p}=0.06$, suggesting that response symmetry did have the expected effect and that the assumption of parameter invariance and in particular, the assumption of an equal proportion of cheaters in both symmetric and asymmetric RRTs seems to be unwarranted. We therefore next allowed π, β, and γ to differ between the symmetric and the asymmetric RRT, which resulted in a saturated model with $\underline{df}=0$. As expected, the model showed perfect fit ($\underline{G^2}=0$) because the four proportions of observable "yes"-responses in the four RRT conditions just equal the number of independent parameters ($\pi_{asymmetric}$, $\beta_{asymmetric}$, $\pi_{symmetric}$, and $\beta_{symmetric}$) that have to be estimated for this model. In the following, we are using this saturated model as the basis for estimating the parameters, and for testing the parameter restrictions relevant to our hypotheses.

When allowing for separate estimates in the symmetric and asymmetric condition, the proportion of cheaters in the asymmetric condition was estimated at $\gamma=21.6\%$, as opposed to only $\gamma=7.1\%$ in the symmetric condition. Thus, as expected, the proportion of cheaters varied as a function of response symmetry; assuming an equal proportion of cheaters in both

conditions significantly worsened the fit of the model, ΔG^2 (df=1)=4.12, p<0.05. Further support for an influence of response symmetry on cheating rate was provided by the fact that γ could be set equal to zero in the symmetric RRT without a significant loss in the goodness of fit of the model, ΔG^2 (df=1)=1.09, ns, whereas the same restriction led to a significantly worsened fit of the model in the asymmetric RRT, ΔG^2 (df=1)=105.05, p<0.001.

We next compared the estimate of the proportion of dishonest exam takers in the asymmetric condition (π=51.6%) with the corresponding estimate of the proportion of dishonest exam takers in the symmetric condition (π=53.5%), and the proportion of honest exam takers in the asymmetric condition (β=26.8%) with the corresponding proportion in the symmetric condition (β=39.4%). In accordance with our hypotheses, there was virtually no difference between the π estimates: Assuming the same π parameter in both the symmetric and asymmetric RRT did not significantly deteriorate the fit of the model, ΔG^2 (df=1)=0.21, ns. Regarding the estimates of β, there was the hypothesized descriptive difference, but assuming no difference across the two RRTs failed to worsen the fit of the model in a significant manner, ΔG^2 (df=1)=1.74, ns.

Discussion

The RRT was introduced to reduce response bias when answering sensitive questions. Different RRT designs, however, offer different degrees of privacy protection to the respondents, and designs offering less protection arguably induce more cheating. In the present study, we extended Clark and Desharnais' (1998) asymmetric CDM to test whether a symmetric variant of the model is capable of increasing the number of respondents who are following the instructions, and proceed as prescribed by the randomization procedure.

As hypothesized, we observed lower rates of non-compliance with the instructions (γ) under conditions of improved privacy protection in a survey of academic dishonesty conducted in a Chinese student sample. The prevalence of non-compliance with the instructions was lower in a symmetric (7.1%) as compared to an asymmetric variant of the RRT (21.6%). Statistically, the prevalence of non-compliance with the instructions was indistinguishable from zero in the symmetric condition. The observed reduction in the proportion of non-compliant respondents was accompanied by an increase in the estimated proportion of respondents who were classified as being honest test takers (26.8% vs. 39.4%), whereas the estimated proportion of respondents that were classified as dishonest test takers was about the same in both the asymmetric (51.6%) and the symmetric design (53.5%). This result suggests that cheating in our survey may have been mainly the result of innocent interviewees who opted to ignore the randomization procedure to avoid being associated with the sensitive characteristic, rather than the result of guilty interviewees who attempted to conceal their true status by intentionally disobeying the instructions. The pattern of results fits in well with the expectation that response symmetry mainly reduces the burden placed upon respondents not carrying the sensitive attribute. Under conditions of response symmetry, these innocent respondents have no longer a reason to cheat, because cheating is no longer helpful to avoid being associated with the sensitive attribute. However, the increased privacy protection we obtained using a symmetrical variant of the RRT came at a cost: First, in line with previous research on the conflict between privacy and efficiency in RRT design, the standard errors of parameter estimates were higher and thus, efficiency was lower under response symmetry. Second, because of an additional sentence needed in the instructions, the symmetric RRT was a tiny bit more time-consuming than an asymmetric RRT. Aside from these minor drawbacks, however, the fact that cheating was reduced to an insignificant level by introducing response

