# Multinomial randomized response models

Inauguraldissertation

zur

Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Morten Moshagen

geboren in Helmstedt

Juni 2008

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

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

Referent: Prof. Dr. Jochen Musch
Korreferent: Prof. Dr. Axel Buchner
Tag der mündlichen Prüfung: 04.07.2008

# Contents

# Zusammenfassung

In vielen sozialwissenschaftlichen Fragestellungen ist der Selbstbericht eine unverzichtbare und oftmals die einzig praktikable Datenquelle. Gleichwohl ist bekannt, dass auf Selbstauskünften beruhende Daten zu sozial erwünschten oder unerwünschten Merkmalen eine fragliche Validität aufweisen. Zwar existieren verschiedene Vorschläge, die Validität von Selbstauskünften bei sensiblen Themen zu erhöhen, etwa die Bogus-Pipeline Technik (Jones & Sigall, 1971), implizite Einstellungsmessung (Greenwald, McGhee, & Schwartz, 1998), psychophysiologische Lügendetektion (Iacono, 2000) sowie der Einsatz von sozialen Erwünschtheitsskalen (z.B. Paulhus, 1984). Jedoch sind diese Verfahren nur mit großem Aufwand zu betreiben, mit ethischen oder rechtlichen Problemen verbunden, oder erlauben nur eine bedingte Kontrolle sozial erwünschter Antworttendenzen. Sie sind deshalb kaum für die Schätzung von Prävalenzen sensibler Merkmale an großen Stichproben geeignet.

Ein vielversprechender und vergleichsweise einfach zu realisierender Ansatz besteht hingegen in der Herstellung von Anonymität. Um die Anonymität über das in direkten Befragungen realisierbare Maß hinaus zu gewährleisten, wurde von Warner (1965) die Randomized Response Technik (RRT) eingeführt. Die zugrundeliegende Idee besteht darin, unter Zuhilfenahme eines Zufallsgenerators die Antworten der Teilnehmer dahingehend zu verschlüsseln, dass die gegebene Antwort nicht mehr direkt mit dem wahren Merkmalsstatus korrespondiert. In der forced-response Variante der RRT entscheidet ein Zufallsgenerator, ob der Befragte gebeten wird die Frage wahrheitsgemäß zu beantworten, oder ob er aufgefordert wird das Vorhandensein des sensiblen Merkmals inhaltsunabhängig zu bejahen. Da eine “Ja”-Antwort somit nicht mehr eindeutig mit dem sensiblen Merkmal assoziiert ist, fördert dieses Verfahren die Bereitschaft der Befragten, das Vorhandensein sensibler Merkmale einzugestehen. Bei bekannter Verteilung des

Zufallsgenerators ist auf Gruppenebene die Schätzung der Prävalenz bei gleichzeitiger Wahrung der Vertraulichkeit der Befragung möglich.

Obgleich in einer Reihe von Validierungsstudien die Überlegenheit der RRT gegenüber traditionellen direkten Befragungsarten wie Fragebögen und Interviews belegt werden konnte (für eine Metaanalyse, siehe Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005), wird die RRT eher selten in der substanzwissenschaftlichen Forschung angewendet. Dafür scheint es mehrere Gründe zu geben. Zum einen besteht in traditionellen Randomized-Response-Befragungen das Problem, dass sich ein unbekannter Anteil der Teilnehmer eventuell nicht an die Instruktionen hält (z.B. Campbell, 1987), mit der Folge, dass die RRT die wahre Prävalenz von sozial erwünschten Merkmalen unterschätzt. Zum anderen entstehen praktische Probleme, wenn es die Forschungsfragestellung erfordert, in einer Studie mehrere sensible Merkmale gleichzeitig zu erfassen. Darüber hinaus gibt es derzeit keine frei verfügbare und einfach zu bedienende Software, welche die Analyse gängiger Randomized-Response-Modelle ermöglicht.

Die vorliegende Dissertation ist diesen Problemen gewidmet; sie hat das Ziel, die Randomized-Response-Technik einem breiteren Anwenderkreis zugänglich zu machen. In Studie I wird dazu ein multinomiales Modell einer Verweigererdetektionsvariante der RRT (Clark & Desharnais, 1988) zur Bestimmung der Prävalenz von ungenügender Zahnhygiene bei Studenten und Studentinnen einer Pekinger Universität eingesetzt. Es kann gezeigt werden, dass mit Hilfe der Verweigererdetektionsvariante signifikant höhere und mutmaßlich validere Prävalenzschätzungen als in einer direkten Befragung erzielt werden können. In Studie II wird mit Hilfe von Computersimulationen die Robustheit des Verweigererdetektionsmodells gegenüber Verletzungen der ihm zugrundeliegenden Modellannahmen überprüft. Die Ergebnisse zeigen, dass Verletzungen der Modellannahmen substantielle Verzerrungen in den Parameterschätzungen zufolge haben können.

Daher wird eine empirisch testbare Erweiterung der Verweigererdetektionsvariante vorgeschlagen, deren Power, Verletzungen der Modellannahmen aufzudecken, ebenfalls bestimmt wird. Studie III behandelt das praktische Problem, mehrere sensible Merkmale in einer einzigen Studie zu erfassen. Damit die Anonymität der Befragten aufrechterhalten wird, scheint die Erfassung mehrerer sensibler Merkmale multiple Randomisierungsprozesse zu erfordern. In Studie III wird jedoch gezeigt, dass ein geeignet gewähltes Befragungsschema die Durchführung wiederholter Zufallsziehungen bei der Erfassung der Antworten auf mehrere sensible Fragen zu vermeiden erlaubt und dabei gleichwohl die Vertraulichkeit der Angaben der Befragten vollständig zu wahren ermöglicht. Studie IV behandelt das Problem, dass in der Verweigererdetektionsvariante der RRT keine Aussage über den wahren Merkmalsstatus von Befragten, die sich nicht an die Instruktion halten, getroffen werden kann. Vorgeschlagen wird eine Erweiterung von Mangats (1994) Variante der RRT, welche es ermöglicht, den Anteil der unehrlich antwortenden Merkmalsträger zu bestimmen. In Studie V wird gezeigt, dass die meisten Randomized Response Modelle als Spezialfall der allgemeineren Klasse der multinomialen Verarbeitungsbaummodelle aufgefasst werden können. Basierend auf dieser multinomialen Reformulierung der Modelle wird ein Programm entwickelt, welches die Analyse von dreizehn verschiedenen Randomized-Response-Modellen in einzelnen und mehreren Gruppen ermöglicht und überdies in der Lage ist, a priori und post-hoc Poweranalysen durchzuführen. Zusammengefasst werden im Rahmen der vorliegenden Dissertation eine Reihe von Problemen gelöst, die einer breiteren Anwendung der Randomized-Response-Technik bislang entgegenstanden. Der Umfrageforschung wird somit ein wirksames und verbessertes Instrument zur Kontrolle von Antwortverzerrungen bei Selbstauskünften zur Verfügung gestellt.

## Summary

Although interviews and questionnaires are widely used in behavioral research, the validity of self-reports of sensitive attitudes and behaviors suffers from the tendency of individuals to distort their responses towards their perception of what is socially acceptable. As a consequence, studies self-report measures consistently underestimate the prevalence of undesirable attitudes or behaviors and overestimate the prevalence of desirable attitudes or behaviors. The randomized response technique (RRT) was developed as a means to overcome this problem by adding random noise to the responses such that there is no direct link between the response an individual provides and his or her true status. Owing to the randomization, the RRT guarantees the confidentiality of responses and encourages more honest responding. Although the superiority of the RRT over more traditional data collection techniques has been repeatedly demonstrated, the RRT is rarely used in substantive research on sensitive issues. Reasons for this dearth of RRT applications include the susceptibility of the RRT to respondents that fail to comply with the instructions, practical problems when assessing multiple attributes in a single study, and the lack of a freely available and easy to use software program implementing randomized response models. The present thesis addressed these problems by extending and validating a cheating detection modification (CDM) of the RRT, showing how to assess multiple attributes with just a single randomization process, and developing a software program tailored for the needs of a wider audience wishing to use RRT models in practice. In Study I, a multinomial reformulation of the CDM was utilized to obtain information about dental hygiene habits among male and female Chinese college students. The results showed that the RRT can substantially improve the validity of prevalence estimates of sensitive behaviors as compared to a traditional direct questioning format. In Study II, computer simulations were performed in order to examine the

statistical efficiency, the statistical power, and the robustness to violations of assumptions of the CDM. It was demonstrated that violations of the assumptions underlying the model lead to biased parameter estimates. Given that the CDM is just-identified and will fit the data perfectly irrespective of violations of assumptions, an extension of the CDM that has the capability to detect violations of assumptions was proposed and examined. The simulation studies further called attention to the importance of choosing diverging randomization probabilities to improve statistical efficiency and power. Study III addressed the problem of assessing multiple attributes in a single study. Using the RRT with multiple attributes seems to either require multiple initializations of the randomization device rendering the administration of the RRT tedious and complicated, or effectively cancels the privacy protection feature of the RRT. To overcome this problem, a particular distribution scheme of the outcomes of the randomization device was developed, which simultaneously avoids the need for multiple randomization processes and maintains privacy protection. Study IV was concerned with the shortcoming of the CDM that it is not able to distinguish whether cheating participants have or do not have a critical attribute. To overcome this problem, an extension of Mangat's (1994) variant of the RRT was proposed. This extension is able to estimate the extent of untruthful responding of those participants who unequivocally carry the sensitive attribute. In Study V, thirteen variants of the RRT are reviewed and it is shown how a common multinomial modeling framework can be adopted for these models. Based on the multinomial reformulation of the models, a software program was developed that allows for the analysis of all of the RRT models in single and multiple groups. Additionally, the program includes an option to perform a-priori and post hoc power analyses. Taken together, the present thesis tackled a series of concerns precluding a wider use of models suitable to gain more valid estimates of the prevalence of sensitive attitudes and behaviors.

# 1 Introduction

In behavioral and survey research, it is often desired to estimate the proportion of respondents holding a certain attitude or behaving in a certain way. To this end, researchers usually ask participants directly on the issue under consideration and utilize the observed proportion of a particular response as an estimate of the prevalence of the respective attribute. It is well known, however, that survey responses do not necessarily reflect an individual's true status. The tendency to present oneself in the best possible light systematically biases responses to sensitive, incriminating, or illegal issues towards a respondent's perception of what is socially acceptable (e.g. Lee, 1993; Tourangou & Yan, 2007). As a consequence, self-report measures consistently underestimate the prevalence of socially undesirable attitudes and behaviors (e.g., doping, drug use, academic cheating, software piracy, tax evasion) and overestimate the prevalence of socially desirable attitudes and behaviors (e.g., general health behavior, hygiene practices, physical activity, moral courage, xenophilia). Several methods have been proposed to overcome this bias, including the bogus pipeline procedure (Jones & Sigall, 1971), implicit attitude measurement (Greenwald, McGhee, & Schwartz, 1998), psychophysiological lie detection (Iacono, 2000), and the use of scales measuring individual differences in the tendency to provide socially desirable responses (e.g., Paulhus, 1984).

Providing confidentiality and anonymity is a simpler but probably the most promising way to encourage truthful and honest responding (e.g., Ong & Weiss, 2000). It has repeatedly been shown that anonymous questionnaires enhance the validity of responses compared to more public modes of administration such as face-to-face interviews. Generally, however, this approach has yielded limited success. Participants may fear that their answer might become known to the researchers conducting the study, and may still decide to mask their true status on the respective attribute by providing supposedly so-

cially acceptable responses. As an attempt to maximize anonymity and confidentiality of responses, random noise is added to the responses in the randomized response technique (RRT; Warner, 1965). Information is thus requested on a probability basis rather than by direct questioning. The confidentiality of responses is increased by ensuring that an individual's status cannot be determined on grounds of his or her response. Since individuals are more likely to be honest when there is no direct link between their attitude or behavior and their response, it is possible to yield more valid prevalence estimates of sensitive or incriminating issues.