symmetry suggests that the symmetrical variant of the CDM is a considerable improvement over existing methods addressing the problem of cheating in randomized-response surveys. Some limitations of the present investigation should be acknowledged, however. First, while providing descriptively higher estimates than the DQ control condition (53.4% vs. 49.9%), statistical power was insufficient to secure significant evidence supporting the notion that the RRT is always providing higher and thus, presumably more valid estimates for the proportion of respondents holding a sensitive attribute. However, the question of whether the RRT is capable of reducing response bias in principle has already been answered convincingly in a large number of surveys (Lensvelt-Mulders et al., 2005), and it is not central to the symmetry hypothesis we investigated in the present study. Moreover, the estimated proportion of truly honest test takers (β according to the model) was significantly lower under RRT (26.4%) as compared to DQ conditions (50.1%). Second, regardless of whether response symmetry is being employed or not, the use of the present CDM is restricted to the assessment of dichotomous attributes only. Even though an extension of the present approach to allow for an assessment of quantitative attributes (Greenberg, Abernathy & Horvitz, 1969) is not impossible in principle, it is not a straightforward exercise and beyond the scope of the present study. Third, the present RRT models do not allow for the assessment of single individuals; they can only be applied to analyze group means. However, in a sense this is both a problem and a virtue. The method cannot be used to determine the behavior of individual participants, but it is exactly this feature that lends credibility to the confidentiality assurance, and encourages respondents to answer more honestly.

To summarize, based on the present findings we can give the following advice to researchers interested in surveying sensitive topics: First, we recommend the use of cheating detection extensions to traditional RRT designs, because only such extended models based on an experimental manipulation of randomization probabilities allow for the assessment of the

proportion of respondents who are disobeying the instructions (Clark & Desharnais, 1998; Musch et al., 2001). Second, when discouraging cheating behavior to increase validity is of primary importance, and when a sufficient number of respondents is available to compensate for the loss of efficiency, we recommend the use of our symmetric variant of the cheater detection design proposed by Clark and Desharnais (1998). As the present results have shown, establishing response symmetry successfully reduces cheating and thus, allows us to considerably improve the assessment of the prevalence of critical attributes. Finally, we recommend using the multinomial modeling techniques suggested by Riefer and Batchelder (1988) and Batchelder and Riefer (1999) when conducting RRT surveys. Multinomial models of the RRT allow convenient parameter estimates using readily available software (Stahl & Klauer, 2007), provide out-of-the-box procedures to flexibly conduct statistical tests of substantive hypotheses (Riefer & Batchelder, 1988), and can easily be adapted to construct new models which improve upon existing RRT designs.

References

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, 99-145.

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

Bourke, P.D. (1984). Estimation of proportions using symmetric randomized response designs. Psychological Bulletin, 96, 166-172.

Campbell, A.A. (1987). Randomized response technique. Science, 236, 1049.

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

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

Greenberg, B.G., Abernathy, J.R., & Horvitz, D.G. (1969). Application of the randomized response technique in obtaining quantitative data. Proceedings of the Social Statistics Section, 1969, 40-44.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R., & Horvitz, D.G. (1969). The unrelated question randomized response model. Theoretical framework. Journal of the American Statistical Association, 64, 520-539.

Guerriero, M., & Sandri, M.F. (2007). A note on the comparison of some randomized response procedures. Journal of Statistical Planning and Inference, 173, 2184-2190.

Hu, X., & Batchelder, W.H. (1994). The statistical analysis of general processing tree models with the EM algorithm. Psychometrika, 59, 21-47.

Hyman, H. (1944). Do they tell the truth? Public Opinion Quarterly, 8, 557-559.

Lanke, J. (1976). On the degree of protection in randomized interviews. International Statistical Review, 44, 197-203.

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.

Lensvelt-Mulders, G., Hox, J., van der Heijden, P., & Maas, C. (2005). Meta-analysis of randomized-response research. Thirty-five years of validation. Sociological Methods & Research, 33, 319-348.

Ljungqvist, L. (1993). A unified approach to measures of privacy in randomized response models: A utilitarian perspective. Journal of the American Statistical Association, 88, 97-103.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. Journal of the American Statistical Association, 71, 269-175.

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: Pabst.

Ostapczuk, M., Musch, J., & Moshagen, M. (2008). A randomized-response investigation of the education effect in attitudes towards foreigners. Manuscript submitted for publication.

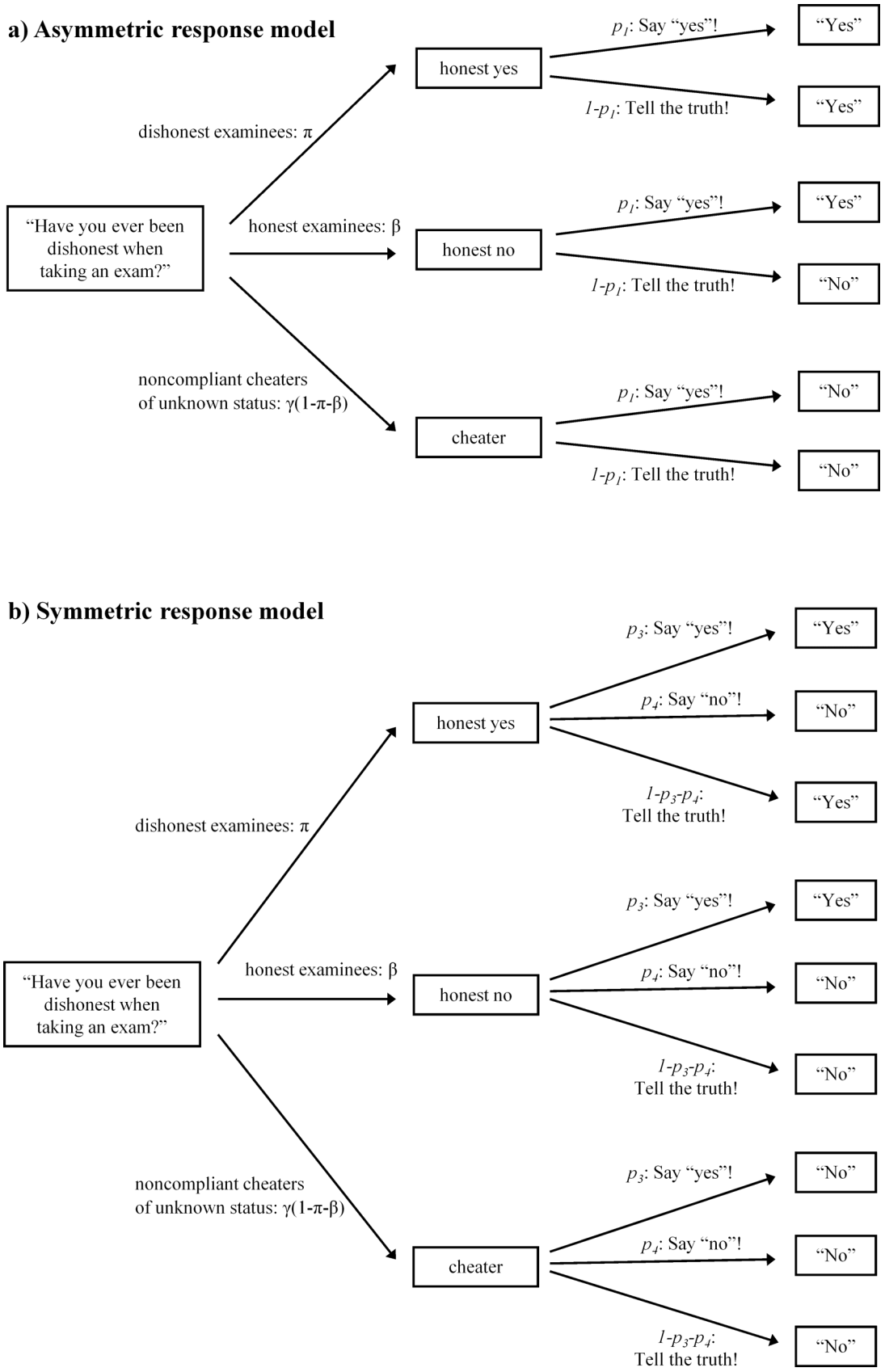
Riefer, D.M., & Batchelder, W.H. (1988). Multinomial modeling and the measurement of cognitive processes. Psychological Review, 95, 318-339.

Stahl, C., & Klauer, K.C. (2007). HMMTree: A computer program for hierarchical multinomial processing tree models. Behavior Research Methods, 39, 267-273.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63-69.

Figures

Figure 1. Multinomial model of the symmetric and asymmetric variant of Clark and Desharnais' (1998) cheating detection extension to the RRT.



Tables

Table 1. Estimated proportions of dishonest examinees, honest examinees and non-compliant cheaters by questioning mode

Total	N=2254		
	Direct Questioning (DQ) N=463		
% yes	49.9%		
(SE)	(3.3)		
% no	50.1%		
(SE)	(3.3)		
	Total sample N=1791	Asymmetric RRT (RRT1 and RRT2) N=901	Symmetric RRT (RRT3 and RRT4) N=890
Honest yes (π)	53.4%	51.6%	53.5%
(SE)	(2.1)	(2.9)	(3.0)
Honest no (β)	26.4%	26.8%	39.4%
(SE)	(3.6)	(4.6)	(8.4)
Cheaters (γ)	20.2%	21.6%	7.1%
(SE)	(2.4)	(2.6)	(6.7)
ΔG^2 (df=1): $\gamma=0$[1]	116.26^{***}		
ΔG^2 (df=1): % yes$=\pi$	1.31		
ΔG^2 (df=1): % no$=\beta$	32.11^{***}		
ΔG^2 (df=1): $\gamma_{asymmetric} / \gamma_{symmetric}=0$		105.05^{***}	1.09
ΔG^2 (df=1): $\gamma_{asymmetric}=\gamma_{symmetric}$		4.12^{*}	
ΔG^2 (df=1): $\pi_{asymmetric}=\pi_{symmetric}$		0.21	
ΔG^2 (df=1): $\beta_{asymmetric}=\beta_{symmetric}$		1.74	

[1]High ΔG^2-values indicate that the fit of the model worsens when the respective restriction is being applied.

$^{*}p<0.05$, $^{***}p<0.001$

Address correspondence to:

Martin Ostapczuk

Heinrich-Heine-Universitaet Duesseldorf

Institute of Experimental Psychology

Universitaetsstr. 1

D-40225 Duesseldorf

Germany

Phone: +49 211 81 10524

Fax: +49 211 81 11753

E-Mail: martin.ostapczuk@uni-duesseldorf.de

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den

(Martin Stefan Ostapczuk)