# 2 The Randomized Response Technique (RRT)

In the historically first randomized response model propounded by Warner (1965), respondents are asked to answer either the sensitive question (e.g., "I have used cocaine") with probability $p$ or its negation ("I have never used cocaine") with probability $1 - p$. Because it is not known to the interviewer which of these questions was answered, a "yes" answer may indicate that a cocaine-user answered the sensitive question with probability $p$, or that a non-user answered the negation of the sensitive question with the complementary probability $1 - p$. Hence, the proportion of "yes" responses ($\lambda$) is $\lambda = \pi p + (1 - \pi)p$ . A simple algebraic rearrangement yields a maximum likelihood estimator of the prevalence of the sensitive attribute ($\pi$), that is, in the present example, the lifetime prevalence of cocaine use:

$$\hat{\pi} = \frac{\hat{\lambda} + (p - 1)}{(2p - 1)}$$

with variance

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

The variance estimator includes two terms: the first term is the usual sampling variance of proportions; the second term represents the variance added by the randomization procedure. Because the second term is always greater than zero, the RRT suffers a considerable loss of efficiency compared to direct questioning techniques (e.g., Lensvelt-Mulders, Hox, & van der Heijden, 2005). Accordingly, many efforts of developing variants of the Warner model were directed to reduce the variance added by the randomization procedure.

In one of the most efficient variants of the RRT (the forced-response model; Dawes & Moore, 1980), each participant is confronted with the sensitive question, but a certain proportion of respondents is asked to disregard the question entirely and to provide a pre-specified response. Depending on the outcome of the randomization device, respondents are prompted to reply "yes" with probability $p_y$ independently of the content of the question or to answer truthfully with probability $1 - p_y$. For example, the participant's month of birth, unknown to the experimenter, may be used to determine whether participants are prompted to respond truthfully to a certain sensitive question (e.g., "Have you ever used cocaine?"). Depending on their month of birth, some participants are asked to respond truthfully, whereas others are prompted to answer "yes" irrespectively of whether they have used cocaine. Using official birth statistics as a proxy for the probability distribution of the randomization device, it is possible to estimate the proportion of non-forced "yes"-responses, that is, the prevalence of the sensitive attribute ($\pi$), by

$$\hat{\pi} = \frac{\hat{\lambda} - p_y}{(1-p_y)}$$

The variance is given by

$$var(\hat{\pi}) = \frac{\hat{\lambda}(1-\hat{\lambda})}{n(1-p_y)^2}$$

and can be shown to be always smaller than the variance of the Warner model.

Since the randomization procedure guarantees that a "yes"-response is no longer unequivocally indicative of a socially undesirable attribute and therefore no longer stigmatizing to the participants, the RRT encourages more honest responding and, in turn, provides more valid prevalence estimates of sensitive issues. A variety of variants of the RRT have been suggested and successfully employed to obtain information about attitudes and behaviors as diverse as employee theft (Wimbush & Dalton, 1997), doping in fitness sports (Simon, Striegel, Aust, Dietz, & Ulrich, 2006), medication non-adherence (Ostapczuk, Musch, & Moshagen, 2008b), rape (Soeken & Damrosch, 1986), smuggle (Nordlund, Holme, & Tamsfoss, 1994), social security fraud (Lensvelt-Mulders, van der Heijden, & Laudy, 2006), and xenophobia (Ostapczuk, Musch, & Moshagen, 2008a). Likewise, a recent meta-analysis (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005) confirmed that the RRT generally yields more valid prevalence estimates of sensitive attributes than conventional direct questioning formats. Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) concluded that "currently available research has not demonstrated the superiority of *any* [italics added] data collection method to RRT" (p.343).

Successful applications notwithstanding, the RRT has been criticized as being susceptible to respondents who are not answering as directed by the randomization device (Campbell, 1987). When employing the forced response variant of the RRT, two types of non-compliance with the instructions may occur (Antonak & Livneh, 1995). First, respondents may refuse to answer truthfully when prompted by the randomization de-

vice (respondent jeopardy). Although the superiority of the RRT over the traditional direct questioning format is owed to the fact that respondents are more likely to admit carrying a sensitive attribute, the RRT may rather reduce than eliminate this type of non-compliance. Second, the randomization procedure introduces another type of non-compliance, namely the denial to comply with the instruction of answering "yes" to a sensitive question regardless of its content (risk of suspicion). Both types of non-compliance, respondent jeopardy and risk of suspicion, lead to a "no"-response, although the randomization device asks respondents to answer in the affirmative. In fact, evidence suggests that cheating occurs (Edgell, Duchan, & Himmelfarb, 1992; Lensvelt-Mulders & Boeije, 2007; Soeken & Macready, 1982). Respondents who are prompted by the randomization device to answer truthfully may answer "no" although they carry the sensitive attribute, because they may, for instance, not fully understand the rationale of the RRT, do not feel sufficiently protected when the probability of being asked to answer truthfully is high, or may not trust the integrity of the randomization process (Landsheer, van der Heijden, & van Gils, 1999; Lensvelt-Mulders & Boeije, 2007; Soeken & McReady, 1982). Furthermore, respondents may answer "no" in spite of being prompted by the randomization device to answer in the affirmative irrespectively of the content of the item to avoid even the slightest suspicion that they are carriers of the sensitive attribute (Clark & Desharnais, 1998), or because they may feel uncomfortable when being "forced to be dishonest" (Lensvelt-Mulders & Boeije, 2007, p. 602). Whatever causes non-compliance with the RRT instructions, the RRT underestimates the prevalence of the critical behavior to the extent that participants fail to comply with the instructions and deny adopting the sensitive attribute even though they are asked to attest to it.

## Detecting cheating in the randomized response model

Addressing this issue, Clark and Desharnais (1998) proposed a modification of the forced response model: The cheating detection model (CDM) explicitly assumes that some respondents may not comply with the RRT instructions and answer "no" irrespective of the outcome of the randomization device. Figure 1 illustrates how the CDM can be represented as a special case of the more general family of multinomial processing tree models (Batchelder & Riefer, 1999; Hu & Batchfelder, 1994). The population is divided into three disjoint and exhaustive groups: The first group ($\pi$) represents the proportion of compliant and honest "yes"-respondents, that is, respondents who honestly admit having the sensitive attribute (honest cocaine users). The second group ($\beta$) is the proportion of compliant and honest "no"-respondents, that is, respondents who truthfully deny having the sensitive attribute (honest non-users). The third group ($\gamma = 1 - \pi - \beta$) represents the proportion of non-compliant cheaters who do not comply with the instruction of the RRT and answer "no" to the sensitive question irrespective of the outcome of the randomization process. It is important to note that nothing is assumed about whether non-compliant respondents actually have the sensitive attribute. Conceivably, respondents who are prompted by the randomization device to answer truthfully deny a critical behavior in which they have in fact been engaged; but it is also possible that respondents who have not been engaged in the critical behavior want to avoid even the slightest suspicion that anyone might think that they committed a prohibited or undesirable act, and therefore answer "no" despite being prompted by the randomization device to answer affirmatively. Thus, the true status of a respondent choosing not to follow the instructions remains unknown.

As the proportions $\pi$, $\beta$ and $\gamma$ are constrained to add up to 1, the CDM contains two independent parameters that cannot be estimated on the basis of only one proportion

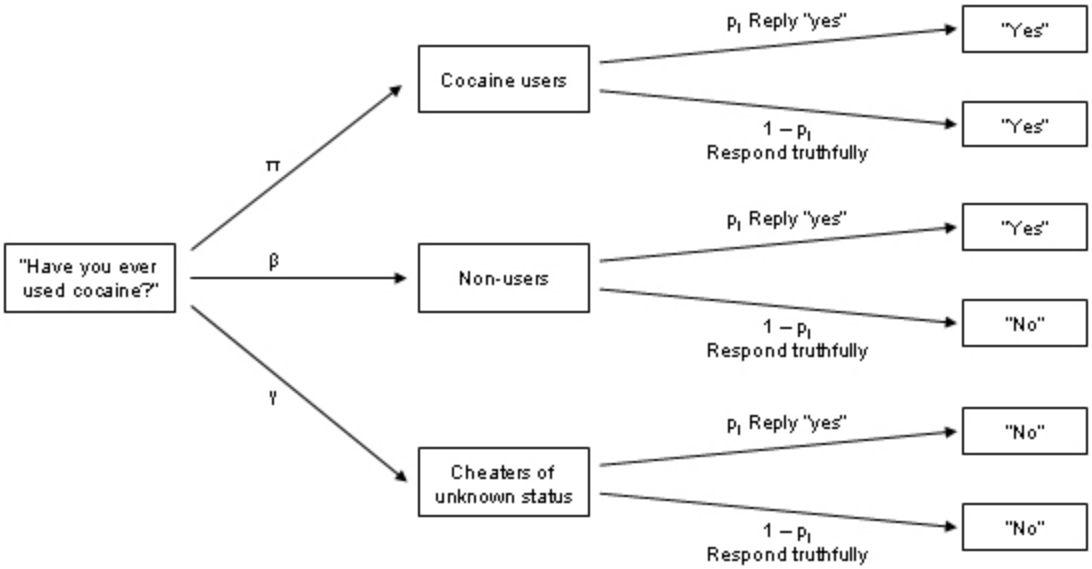



*Figure 1.* A multinomial model of Clark and Desharnais' (1998) cheating detection modification of the randomized response technique. Two independent samples with different randomization probabilities $p_1$ and $p_2$ are needed to make the model identifiable.

of "yes"-responses provided by traditional RRT procedures. An experimental approach is needed to obtain a sufficient data base. More specifically, two independent samples of respondents have to be drawn with different probabilities $p_1$ and $p_2$ of being prompted to reply "yes" by the randomization device (Clark & Desharnais, 1998). Figure 1 shows only one of these conditions, in which probability $p_1$ applies; the second condition could be represented by an identical figure with the sole exception that probability $p_1$ would be replaced with probability $p_2$. Under the assumption that the same proportions apply in both groups when participants are randomly assigned to conditions ($\pi_1 = \pi_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$), the CDM allows to observe two independent proportions of "yes"-responses ($\lambda_1$ and $\lambda_2$), which are sufficient to estimate the two independent parameters $\pi$ and $\beta$

(with $\gamma = 1 - \pi - \beta$). For this particular model, Clark and Desharnais (1998) provide closed-form solutions for unbiased maximum likelihood estimates of the parameters $\pi$, $\beta$, and $\gamma$:

$$\hat{\pi} = \frac{p_{y2}\hat{\lambda}_1 - p_{y1}\hat{\lambda}_2}{(p_{y2} - p_{y1})}$$

$$\hat{\beta} = \frac{\hat{\lambda}_2 - \hat{\lambda}_1}{(p_{y2} - p_{y1})}$$

and $\hat{\gamma}$ can be easily computed by $\hat{\gamma} = 1 - \hat{\pi} - \hat{\beta}$. The asymptotic variance of $\pi$ and $\beta$ are given by

$$var(\hat{\pi}) = \frac{1}{(p_{y2} - p_{y1})^2}\left[\frac{p_{y1}^2 n_{y2} n_{n2}}{{N_2}^3} + \frac{p_{y2}^2 n_{y1} n_{n1}}{{N_1}^3}\right]$$

and

$$var(\hat{\beta}) = \frac{1}{(p_{y2} - p_{y1})^2}\left[\frac{n_{y2} n_{n2}}{{N_2}^3} + \frac{n_{y1} n_{n1}}{{N_1}^3}\right]$$

where $n_{yi}$ and $n_{ni}$ represent the observed frequency of "yes" and "no" responses in the $i$-th sample, respectively.

The CDM offers a unique advantage over both traditional surveys and previous RRT models: If no cheating occurs ($\gamma = 0$), the parameter $\pi$ provides an asymptotically unbiased estimate of the population proportion having the sensitive attribute. If there is a significant proportion of non-compliant respondents, it is possible to compute both an upper and a lower bound for the prevalence of the sensitive attribute by assuming that non-compliant respondents either have or do not have the sensitive attribute (Musch, Bröder, & Klauer, 2001). The CDM may also be considered as a generalization of

the forced response variant of the RRT in that the proportion of cheaters is explicitly modeled, but may also become zero, in which case the CDM is identical to the forced response model extended to two groups.

# 3 Research Questions and Aims of the Studies

Given that research repeatedly demonstrated the superiority of randomized response models over direct questioning formats, it is desirable that the RRT be routinely used in research on sensitive issues. However, as noted by Umesh and Peterson (1991), there is a mismatch between the theoretical development of the RRT and studies using this technique for substantive research questions (see also, Antonak & Livneh, 1995). Possible reasons of the dearth of applications include the susceptibility of the RRT to respondents that fail to comply with the instructions, practical problems when multiple attributes are to be assessed in a single study, and the lack of a freely available and easy to use software program that is implementing randomized response models. The purpose of the present thesis was to tackle these problems in order to increase the use of randomized response models in substantive research. More specifically, the following research questions were addressed:

First, randomized response models have traditionally been analyzed using closed-form solutions. Although mathematically appealing, this approach is limited to specific designs and does not allow for the formulation of more complex models involving additional parameters representing, for example, moderator variables or direct questioning control conditions. Therefore, the possibilities offered by the more general family of multinomial processing tree models were explored and demonstrated.

Second, the statistical properties of Clark and Desharnais' (1998) cheating detection model are largely unknown. Using computer simulations, the efficiency, statistical power,

and the robustness to violations of the underlying assumptions of the model were investigated. A major drawback of the CDM is that it is just-identified and will show a perfect fit to the data irrespectively of violations of assumptions. A modification of the CDM was therefore proposed, which allows observing one additional proportion of "yes"-responses and, thereby, may be falsified empirically.

Third, it is often desired to measure multiple sensitive attributes in a single study. In order to maintain the privacy protection of the RRT, multiple randomization processes seem to be required, rendering the administration of the RRT rather tedious and complicated. A model was developed that allows for assessing multiple attributes with just a single randomization process by using an appropriate answering scheme. This multiple-issues-cheating-detection (MICD) model allows for assessing multiple attributes in a single study while simultaneously maintaining the privacy protection feature of the RRT.

Fourth, a limitation of the CDM is that it is not possible to distinguish between cheating participants that possess or do not possess the sensitive attribute. A modification of Mangat's (1994) variant of the RRT was therefore proposed which has the capability to estimate the proportion of participants who are unequivocally carriers of the sensitive attribute, but fail to respond truthfully.

Finally, an easy to use, platform-independent software program was developed that allows for the estimation of randomized response models in single and multiple groups, including support to analyze moderator variables and to conduct power analyses. Such analyses are difficult to conduct without a thorough knowledge of advanced statistical techniques, but they are of considerable practical importance for researchers using the RRT in applied settings.

In pursuing these issues, the present thesis contributes both to the development of

advanced statistical methods and to the provision of recommendations and tools for researchers wishing to use randomized response models in substantive research.

# 4 Studies

## 4.1 Study I: Employing a multinomial representation of the CDM to investigate gender differences in dental hygiene habits among college students in the PR China

The purpose of this study was to explore the utility of a multinomial representation of Clark and Desharnais' (1998) cheating detection modification of the RRT. To this end, gender differences in dental hygiene habits among Chinese college students were examined. Poor dental hygiene is considered a significant risk factor for a variety of dental diseases such as caries (Bader, Shugars, & Bonito, 2001) and periodontitis (Pihlstrom, Michalowicz, & Johnson, 2005). Since it is difficult to objectively assess the frequency of teeth brushing, previous epidemiological studies mainly relied on self-reported teeth brushing behavior. However, since self-reported hygiene practices are likely to be distorted by socially desirable responding (e.g. Little, Hollis, Stevens, Mount, Mullooly, & Johnson, 1997; Tang, Quinonez, Hallett, Lee, & Whitt, 2005), estimates of the prevalence of appropriate teeth brushing behavior may be overly optimistic. Consequently, assessment of dental hygiene practices was considered a fruitful area of application for randomized response models. The research design comprised a direct questioning control condition in order to obtain an estimate on how much response bias can be reduced by using the CDM. Moreover, because there is evidence suggesting gender differences in teeth-brushing behavior (Lim, Schwarz & Lo, 1994; Petersen, Peng & Tai, 1997), the possibility to include additional parameters representing gender groups in suitably

expanded multinomial models was demonstrated. The results show that only 34.9% of males and 10.4% of females admitted to brush their teeth less than twice a day when questioned directly. Using the CDM, however, 50.7% of males and 20.4% of females admitted this undesirable behavior, indicating that the CDM helps to improve the validity of prevalence estimates of sensitive issues.

## 4.2 Study II: Investigating the robustness to violations of assumptions and the statistical power of the CDM and an enhanced CDM

A major drawback of the CDM is that the model is saturated and will always fit the data perfectly irrespective of whether the assumptions underlying the model are violated. To make the model identifiable, two independent samples using different randomization probabilities have to be drawn, while assuming that the same proportions $\pi$, $\beta$, and $\gamma$ apply in both conditions. From a Bayesian perspective, however, there is reason to suspect that the likelihood to disregard the RRT instructions may be a function of the assigned randomization probabilities (Scheers, 1992; Soeken & McReady, 1982). This is because the likelihood that a "yes"-answer is associated with the sensitive attribute increases with decreasing probability of being prompted to reply "yes". Under these circumstances, the assumption of equal proportions of cheaters across conditions would be violated, which, in turn, might lead to biased parameter estimates. This would be especially critical if the violation of the assumption of equal proportions of cheaters across groups resulted in inflated estimates of the prevalence of the sensitive attribute. Unfortunately, it is unknown how the CDM performs in the presence of an unequal proportion of cheaters across groups.

Given these concerns, an enhancement of the cheating detection model (ECDM) was proposed. The basic idea is to extend the CDM to three different groups, each of

which is questioned with a different randomization probability, while maintaining the assumption of cross-group equality of the parameters $\pi$, $\beta$, and $\gamma$. By this extension, the ECDM provides three independent proportions of "yes"-responses to estimate the two parameters $\pi$ and $\beta$ (with $\gamma = 1 - \pi - \beta$). Thus, the ECDM is overidentified and thereby providing a means for detecting violations of assumptions and model misfit in general. This feature is especially important, as applied researchers will typically not be aware whether an assumption is violated, and may therefore run the risk of obtaining severely biased parameter estimates.

A series of computer simulations was performed to investigate the effects of violations of the assumption of an equal proportion of cheaters across conditions on the accuracy of parameter estimates. Additionally, the statistical power to detect violations of assumptions using the ECDM was determined, and the CDM and the ECDM were compared with respect to the standard errors of the parameter estimates, and the power for analyzing parameter restrictions. Moreover, recommendations were given regarding the optimal choice of the randomization probabilities.

The results demonstrate that violations of the assumption of an equal proportion of cheaters across conditions result in biased estimates using both the CDM and the ECDM. Even though the bias in the parameter estimates may be substantial, it was shown that the models act conservatively by underestimating both the prevalence of the critical behavior and the proportion of cheaters. On the other hand, the models overestimate the proportion of compliant and honest "no"-respondents if the assumption underlying the model are violated. The power of the ECDM to detect violations of the underlying assumptions was found to be rather low, unless the violations are severe or the sample size is large. Moreover, the ECDM was found to suffer from a slight loss of efficiency. Nevertheless, the alternative to using the ECDM – namely, to use a saturated cheater

detection model – is equivalent to accepting that violations of the assumptions and biases in the parameter estimates will not be detectable at all, however large they may be. Furthermore, it was found that power and efficiency may be greatly enhanced by choosing randomization probabilities that lie as far apart as possible.

## 4.3 Study III: Assessing multiple attitudes with a single randomization process

In substantive research, it is often desired to assess multiple sensitive attributes in a single study. If one were to use just a single randomization process for multiple questions, a situation arises where it would be possible to infer from an individual's response pattern whether he or she responded truthfully, or responded as directed by the randomization device. In such a situation, the RRT offers no more privacy protection than traditional direct questioning formats (Tamhane, 1981). A possible solution for this problem is to use a randomization device that allows for multiple initializations (as for example, a die) and re-initialize the randomization process for each question in the survey (e.g., Himmelfarb & Lickteig, 1982). However, this approach is not very attractive because of its complexity, particularly when a large number of questions is to be asked.

This study offered an alternative solution for this problem that does not require multiple randomization processes and simultaneously maintains the privacy protection feature of the RRT. The basic idea is to create as many response patterns solely by the randomization procedure as there are possible response patterns. Consider as an example the participant's month of birth as randomization device. Table 1 clarifies how to distribute the twelve outcomes of the randomization device on the three different questions such that each response pattern (except for denying each question) could be the result of the randomization process. Conversely, there is no response pattern indicating that a

particular participant was prompted to answer truthfully.

*Table 1*

Distribution of the outcomes of the participants' month of birth as a randomization device on three different questions

| | Month of birth | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| #1 | x | x | | x | x | | | | | | | |
| #2 | x | x | x | | | x | | | | | | |
| #3 | x | | x | x | | | x | | | | | |

*Note.* Participants born in a month marked with an 'x' are prompted to provide a pre-specified response (e.g., "yes") to the particular item, the remaining participants are asked to answer truthfully. Each possible response pattern (except for denying every item) may thus be the result from either the randomization procedure, or from truthful responding.

The proposed method can be further enhanced to detect cheaters by sampling two groups for each question. In the first group, a randomization probability of $p_1$ has to be used, whereas in the second group, a randomization probability of $p_2 = 1 - p_1$ is used by inverting the set of months that determines the outcome of the randomization process.

This multiple-issues-cheating-detection (MICD) model was tested in an empirical application with three different questions containing a reference to socially desirable attitudes. Additionally, the performance of the MICD model was compared against a direct question condition, as well as a traditional forced-response RRT not considering cheating. The results showed that the MICD model allows for observing higher and presumably more valid prevalence estimates of the population proportion sharing a socially undesirable attitude as compared to a conventional direct questioning format. It was also shown, that the traditional forced response model not considering cheating leads to

misleading results if substantial non-compliance to the RRT instructions occurs.

## 4.4 Study IV: Development of a model capable of detecting untruthful answering

Albeit Clark and Desharnais' (1998) cheating detection model is a vast improvement compared to both conventional data collection techniques and traditional randomized response variants not considering cheating, the model is not capable of distinguishing whether cheating participants possess or do not possess the sensitive attribute. In fact, Clark and Desharnais (1998) warn against using the proportion of cheaters as a correction factor for the prevalence of the critical behavior. Clark and Desharnais (1998) further note that "neither this model nor any other model is capable of indicating the true behavior of cheaters" (p.166). However, the latter observation only holds true as far as the forced response variant underlying Clark and Desharnais' cheating detection model is concerned. In general, determining whether cheating to the RRT instructions occurred requires the response pattern of cheaters to differ from the response pattern of individuals who truthfully deny having the sensitive attribute. In the forced response variant, respondents truthfully denying to have the sensitive attribute reply "no" if asked to respond truthfully, but reply "yes" if asked to provide a pre-specified response with probability $p_y$; whereas non-compliant cheaters always reply "no" regardless of the outcome of the randomization device. Consequently, it is possible to estimate the extent of non-compliance; however, since both innocent and guilty respondents may decide to disregard the instructions, no assumption can be made about the true status of these cheating participants.

This study showed that another variant of the RRT (Mangat, 1994) also fulfils the basic requirement of different response patterns of respondents disobeying the instructions

and respondents truthfully denying to carry the sensitive attribute. In Mangat's (1994) two-step procedure, each respondent actually carrying the sensitive attribute is asked to answer the sensitive question truthfully. Respondents who do not carry the sensitive attribute are required to use the Warner device, that is, they are asked to answer either the sensitive question with probability $p$ or the negation of the sensitive question with probability $1-p$. Consequently, participants who are carrying the sensitive attribute and respond truthfully ($\pi$) reply "yes", participants who are carrying the sensitive attribute, but fail to respond truthfully ($\gamma$) reply "no", and participants who are not carrying the sensitive attribute ($\beta$) reply "yes" or "no" depending on the outcome of the randomization process (Figure 2). Given these different response patterns, it is shown that the three proportions $\pi$, $\beta$, and $\gamma$ may be estimated by extending Mangat's (1994) procedure to two independent samples with different randomization probabilities. The proposed model has a unique theoretical advantage over Clark and Desharnais' (1998) cheating detection model: The parameter $\gamma$ now unequivocally refers to participants who are carriers of the sensitive attribute, but fail to respond truthfully. This feature obviates the necessity of alternately assuming that all or none of the cheating participants actually carry the sensitive attribute, and directly allows for determining the extent of untruthful answering.

## 4.5 Study V: Development of a multiplatform software program for the analysis and rational design of randomized response models

Randomized response models have traditionally been analyzed using paper and pencil and applying closed form solutions to obtain parameter estimates. This approach suffers from several shortcomings, with the main one being a lack of flexibility. In this study, it was shown that most randomized response models can be subsumed under the

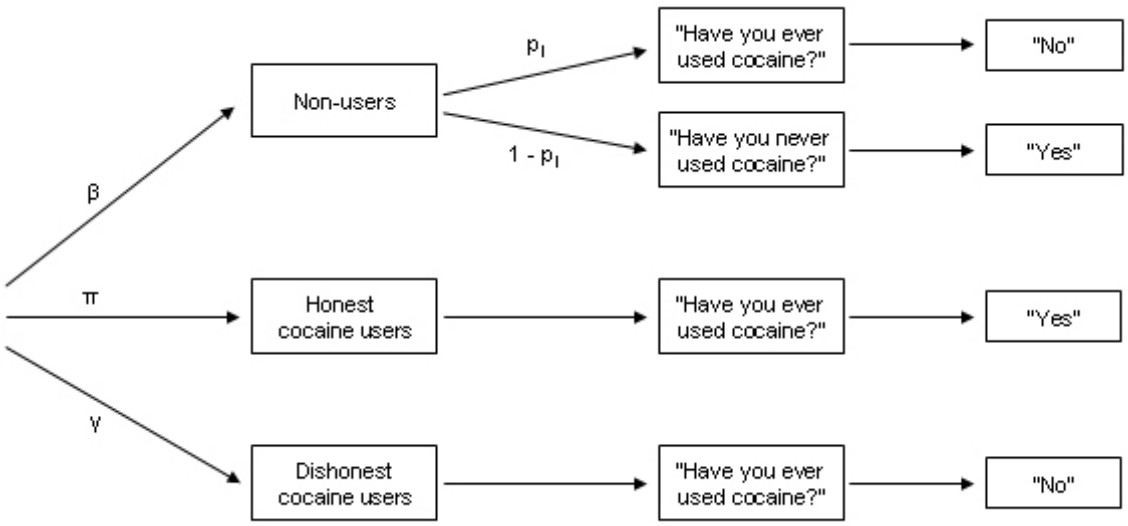



*Figure 2.* A multinomial representation of the proposed modification of Mangat's (1994) variant of the RRT. The population is divided into three disjoint and exhaustive groups: respondents who carry the sensitive attribute and respond truthfully ($\pi$), respondents who carry the sensitive attribute and fail to respond truthfully ($\gamma$), and respondents who do not carry the sensitive attribute ($\beta$). Two independent samples with different randomization probabilities $p_1$ and $p_2$ are needed to make the model identifiable.

more general multinomial processing tree framework (Batchelder & Riefer, 1999; Hu & Batchelder, 1995). A multinomial representation offers several benefits: First, it is easily possible to place constraints on certain parameters, for example, to test whether an RRT-based prevalence estimate of a sensitive attribute differs significantly from the estimate obtained in a conventional direct questioning condition. Second, the models can be extended to the simultaneous analysis of multiple groups with or without cross-group equality constraints on the parameters. Third, more complex models involving additional parameters may be formulated, permitting the estimation of models for which currently there are no closed-form solutions available. Finally, it is possible to perform a priori and post-hoc power analyses, which are necessary for the rational design of studies

employing a particular randomized response model.

The second purpose of this study was to develop a java-based software program called RRTM (randomized response tree modeling) that allows for the analysis of thirteen different randomized response models in single and multiple groups based on multinomial processing tree models as a common framework, additionally including support for performing both, a priori and post-hoc power analyses. Given a randomized response model, the respective design parameters, and the observed response frequencies, RRTM computes maximum likelihood estimates of the parameters of the particular model along with their standard errors, confidence intervals, and significance levels. RRT also makes it easy to analyze multiple-group models. Different groups may represent different subgroups (for example, identified by gender) for which the parameters should be estimated separately, or different RRT questions such as hierarchically ordered questions on a quantity of a sensitive attribute (e.g., "Were you ever involved in a theft from your employer of cash worth from 5$-10$ / 10$-50$ / 50$ and more?"). If more than one group were specified, RRTM provides parameter estimates both with and without cross-group equality restrictions on the parameters along with statistics indicating the applicability of these constraints. Moreover, an option is provided to include a direct questioning control condition. RRTM is not limited to the estimation of the parameters of a particular randomized response model, but also includes the possibility to perform power analyses. The statistical power of a test is defined as the complement of the $\beta$-error probability of falsely retaining an incorrect null hypothesis ($H_0$; Cohen, 1988; Faul, Erdfelder, Lang, & Buchner, 2007). Generally, the power of a test is a function of the probability of an $\alpha$-error, the sample size, and the degree of deviation between the null and alternative hypothesis ($H_1$). Power is calculated by evaluating the non-central $\chi^2$-distribution at a given $\alpha$ with the difference of the $G^2$ fit-statistics of the more restricted $H_0$ model and

the less restricted $H_1$ model as an estimate of the non-centrality parameter. Two types of power-analysis are implemented in RRTM. In a-priori power analyses (Cohen, 1988), the required sample size to reject a false $H_0$ is computed given a fixed significance level $\alpha$ and the desired power. In post-hoc power analyses (Cohen, 1988), the power achieved to reject a false $H_0$ is computed for a given significance level $\alpha$ and a fixed sample size.

# 5 General Discussion

Although it has repeatedly been demonstrated that the RRT helps to improve the validity of prevalence estimates of issues threatened by socially desirable responding (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005), there is an apparent mismatch between the theoretical development of the RRT and studies using this technique for substantive research questions (Antonak & Livneh, 1995; Umesh & Peterson, 1991). Three possible reasons for dearth of RRT applications have been addressed in the present thesis: the susceptibility of the RRT to respondents that fail to comply with the instructions, practical problems when surveying multiple attributes in a single study, and the lack of a freely available and easy to use software program implementing randomized response models.

Regarding the first concern, a cheating detection modification (CDM; Clark & Desharnais, 1998) of the forced response variant of the RRT has been illustrated and validated in Studies I and III. It was demonstrated that as compared to a conventional direct questioning format, the CDM yields higher and presumably more valid prevalence estimates of sensitive issues threatened by over- and underreporting. It was also shown in Study III that the traditional forced response variant of the RRT which is not considering cheating provides misleading results if non-compliance with the instructions does occur. Using computer simulations (Study II), the power of the CDM to reject the null

hypothesis that a parameter does not differ significantly from zero was determined, and power curves were provided as a reference guide for researchers wishing to employ the CDM for their substantive research questions. Study II also investigated the robustness of the CDM to violations of the assumption of cross-group equality of the proportion of cheaters. In the presence of a violation of the assumptions, the CDM was found to act conservatively by underestimating both, the proportion of cheaters and the prevalence of the sensitive attribute. However, given that the CDM is just identified and thus will show a perfect fit irrespective of any violations of the underlying assumptions, researchers might run the risk of obtaining severely biased parameter estimates. For this reason, an enhancement of the CDM called ECDM was proposed. The ECDM allows observing one additional independent proportion of "yes"-responses and, thereby, becomes empirically falsifiable. Even though the power of the ECDM to detect small to moderate violations of the assumptions was found to be rather low and the ECDM suffers a slight loss of efficiency compared to the CDM, it is still preferable to the CDM unless efficiency is a major concern. A further limitation of Clark and Desharnais' (1998) cheating detection model was addressed in Study IV by proposing a model that is capable of estimating the proportion of participants who unequivocally are carriers of the sensitive attribute, but fail to respond truthfully. To the best of my knowledge, the proposed procedure is currently the only randomized response model and, more generally, the only survey method so far, that not only provides incentives for respondents to reply more honestly by increasing their privacy, but also permits to determine the extent of untruthful responding. Although the statistical properties of the proposed model have yet to be established and the validity of the model has yet to be demonstrated, the model may be considered as a vast theoretical improvement over both conventional data collection techniques and previous randomized response models.

In Study III, a particular answering scheme of the outcomes of the randomization device was proposed to address the commonly encountered problem of utilizing the RRT to obtain information about more than one sensitive attribute. This answering scheme permits to assess multiple attributes with just a single randomization process, while maintaining the privacy protection feature of the RRT, consequently facilitating the application of the RRT in substantive research.

Moreover, it has been shown in Study V and demonstrated in Studies I, II, and III, that most randomized response models can be subsumed under the more general family of multinomial processing tree models (Batchelder & Riefer, 1999; Hu & Batchelder, 1994). This modeling approach offers several benefits, the main one being increased flexibility compared to the use of closed-form solutions that have to be computed for each model anew. For instance, a particular model may be augmented with additional parameters representing different subgroups. This approach was illustrated in Study I by comparing the prevalence estimates for appropriate dental hygiene habits across gender groups. As it is often desired to include possible moderator and background variables of theoretical interest, it is believed that this increase in flexibility offered by the multinomial modeling framework is a major improvement for researchers using the RRT.

However, given that employing a multinomial modeling framework to estimate the parameters of a randomized response model requires substantial statistical knowledge, it was deemed necessary to provide a simpler means to analyze these models. Consequently, an easy to use, platform-independent software program was developed (Study V). Based on the multinomial modeling framework, the program is capable of estimating a variety of different randomized response models. Additionally, it includes support for moderator variables and power analyses. To my knowledge, this is the first software program

available that provides a means to analyze the most commonly used randomized response models, thereby rendering the necessity of by-hand calculations obsolete. Furthermore, a rational design of randomized response studies, and an a priori computation of the sample size needed to detect a hypothesized effect becomes possible by the help of the power module included in the software program.

The RRT has numerous strengths, but some the limitations of the technique should also be acknowledged. First, owing to the randomization procedure, the RRT may only be applied at group-level and cannot be used to obtain information about the status of single individuals. As a consequence, it is difficult (though not impossible, see Maddala 1983; Scheers & Dayton, 1986; van den Hout, van der Heijden, & Gilchrist, 2007) to compute measures of association between a sensitive attribute and other variables of interest. As demonstrated in the present thesis, a possible workaround is to utilize the multinomial modeling framework to perform moderator analyses; however, this approach comes to its limits if the number of moderator variables is large. Second, because the RRT adds random noise to the responses, it suffers a greater sampling variation and requires more participants compared to a traditional direct questioning format (e.g., Lensvelt-Mulders, Hox, & van der Heijden, 2005). The resulting loss of efficiency is only outweighed by a gain in precision, if the attribute under consideration is of a sufficiently sensitive nature. Third, employing the RRT is more complicated and tedious than more simple answering formats, as it requires a careful selection of the randomization device used, the sensitive question asked, and particular attention in the development of the instructions given. Finally, the randomized response models considered in the present thesis are only applicable to qualitative attributes that can be assessed by dichotomous items. Using qualitative randomized response models to obtain information about a quantitatively ordered attribute requires the researcher to a priori categorize

the quantities of the attribute under consideration and ask a series of randomized response questions (e.g., "Were you ever involved in a theft from your employer of cash worth from 5$-10$ / 10$-50$ / 50$ and more?"; Wimbush & Dalton, 1997). However, there are extensions of the RRT that allow for an assessment of quantitative attributes (e.g., Greenberg, Kuebler, Abernathy & Horvitz, 1969). Albeit it might be possible to extend these models by adding a cheating detection feature in analogy to Clark and Desharnais' (1998) cheating detection modification of the forced response technique or the modification of Mangat's (1994) variant of the RRT proposed in Study IV, this is not a straightforward exercise and goes beyond the scope of the present thesis.

Limitation notwithstanding, the RRT shows considerable promise as a means to reduce the problem of socially desirable responding. Even though a careful decision must be made prior to adopting this methodology, the RRT may often be a helpful way to obtain meaningful results when surveying sensitive attributes. The present thesis aimed at increasing the impact of the RRT on substantive research questions by addressing various concerns that previously precluded a wider use of these models. To summarize, the main points made in the present thesis are the following:

1) Socially desirable responding leads to biased prevalence estimates of sensitive, illegal, or incriminating issues. The randomized response technique (RRT) is a promising technique for improving the validity of prevalence estimates of issues threatened by socially desirable responding.

2) Most randomized response models may be subsumed under the more general family of multinomial processing tree models. A software program permitting to estimate RRT models based on the multinomial modeling framework is presented.

3) Multiple sensitive attributes may be assessed with just a single randomization process

using an appropriately designed answering scheme.

4) A major drawback of the RRT is its susceptibility to non-compliance to the instructions. However, it is possible to estimate the extent of this non-compliance to the instructions by using a cheating detection modification (CDM; Clark & Desharnais, 1998) of the RRT. The CDM was shown to improve prevalence estimates of sensitive issues compared to both direct questioning formats and a traditional forced response variant of the RRT not considering cheating.

5) For the purpose of identification, the CDM requires two conditions with different randomization probabilities, but assumes the proportion of cheaters to be equal across conditions. It was shown that violations of this assumptions lead to biased parameter estimates. It was also shown that CDM acts conservatively by underestimating both the proportion of cheaters and the prevalence of the critical behavior.

6) The CDM is just-identified and therefore provides no means to detect violations of assumptions when they are present. An enhanced CDM is proposed that can be falsified on the basis of empirical data.

7) A limitation of the CDM is that it is not capable of distinguishing whether cheating respondents do or do not carry the sensitive attribute. An alternative cheating detection model based on Mangat's (1994) variant of the RRT is proposed that allows us to estimate the proportion of participants who unequivocally are carriers of the sensitive attribute, but failed to respond truthfully.

# 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*, 97-145.

Bader, J. D., Shugars, D. A., & Bonito, A. J. (2001). A systematic review of selected caries prevention and management methods. *Community Dentistry and Oral Epidemiology, 29*, 399-411. doi: 10.1034/j.1600-0528.2001.290601.x.

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

Campbell, A. (1987). Randomized response technique. *Science, 236*, 1049. doi: 10.1126/science.3576216.

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160-168. doi: 10.1037/1082-989X.3.2.160.

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

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

Edgell, S. E., Duchan, K. L., & Himmelfarb, S. (1992). An empirical test of the unrelated question randomized response technique. *Bulletin of the Psychonomic Society, 30*, 153-156.

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

*Research Methods, 39*, 175-191.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., & Horvitz, D. G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association, 66*, 243-250.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464-1480. doi: 10.1037/0022-3514.74.6.1464.

Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology, 43*, 710-717. doi: 10.1037/0022-3514.43.4.710.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of engineering processing tree models with the EM algorithm. *Psychometrika, 59*, 21-47. doi: 10.1007/BF02294263.

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

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76*, 349-364. doi: 10.1037/h0031617.

Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality & Quantity, 33*, 1-12.

Lee, R. (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. doi: 10.1016/j.chb.2004.11.001.

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. doi: 10.1007/s11135-004-0432-3.

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

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, Series A, 169*, 305-318. doi: 10.1111/j.1467-985X.2006.00404.x.

Lim, L. P., Schwarz, E., & Lo, E. C. M. (1994). Chinese health beliefs and oral health practices among the middle-aged and the elderly in Hong Kong. *Community Dentistry and Oral Epidemiology, 22*, 364-368. doi: 10.1111/j.1600-0528.1994.tb01594.x.

Little, S. J., Hollis, J. F., Stevens, V. J., Mount, K., Mullooly, J. P., & Johnson, B. D. (1997). Effective group behavioral intervention for older periodontal patients. *Journal of Periodontal Research, 32*, 315-325. doi: 10.1111/j.1600-0765.1997.tb00540.x.

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

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

Musch, J., Bröder, A., & Klauer, K. C. (2001). Improving survey research on the world-wide web using the randomized response technique. In U. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 172-192). Lengerich: Pabst.

Nordlund, S., Holme, I., & Tamsfoss, S. (1994). Randomized response estimates for the purchase of smuggled liquor in Norway. *Addiction, 89*, 401-405. doi: 10.1111/j.1360-

0443.1994.tb00913.x.

Ong, A. D., & Weiss, D. J. (2000). The impact of anonymity of responses to sensitive questions. *Journal of Applied Social Psychology, 30*, 1691-1708. doi: 10.1111/j.1559-1816.2000.tb02462.x.

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

Ostapczuk, M., Musch, J., & Moshagen, M. (2008b). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique. *Manuscript sutmitted for publication.*

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609. doi: 10.1037/0022-3514.46.3.598.

Petersen, P. E., Peng, B., & Tai, B. J. (1997). Oral health status and oral health behaviour of middle-aged and elderly people in PR China. *International Dental Journal, 47*, 305-312.

Pihlstrom, B. L., Michalowicz, B. S., & Johnson, N. W. (2005). Periodontal diseases. *Lancet, 366*, 1809-1820. doi: 10.1016/S0140-6736(05)67728-8.

Scheers, N. J., & Dayton, C. M. (1986). RRCOV: Computer program for covariate randomized response models. *American Statistician, 40*, 229.

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

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

Soeken, K. L., & Damrosch, S. P. (1986). Randomized response technique: Appli-

cations to research on rape. *Psychology of Women Quarterly, 10*, 119-125. doi: 10.1111/j.1471-6402.1986.tb00740.x.

Soeken, K. L., & Macready, G. B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin, 92*, 487-489. doi: 10.1037/0033-2909.92.2.487.

Tamhane, A. (1981). Randomized response techniques for multiple sensitive attributes. *Journal of the American Statistical Association, 1976*, 916-923.

Tang, C., Quinonez, R. B., Hallett, K., Lee, J. Y., & Whitt, J. K. (2005). Examining the association between parenting stress and the development of early childhood caries. *Community Dentistry and Oral Epidemiology, 33*, 454-460. doi: 10.1111/j.1600-0528.2005.00249.x.

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

Umesh, U., & Peterson, R. (1991). A critical evaluation of the randomized response method: Applications, validation and research agenda. *Sociological Methods & Research, 20*, 104-138. doi: 10.1177/0049124191020001004.

van den Hout, A., van der Heijden, P., & Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics & Data Analysis, 51*, 6060-6069. doi: 10.1016/j.csda.2006.12.002.

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

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology, 82*, 756-763. doi: 10.1037/0021-9010.82.5.756.

# Appendices

**Study I:**

Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2008). Reducing socially desirable responding in epidemiological surveys using a cheating detection extension of the randomized-response technique. *Manuscript submitted for publication.*

**Study II:**

Moshagen, M., Ostapczuk, M., Musch, J., Mischke, R., Bröder, A., & Erdfelder, E. (2008). Making compliance testable: How to improve cheating detection in the randomized response technique. *Manuscript submitted for publication.*

**Study III:**

Moshagen, M. & Musch, J. (2008). Surveying multiple sensitive attitudes using a cheating detection extension of the randomized response technique. *Manuscript submitted for publication.*

**Study IV:**

Moshagen, M. & Musch, J. (2008). A note on untruthful responding in randomized response surveys. *Manuscript submitted for publication.*

**Study V:**

Moshagen, M. & Musch, J. (2008). Randomized response models: A review and a software program. *Manuscript submitted for publication.*

Running head: Randomized-response-technique

# Reducing socially desirable responding in epidemiological surveys using a cheating detection extension of the randomized-response technique

Morten Moshagen, Jochen Musch, Martin Ostapczuk & Zengmei Zhao

University of Duesseldorf, Germany

Correspondence should be addressed to:

Morten Moshagen

Institute of Experimental Psychology

University of Duesseldorf

Universitaetsstr. 1

40225 Duesseldorf

Germany

Phone: +49 211 81 13494

Fax: +49 211 81 11753

Email: morten.moshagen@uni-duesseldorf.de

Summary

Background. Even though the validity of self-reports of sensitive behaviours is threatened by social desirability bias, interviews and questionnaires are widely used as data in epidemiological surveys.

Methods. In order to reduce the problem of socially desirable responding, the confidentiality of responses may be enhanced by guaranteeing that the true status of a respondent cannot be identified on grounds of his or her response to a sensitive question. In the randomized-response-technique (RRT), a randomization device is therefore used to determine whether respondents are asked to answer truthfully, or whether they are prompted to provide a prespecified response. Based on the known probability distribution of the randomization device, it is possible to estimate the true population value of sensitive attributes. In the present study, the RRT approach was further extended by employing an experimental cheating detection extension to obtain more valid data on the dental hygiene habits of Chinese college students.

Results. Whereas only 34.9% of males and 10.4% of females admitted to brushing their teeth less than twice a day when questioned directly, 50.7% of males and 20.4% of females attested to this hygienically questionable and hence, socially undesirable behaviour in a randomized-response survey.

Conclusions. The results show that the cheating detection extension of the RRT encourages more honest responding and leads to more valid prevalence estimates than direct questioning. Given the considerable discrepancy between the results obtained by direct questioning and by using the RRT, we propose to routinely consider using the RRT in epidemiological self-reports of sensitive behaviours.

Keywords: Randomized-response technique, cheating detection, underreporting bias, social desirability, sensitive topics, dental hygiene

## Introduction/Background

Poor dental hygiene is considered a significant risk factor for a variety of dental diseases such as caries (Bader et al, 2001) and periodontitis (Piehlstrom et al, 2005). A survey aimed at determining the prevalence of insufficient dental hygiene in China reported that 31% of 20-29 year-olds admit to brushing their teeth less than twice a day (Peng, Petersen, Tai, Yuan & Fan, 1997). This alarming figure seems to be moderated by several demographic variables, however; consistent with epidemiological data on the inhabitants of western countries, improved dental hygiene practices were found in China for females (Petersen, Peng & Tai, 1997; Lim, Schwarz & Lo, 1994), the higher educated (Petersen, Peng & Tai, 1997), and the inhabitants of urban areas (Lin et al., 2001; Zhu et al., 2003). Whereas the prevalence estimates obtained in these studies are already high enough to raise concerns regarding the lack of a sufficient dental hygiene in contemporary China, it is important to note that it is difficult to objectively assess the frequency of teeth brushing relying on self-reported teeth brushing behaviour only. This approach was however taken in all studies that have been conducted in China as yet. The validity of the estimates obtained in this manner may be questioned because self-reported hygiene practices are likely to be distorted owing to socially desirable responding (Little et al, 1997; Tang et al., 2005). Epidemiological validation studies comparing self-report data against gold standard measures have repeatedly shown that the repondent's tendency to provide socially desirable answers results in overreporting desirable behaviours such as physical activity (Adams et al, 2005), and in underreporting undesirable behaviours such as drug use (Colon et al, 2001; Johnson, 2005), energy intake (Hebert et al, 1997; Subar et al, 2003), and sexual risk behaviour (Fennema et al, 1995). By analogy, there is reason to suspect that previous prevalence estimates of appropriate dental hygiene habits in China may have been overly optimistic. But what can be done to obtain more valid estimates?

## The Randomized-Response Technique

In an attempt to overcome the problem of social desirability bias, Warner (1965) proposed the randomized-response technique (RRT) to increase the confidentiality of responses to sensitive issues. To encourage more honest responding and, consequently, to yield more valid prevalence estimates than a direct question, the RRT uses a randomization procedure to ensure that an individual's status cannot be identified on grounds of the response he or she is providing (van der Heijden et al., 2000; Lamb & Stem, 1978; Shotland & Yankowski, 1982). The purpose of the present study was to use an extended cheating detection extension of the RRT (Clark & Desharnais, 1998) to obtain more valid data than a direct question provides on the dental hygiene habits of Chinese college students.

The main idea of the RRT is to improve prevalence estimates of sensitive behaviours by enhancing the confidentiality of responses. There are several variants of the RRT (for a taxonomy, see Antonak & Livneh, 1995) which all rest on the assumption that responders are more likely to be honest when they believe that their true status cannot be determined from their response. In the "forced-response" variant of the RRT (Greenberg et al., 1969), a randomization device (e.g., a die) is used to determine whether participants have to answer a sensitive question truthfully (e.g. "Have you ever used cocaine?") or whether they are prompted to provide a prespecified response (e.g., "yes") irrespective of their true status. Because the outcome of the randomization process is solely known to the participant, the investigator never knows whether a "yes"-response resulted from truthful answering, or from the randomization process. However, the proportion of "yes"-responses which have not been prompted by the randomization procedure (cocaine users in the present example) may be estimated because the probability distribution of the randomization device is known. As a randomization device, the participant's month of birth, unknown to the experimenter, may be

used. Depending on their month of birth, some participants are then asked to respond truthfully, while others are prompted to answer "yes" regardless of their true status. The probability distribution of the randomization device can easily be approximated on the basis of official birth statistics, and the proportion of non-forced "yes"-responses may be estimated by straightforward probability calculations. Hence, the RRT allows to estimate the prevalence of sensitive behaviours at group level without exposing the true status of any individual respondent. The confidentiality of responses is guaranteed by this randomization procedure; owing to the forced "yes"-responses of some of the participants, a respondent no longer unequivocally associates himself with an undesirable behaviour by answering in the affirmative to a sensitive question.

Experimental cheating detection extension of the Randomized-Response-Technique

In spite of their many successful applications (see Lensvelt-Mulders et al., 2005; Antonak & Livneh, 1995; and Fox & Tracy, 1986 for reviews), traditional RRT models have been criticized as being susceptible to non-compliant participants, that is, respondents who are not answering as directed by the randomization device (Campbell, 1987). The prevalence of critical behaviours is underestimated to the extent that participants fail to comply with the instructions and deny being carriers of a sensitive attribute even though they are being asked by the randomization device to attest to it. To address this issue, Clark and Desharnais (1998) proposed an inventive extension of the "forced-response" model: In what we will refer to as the cheating detection model, it is assumed that some respondents may not comply with the RRT rules and answer "no" irrespective of the outcome of the randomization device. For example, respondents who are prompted by the randomization device to answer truthfully may answer "no" in spite of having performed the critical behaviour. On the other hand, innocuous respondents may answer "no" in spite of being prompted by the randomization device to answer in the affirmative, just to avoid associating themselves with a sensitive

attribute. It is important to note that nothing is assumed about whether non-compliant respondents actually performed the sensitive behaviour. It is conceivable that respondents who have been prompted to answer truthfully by the randomization device deny a critical behaviour in which they have in fact been engaged; but it is also possible that respondents who have not been engaged in the critical behaviour want to rule out even the slightest suspicion that they committed a prohibited or undesirable act despite in spite of being asked by the randomization device to answer in the affirmative. Thus, the true status of a respondent choosing not to follow the instructions necessarily remains unknown, and their number cannot simply be added as a correction factor to the number of truthful "yes"-responses (Clark & Desharnais, 1998). Adding all of them to the number of truthful "yes"-respondents is however equivalent to computing an upper bound according to a worst-case scenario which assumes that all incompliant respondents have actually engaged in the critical behaviour.

### please insert figure 1 about here ###

Figure 1 illustrates that the cheating detection model can be depicted as a multinomial model dividing the population into three disjoint groups: The first group ($\pi$) consists of compliant and honest "yes"-respondents, that is, respondents honestly admitting the critical behaviour (e.g., honest cocaine users). The second group ($\beta$) consists of honest "no"-respondents, that is, respondents truthfully denying the critical behaviour (honest non-users). The third group ($\gamma = 1-\pi-\beta$) consists of non-compliant cheaters who do not conform to the rules of the RRT and answer "no" to the sensitive question irrespective of the outcome of the randomization process. As explained above, nothing can be said about the true status of these non-compliant respondents. However, the cheating detection model allows for computing both an upper and a lower bound for the prevalence of the sensitive attribute by assuming that these non-

compliant respondents either did, or did not engage in the critical behaviour (Musch et al., 2001).

As the proportions $\pi$, $\beta$ and $\gamma$ are constrained to add up to 1, the cheating detection model comprises two independent parameters. These cannot be estimated on the basis of the only one proportion of "yes"-responses that is provided by traditional RRT procedures and thus, the model is not identified. To make the model identifiable, 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 in the affirmative to the sensitive question (Clark & Desharnais, 1998). 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 be replaced with probability $\underline{p_2}$. Under the assumption that the same proportions $\pi$, $\beta$, and $\gamma$ apply in both groups when participants are randomly assigned to conditions, the cheating detection model allows to observe two independent proportions of "yes"-responses which are sufficient to estimate the two independent parameters $\pi$ and $\beta$ (with $\gamma=1-\pi-\beta$). For this particular model, Clark and Desharnais (1998) provide closed-form solutions for maximum likelihood estimates of the parameters $\pi$, $\beta$, and $\gamma$, as well as a statistical test of the null hypothesis that no cheating occurs. However, their cheating detection model can actually be regarded as a special case of the more general family of multinomial models (Batchelder & Riefer, 1999; Riefer & Batchelder, 1988). Converting the ternary tree model into a statistically equivalent binary tree representation allows using established procedures of parameter estimation for multinomial models (Batchelder & Riefer, 1999; Hu & Batchelder, 1994), and also allows to test the applicability of restrictions on the parameters such as the assumption that no cheating occurs ($\gamma=0$). The latter can be done because the difference of the fit of a restricted and an unrestricted model follows the asymptotically $\chi^2$ distributed log-likelihood ratio statistic $\underline{G^2}$. Importantly, using a multinomial modelling framework, it is also

possible to formulate more complex models incorporating additional parameters. These may represent, for example, subgroups for which parameters have to be estimated separately (e.g., parameter estimates for different sexes or age groups).

The cheating detection model offers a unique theoretical advantage over both, traditional surveys and previous RRT models: If no cheating occurs (that is, if the proportion $\gamma$ of non-compliant respondents can be set equal to zero without a significant loss in the goodness of fit of the model to the data), the parameter $\pi$ provides an asymptotically unbiased estimate of the population proportion engaged in the sensitive behaviour. Morevoer, if there is a significant proportion of non-compliant respondents, it is possible to compute both an upper and a lower bound for the prevalence of the sensitive attribute by assuming that non-compliant respondents either all did, or did not engage in the critical behaviour.

In order to explore the magnitude of a potential bias in self-reported hygiene habits due to socially desirable responding, we performed an RRT study on teeth brushing behaviour among Chinese college students. By breaking down the sample by sex to investigate a possible influence of this variable, we took advantage of the possibility to include additional parameters for different subgroups in our multinomial randomized response model. Additionally, we included a direct questioning control condition to obtain an estimate of how much response bias can be reduced by using the cheating detection extension of the RRT.

## Methods

### Participants

A total of 2254 undergraduates, 1023 of which were female, from various faculties of the Beijing Normal University, Beijing volunteered to participate in this study. Age ranged from 18 to 24 years. Students completed the questionnaire on an anonymous and voluntary basis during regular classes.

Measures and Procedures

Participants completed a questionnaire comprising demographic information, several questions not pertinent to the present study, and the sensitive question concerning their dental hygiene habits: "Do you brush your teeth at least twice a day?" In previous applications of the RRT, the socially undesirable response to a sensitive question typically was to answer in the affirmative (e.g., "Yes, I did use cocaine"). Because in the present investigation, the socially undesirable response was to admit brushing one's teeth less than twice a day by answering "no", we adapted the randomized response procedure to this reverse direction of social desirability by making sure that the randomization procedure required some of the participants to provide a forced "no"-answer, to protect the answer that would otherwise be regarded as stigmatizing.

Participants were randomly assigned to one of three conditions, with the restriction that a higher number of participants was assigned to the randomized-response conditions to compensate for the loss of efficiency in parameter estimation associated with the use of the randomization procedure. In the direct questioning control condition (N=463, 251 female), participants were simply asked to answer truthfully to the sensitive question. In the two RRT conditions ($p_1$: N=900, 501 female; $p_2$: N=891, 478 female), the sensitive question was asked in randomized-response format. In order to keep the randomization procedure simple and transparent, we used the participants' month of birth as a randomization device. The RRT instructions presented to the low probability group ($p_1$) read as follows: "If you were born in January or February, then please answer 'no' to the question independently of its content. If you were born in another month, then please answer truthfully". Instructions for the high probability group ($p_2$) were identical with the exception that the months of births were reversed; participants in this group were asked to respond truthfully when they were born in

January or February, and to answer 'no' regardless of the question content if they were born in another month. According to birth statistics provided by the National Bureau of Statistics of China, the probabilities $p_1$ and $p_2$ of being forced to say "no" in the two RRT conditions thus approximated .17 and .83, respectively. Detailed instructions explained that owing to the randomization, the procedure guaranteed the confidentiality of responses.

### please insert table 1 about here ###

## Results

Maximum likelihood estimates were computed for the multinomial model parameters using the EM-algorithm (Hu & Batchelder, 1994) implemented in the freely available software program HMMTree (Stahl & Klauer, 2007). When questioned directly, 34.9% of males and 10.4% of females admitted to brushing their teeth less than twice a day (i.e., answered "no" to the sensitive question). Because the fit of the model significantly deteriorated when these two proportions were restricted to be equal ($\Delta G^2$ (df=1)=41.78, $P< .01$), we estimated the parameters of the model for both sexes separately (see Table 1 for the parameter estimates for the whole sample). In general, the figures we obtained in the direct questioning condition were consistent with the proportions reported in the prior study of Peng, Petersen, Tai, Yuan and Fan (1997). However, when questioned using the randomized-response model, 50.7% of males and 20.4% of females admitted to brushing their teeth less than twice a day (Table 1). These proportions are significantly higher than those in the direct questioning condition: Assuming that the proportion of "no"-responses to the direct question does not differ from the estimated proportion of honest "no"-responses in the RRT condition (% "no" = $\pi$) significantly worsened the fit of the model both for males ($\Delta G^2$ (df=1)=11.70, $P<.01$) and for females ($\Delta G^2$ (df=1)=9.01, $P< .01$). Similarly, constraining $\pi$ to be equal across sexes also

resulted in a significant loss in model fit ($\Delta G^2$ (df=1)=51.67, P<.01), indicating that males and females are found to differ with respect to dental hygiene habits when questioned in randomized response format. Finally, a noteworthy proportion of noncompliance to instructions was observed for both males ($\gamma$=10.1%) and females ($\gamma$=13.0%). These proportions of non-compliant respondents differed significantly from zero in both groups ($\Delta G^2$ (df=1)=23.12, P<.01 for males, and $\Delta G^2$ (df=1)=33.21, P<.01 for females). Restricting the $\gamma$ parameters to be equal across sexes did not significantly worsen model fit ($\Delta G^2$ (df=1)=0.70, ns), indicating that males and females did not differ with respect to their tendency to be noncompliant to the RRT rules. Depending on whether non-compliant respondents were considered to have or have not been engaged in insufficient teeth brushing, we computed a lower-bound and an upper-bound estimate for the proportion of respondents admitting to brushing their teeth less than twice a day. The lower-bound estimate for this proportion was $\pi$ = 50.7% for males and $\pi$ = 20.4% for females; the respective upper-bound estimate was $\pi + \gamma$ = 60.8% for males and of $\pi + \gamma$ = 33.4% for females, respectively. According to the model, the proportion of males and females that can unequivocally be classified as brushing their teeth at least twice a day was thus estimated at only 39.2% for males and 66.6% for females, respectively.

## Discussion

It has often been argued that survey data reflect what respondents tell investigators, rather than the respondents' actual behaviour. The present paper used a cheating detection extension of the RRT as a means to improve the validity of an epidemiological survey on dental hygiene habits among Chinese college students. Only 34.9% of males and 10.4% of females admitted to brushing their teeth less than twice a day when questioned directly. When the cheating detection extension of the RRT was employed, the respective proportions increased

substantially for males, and almost doubled for females. Assuming that all non-compliant respondents do in fact not brush their teeth at least twice a day, it was possible to compute an upper-bound prevalence of insufficient dental hygiene for males and females of 60.8% and 33.4%, respectively. Taken together, these figures clearly demonstrate the superiority of the cheating detection variant of the RRT over direct questioning formats, which may provide strongly distorted prevalence estimates when the behaviour in question is subject to social desirability bias, and also over traditional variants of the RRT not capable of detecting cheating. The latter underestimate the true population proportion of the carriers of a sensitive attribute to the extent that there are non-compliant respondents. In order to yield more valid prevalence estimates of sensitive issues, we therefore strongly suggest to routinely consider the cheating detection extension of the RRT in future epidemiological surveys on sensitive issues.

Some limitations of the RRT should however be acknowledged. First, due to the randomization procedure, randomized-response models introduce random error and therefore induce greater sampling variance; this can only be compensated by increasing the sample size. Second, RRT surveys are somewhat more time-consuming and slightly more complicated to administer. Third, the cheating detection extension of the forced response variant of the RRT is only applicable to dichotomous items. Although the cheating detection model can readily be applied to any other variant of the RRT, those allowing for the quantitative assessment of attributes (e.g. Greenberg et al., 1971; Pollock & Bek, 1976) are characterized by a considerably lower efficiency (Lensvelt-Mulders et al., 2005b). Finally, the true dental hygiene habits of any single individual necessarily remains unknown, since the RRT can only be applied at group level. However, it is exactly this feature of the RRT that enhances the confidentiality of responses and encourages participants to answer more honestly. Despite these limitations, we argue that the cheating detection extension of the RRT shows considerable promise as a means to improve prevalence estimates of sensitive behaviors.

Given the considerable discrepancy between the results we obtained in the direct questioning and the randomized-response condition, we strongly recommend the use of the cheating detection variant of the RRT in future epidemiological surveys on sensitive behaviours.

Author note

This work was supported by a grant of the German Research Foundation (DFG, Mu 2674/1-1). Correspondence concerning this article should be addressed to Morten Moshagen (E-Mail: morten.moshagen@uni-duesseldorf.de) or Jochen Musch (E-Mail: jochen.musch@uni-duesseldorf.de), Institute for Experimental Psychology, University of Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany.

References

1. Adams SA, Matthews CE, Ebbelin CB et al. The effect of social desirability and social approval on self-reports on physical activity. Am J Epidemiol 2005; **161**: 389-98.

2. Antonak RF, Livneh H. Randomized response technique: A review and proposed extension to disability attitude research. Genet Soc Gen Psych 1995; **121**: 97-145.

3. Bader JD, Shugars DA, Bonito AJ. A systematic review of selected caries prevention and management methods. Community Dent Oral Epidemiol 2001; **29**: 399-411.

4. Batchelder WH, Riefer DM. Theoretical and empirical review of multinomial process tree modeling. Psychon Bull Rev 1999; **6**: 57-86.

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

6. Clark SJ, Desharnais RA. Honest answers to embarrassing questions: Detecting cheating in the randomized response model. Psychol Methods 1998; **3**: 160-8.

7. Colón HM, Robles RR, Sahai H. The validity of drug use responses in a household survey in Puerto Rico: Comparison of survey responses of cocaine and heroin use with hair tests. Int J Epidemiol 2001; **30**: 1042-9.

8. Fennema JSA, van Ameijden EJC, Coutinho RA, van den Hoek JAR. Validity of self-reported sexually transmitted diseases in a cohort of drug-using prostitutes in Amsterdam: Trends from 1986-1992. Int J Epidemiol 1995; **24**: 1034-41.

9. Fox JA, Tracy PE. Randomized response: A method for sensitive surveys. Beverly Hills, CA: Sage, 1986.

10. Greenberg BG, Abul-Ela A-LA, Simmons WR, Horvitz DG. The unrelated question randomized response model. Theoretical framework. J Am Stat Assoc 1969; **64**: 520-39.

11. Hebert JR, Ma Y, Clemow L et al. Gender differences in social desirability and social approval bias in dietary self-report. Am J Epidemiol 1997; **146**: 1046-55.

12. Hu X. Multinomial processing tree models: An implementation. Behav Res Methods 1999; **31**: 689-95.

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

14. Johnson T, Fendrich M. Modeling sources of self-report bias in a survey of drug use epidemiology. Ann Epidemiol 2005; **15**: 381-9.

15. Lamb CW, Stem DE. An empirical validation of the randomized response technique. J Marketing Res 1978; **15**: 616-21.

16. Lee RM. Doing Research on Sensitive Topics. London: Sage, 1993.

17. Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM. How to improve the efficiency of randomised response designs. Qual Quant 2005; **39**: 253-65.

18. Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM, Maas C. Meta-analysis of randomized-response research. Thirty-five years of validation. Sociol Method Res 2005; **33**: 319-48.

19. Lim LP, Schwarz E, Lo ECM. Chinese health beliefs and oral health practices amont the middle-aged and the elderly in Hong Kong. Community Dent Oral Epidemiol 1994; **22**: 364-8.

20. Lin HC, Wong MCM, Wang ZJ, Lo ECM. Oral health knowledge, attitudes, and practices of Chinese Adults. J Dent Res 2001; **80**: 1466-70.

21. Little SJ, Hollis JF, Stevens VJ, Mount K, Mullooly JP, Johnson BD. Effective group behavioral intervention for older periodontal patients. J Periodont Res 1997; **32**: 315-325.

22. Madalla GS. Limited dependent and qualitative variables in econometrics. Cambridge: Cambridge University Press, 1983.

23. Musch J, Bröder A, Klauer KC. Improving survey research on the World-Wide Web using the randomized response technique. In: Reips U-D, Bosnjak M eds). Dimensions of Internet Science. Lengerich: Pabst, 2001.

24. Peng B, Petersen PE, Tai BJ, Yuan BY, Fan MW. Changes in oral health knowledge and behaviour 1987-95 among inhabitants of Wuhan city, PR China. Int Dent J 1997; **47**: 142-7.

25. Petersen PE, Peng B, Tai BJ. Oral health status and oral health behaviour of middle-aged and elderly people in PR China. Int Dent J 1997; **47**: 305-12.

26. Pihlstrom BL, Michalowicz BS, Johnson NW. Periodontal diseases. Lancet 2005; **366**: 1809-20.

27. Pollock KH, Bek Y. A comparison of three randomized response models for quantitative data. J Am Stat Assoc 1976; **71**: 884-6.

28. Riefer DM, Batchelder WH. Multinomial modeling and the measurement of cognitive processes. Psychol Rev 1988; **95**: 318-39.

29. Scheers NJ, Dayton CM. RRCOV: Computer program for covariate randomized response models. Am Stat 1986; **40**: 229.

30. Shotland RL, Yankowski LD. The random response method: A valid and ethical indicator of the "truth" in reactive situations. Pers Soc Psychol Bull 1982; **8**: 174-9.

31. Stahl C, Klauer KC. HMMTree: A computer program for hierarchical multinomial processing tree models. Behav Res Methods 2007; **39**: 267-273.

32. Subar AF, Kipnis V, Troiano RP et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study. Am J Epidemiol 2003; **158**: 1-13.

33. Tang C, Quinonez RB, Hallett K, Lee JY, Whitt JK. Examining the association between parenting stress and the development of early childhood caries. Community Dent Oral Epidemiol 2005; **33**: 454-60.

34. van der Heijden PGM, van Gils G, Bouts J, Hox JJ. A comparison of randomized response, CASI and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. Sociol Method Res 2000; **28**: 505-37.

35. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. J Am Stat Assoc 1965; **60**: 63-.9.

36. Zhu L, Petersen PE, Wang H-Y, Bian J-Y, Zhang B-X. Oral health knowledge, attitudes and behaviour of children and adolescents in China. Int Dent J 2003; **53**: 289-98.

Tables

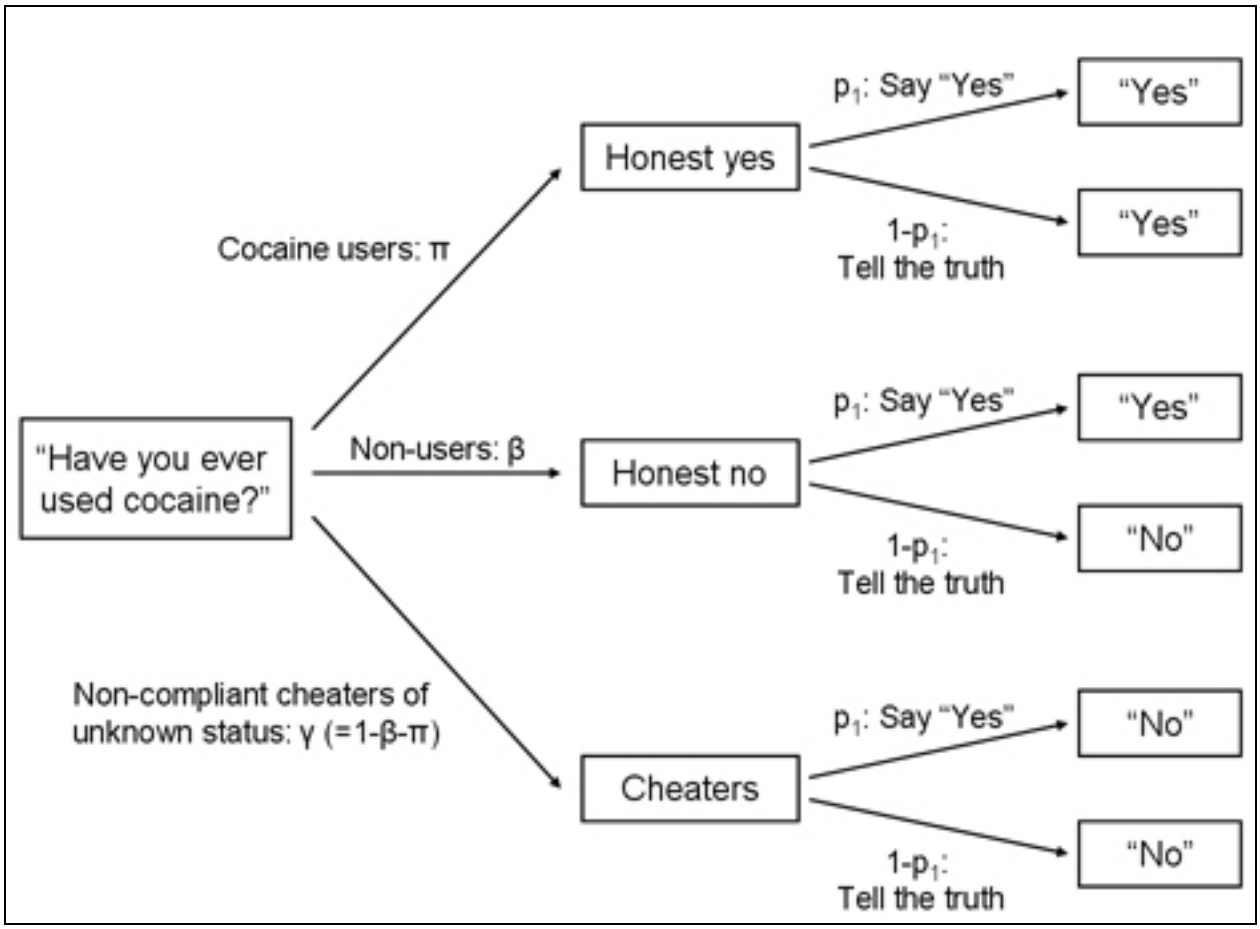
Table 1. Observed and estimated proportions (%) of "yes" and "no"-responses to the question: "Do you brush your teeth at least twice a day?".

| | | Total | Females | Males |
|---|---|---|---|---|
| DQ | yes | 78.40 | 89.64 | 65.09 |
| | [95% CI] | [74.65-82.15] | [85.87-93.41] | [58.67-71.51] |
| | no | 21.60 | 10.36 | 34.91 |
| | [95% CI] | [17.85-25.35] | [6.59-14.13] | [28.49-41.32] |
| RRT | Honest yes (β) | 54.71 | 66.63 | 39.22 |
| | [95% CI] | [48.36-61.06] | [58.14-75.11] | [30.08-48.37] |
| | Honest no (π) | 33.82 | 20.42 | 50.73 |
| | [95% CI] | [29.69-37.94] | [15.20-25.64] | [44.56-56.90] |
| | Non-compliant cheaters (γ) | 11.47 | 12.95 | 10.05 |
| | [95% CI] | [8.02-14.92] | [8.01-17.90] | [5.35-14.74] |

Notes. DQ=direct questioning control condition. RRT=randomized-response conditions. All proportions are significant at P<.01.

Figures

Figure 1: A multinomial representation of the cheating detection extension of the randomized-response-technique



Running Head: RANDOMIZED RESPONSE TECHNIQUE

# Making compliance testable:

# How to improve cheating detection in the randomized response technique

Morten Moshagen, Martin Ostapczuk, and Jochen Musch

University of Duesseldorf, Germany

Robert Mischke and Arndt Bröder

University of Bonn, Germany

Edgar Erdfelder

University of Mannheim, Germany

Abstract

The randomized response technique (RRT) encourages more honest responses to sensitive questions by requesting information on a probability basis, but underestimates the prevalence of the critical behavior to the extent that respondents fail to comply with the instructions. The cheating detection modification (CDM) provides a means to estimate the extent of non-compliance with the RRT instructions by questioning two independent groups with different randomization probabilities. The CDM assumes the proportion of cheaters to be equal across groups; however, there is reason to assume that the extent of cheating may depend on the assigned randomization probabilities. Using computer simulations, we demonstrate that violations of this assumption lead to biased parameter estimates. Because the CDM is a saturated model always perfectly fitting the data, we propose an enhancement of the cheating detection model (ECDM). The ECDM is overidentified and allows detecting violations of assumptions and model misfit in general. We recommended to use the ECDM with diverging randomization probabilities to enhance statistical power and efficiency.

Keywords:

Randomized response technique, cheating, social desirability, computer simulation, power

# Making compliance testable:
# How to improve cheating detection in the randomized response technique

Questionnaires and interviews are commonly used in behavior research to study a variety of attitudes and behaviors. Unfortunately, the tendency to present oneself in the best possible light systematically biases responses on sensitive, incriminating, or illegal issues towards respondents' perceptions of what is socially acceptable. As a consequence of the desire to conform to societal norms and to avoid being embarrassed, self-report measures consistently underestimate the prevalence of undesirable attitudes and behaviors and overestimate the prevalence of desirable attitudes and behaviors.

Providing confidentiality and anonymity to respondents is probably the most promising way to encourage honest and truthful responding. However, this strategy has yielded limited success. Respondents may not always completely trust the investigators or may fear that their answer becomes known at the very least to the researchers conducting the survey. As an attempt to overcome this problem, Warner (1965) developed the randomized response technique (RRT), which requests information on a probability basis rather than by direct questioning. The confidentiality of responses is increased by ensuring that an individual's status cannot be determined by his or her response. Since individuals are more likely to be honest when their true status cannot be identified from their response, it is possible to yield more valid prevalence estimates of sensitive or incriminating issues. Several variants and extensions of the original RRT have been proposed (for a taxonomy, see Antonak & Livneh, 1995), among which the forced response variant (Dawes & Moore, 1980) has been shown to be most efficient under a variety of conditions (Lensvelt-Mulders, Hox, & van der Heijden, 2005).

In the forced response variant of the RRT, each respondent is confronted with a sensitive question (e.g., "Have you ever used cocaine?"). Before answering, a randomization

device (e.g., a die) is used to determine whether participants are asked to respond truthfully or whether they are prompted to provide a prespecified response (e.g., "yes") irrespective of their true status. Because the outcome of the randomization process is solely known to the participants, the investigator cannot determine whether a "yes"-response resulted from truthful answering or from the randomization process. However, the proportion of "yes"-responses that have not been prompted by the randomization procedure (the lifetime prevalence of cocaine use in the present example) can be estimated by straightforward probability calculations based on the a-priori known probability distribution of the randomization device. Suppose that participants are asked to answer in the affirmative to the question on cocaine use if the die shows 1-5 and are asked to respond truthfully if the die shows a 6. Assuming a fair die (i.e., the probability of being prompted to answer in the affirmative is equal to $\underline{p} = 5/6$), the observed proportion of "yes"-responses ($\lambda$) equals the sum of the proportion of cocaine users ($\pi$) and the proportion of non-users ($1 - \pi$) which were asked to answer in the affirmative with the randomization probability $\underline{p}$:

$$\lambda = \pi + \underline{p} (1 - \pi)$$

A simple algebraic rearrangement yields an unbiased maximum likelihood estimate of the prevalence of the critical behavior given an observed proportion of "yes"-responses and the randomization probability:

$$\pi = (\lambda - \underline{p}) / (1 - \underline{p})$$

with the variance of $\pi$ given by

$$\text{var}(\pi) = \lambda (1 - \lambda) / N [(1 - p)]^2$$

### insert figure 1 about here ###

As Figure 1 shows, the forced response variant of the RRT can also be represented as a special case of the more general family of multinomial processing tree models (Batchelder

& Riefer, 1999; Hu & Batchfelder, 1994; Riefer & Batchelder, 1988). The population is divided into two groups: those, who have been engaged in the critical behavior ($\pi$) and those, who have not ($\beta = 1 - \pi$). The former group will answer "yes" to the sensitive question irrespective of the outcome of the randomization device, whereas the latter group will only answer "yes" if prompted by the randomization device. This multinomial representation is advantageous in that established procedures of multinomial modeling may be used to estimate parameters and to test the applicability of parameter restrictions, such as a test of the null-hypothesis that the prevalence of the critical behavior does not differ from zero ($\pi = 0$). Using the multinomial modeling framework, it is also possible to formulate more complex models incorporating additional parameters. For example, one could estimate a multigroup model with and without cross-group equality restrictions on the parameters to examine whether the prevalence of the critical behavior differs across gender or various age groups. Similarly, one may wish to compare prevalence estimates obtained by using traditional direct questioning formats with RRT-based prevalence estimates. This may be achieved by setting equality constrains on the $\pi$ parameters and comparing the difference in model fit in relation to the unconstrained model as assessed by the asymptotically $\chi^2$ distributed log-likelihood ratio statistic $\underline{G}^2$.

Given that the randomization procedure of the RRT guarantees that a "yes"- response is no longer unequivocally indicative of an undesirable attitude or behavior, the RRT allows for group-based prevalence estimates without revealing an individual's status on the respective attribute. More honest responses are thereby encouraged and have, in fact, been observed. As demonstrated in studies where the true status of each individual on the sensitive attribute was known, RRT-based prevalence estimates are less biased than those based on conventional data collection techniques. Likewise, comparative surveys on such diverse issues as academic dishonesty, illegal drug use, employee theft, shoplifting, and rape have repeatedly shown that the RRT yields higher and presumably more valid prevalence estimates

than traditional surveys (for reviews, see Chaudhuri & Mukerjee, 1988; Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005).

Successful applications notwithstanding, the RRT has been criticized as being susceptible to respondents who are not answering as directed by the randomization device (Campbell, 1987). When employing the RRT, two types of non-compliance with the RRT instructions may occur. First, respondents may refuse to answer truthfully when prompted by the randomization device (respondent jeopardy). The superiority of the RRT over traditional direct questioning formats is owed to the fact that respondents are more likely to admit a critical behavior. However, the RRT may be able to reduce this type of non-compliance rather than to eliminate it altogether. Second, the randomization procedure of the RRT introduces another type of non-compliance, namely the denial to comply with the RRT instruction of answering "yes" to a sensitive question regardless of its content (risk of suspicion). Both types of non-compliance, respondent jeopardy and risk of suspicion, lead to a "no"-response, although the randomization device asks respondents to answer in the affirmative. In fact, there is evidence that cheating occurs (Edgell, Duchan, & Himmelfarb, 1992; Locander, Sudman, & Bradburn, 1976; Lensvelt-Mulders & Boeije, 2007; Shimizu & Bonham, 1978; Shotland & Yankowski, 1982; van der Heijden, van Gils, Bouts, & Hox, 2000). Respondents who are prompted by the randomization device to answer truthfully may answer "no" although they have performed the critical behavior, because, for example, they may not fully understand the rationale of the RRT, do not feel sufficiently protected when the probability of being asked to answer truthfully ($1 - \underline{p}$) is high, or may not trust the integrity of the randomization process (Landsheer, van der Heijden, & van Gils, 1999; Lensvelt-Mulders & Boeije, 2007; Soeken & McReady, 1982). Furthermore, respondents may answer "no" in spite of being prompted by the randomization device to answer in the affirmative irrespective of the item's content to avoid even the slightest suspicion that they have been engaged in a critical behavior or because they may feel uncomfortable when being "forced to be dishonest" (Lensvelt-Mulders

& Boeije, 2007, p. 600). Whatever causes non-compliance with the RRT instructions, the RRT underestimates the prevalence of the critical behavior to the extent that participants fail to comply with the instructions and deny the critical behavior even though they are asked to attest to it.

### insert figure 2 about here ###

Addressing this issue, Clark and Desharnais (1998) proposed a modification of the forced response model: In what we will refer to as the cheating detection model (CDM), it is assumed that some respondents may not comply with the RRT instructions and answer "no" irrespective of the outcome of the randomization device. Figure 2 illustrates how the CDM can be depicted as a multinomial model dividing the population into three distinct and exhaustive groups: The first group ($\pi$) represents the proportion of compliant and honest "yes"-respondents, that is, respondents who honestly admit the critical behavior (honest cocaine users). The second group ($\beta$) is the proportion of compliant and honest "no"-respondents, that is, respondents truthfully denying the critical behavior (honest non-users). The third group ($\gamma = 1 - \pi - \beta$) represents the proportion of non-compliant cheaters who do not comply with the instruction of the RRT and answer "no" to the sensitive question irrespective of the outcome of the randomization process. It is important to note that nothing is assumed about whether non-compliant respondents have actually engaged in the sensitive behavior. Conceivably, respondents who are prompted by the randomization device to answer truthfully deny a critical behavior in which they have in fact been engaged; but it is also possible that respondents who have not been engaged in the critical behavior want to avoid even the possibility of anyone thinking that they committed a prohibited or undesirable act and answer "no" despite being prompted by the randomization device to answer affirmatively. Thus, the true status of a respondent choosing not to follow the instructions remains unknown.

As the proportions $\pi$, $\beta$ and $\gamma$ are constrained to add up to 1, the CDM contains two independent parameters that cannot be estimated on the basis of only one proportion of "yes"-responses provided by traditional RRT procedures. An experimental approach is needed to obtain at least two degrees of freedom for the purpose of identification. More specifically, 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 say "yes" (Clark & Desharnais, 1998). Figure 2 shows 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 be replaced with probability $\underline{p}_2$. Under the assumption that the same proportions apply in both groups when participants are randomly assigned to conditions ($\pi_1 = \pi_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$), the CDM allows to observe two independent proportions of "yes"-responses ($\lambda_1$ and $\lambda_2$), which are sufficient to estimate the two independent parameters $\pi$ and $\beta$ (with $\gamma = 1 - \pi - \beta$). For this particular model, Clark and Desharnais (1998) provide closed-form solutions for unbiased maximum likelihood estimates of the parameters $\pi$, $\beta$, and $\gamma$ as well as a statistical test of the null hypothesis that no cheating occurs. When $\underline{y}$ is the frequency of "yes"-responses, $\pi$ and $\beta$ are given by

$$\pi = (\underline{p}_2\, \underline{y}_1 / \underline{n}_1) / (\underline{p}_2 - \underline{p}_1)$$

$$\beta = (\underline{y}_2 / \underline{n}_2 - \underline{y}_1 / \underline{n}_1) / (\underline{p}_2 - \underline{p}_1)$$

$$\pi = (\underline{p}_2\, \lambda_1) / (\underline{p}_2 - \underline{p}_1)$$

$$\beta = (\lambda_2 - \lambda_1) / (\underline{p}_2 - \underline{p}_1)$$

and $\gamma$ may be easily computed by $\gamma = 1 - \pi - \beta$. Again, it is also possible (and probably preferable) to estimate the CDM using the multinomial modeling framework. Converting the tree model depicted in Figure 2 into a statistically equivalent binary tree representation (see Appendix for details on the reparameterization) allows estimating the parameters using the EM algorithm, testing parameter restrictions, and formulating more complex models incorporating additional parameters. Employing the EM algorithm has the additional

advantage that the parameter estimates, which are proportions, will always be within the range of 0 - 1.

The cheating detection model offers a unique theoretical advantage over both traditional surveys and previous RRT models: If no cheating occurs ($\gamma = 0$), the parameter $\pi$ provides an asymptotically unbiased estimate of the population proportion engaged in the sensitive behavior. If there is a significant proportion of non-compliant respondents, it is possible to compute both an upper and a lower bound for the prevalence of the sensitive attribute by assuming that non-compliant respondents either did or did not engage in the critical behavior (Musch, Bröder, & Klauer, 2001). The CDM may also be considered as a generalization of the forced response variant of the RRT in that the proportion of cheaters is explicitly modeled, but may also become zero, in which case the CDM is identical to the forced response model extended to two groups.

However, the CDM is also subject to certain problems. Recall that for the purpose of identification it is necessary to question two independent samples using different randomization probabilities, while assuming the model parameters $\pi$, $\beta$, and $\gamma$ to be equal across groups. Whereas $\pi$ and $\beta$ are unlikely to differ systematically across groups over and above sampling fluctuations when participants are randomly assigned to conditions, the assumption of equal proportions of cheaters across conditions ($\gamma_1 = \gamma_2$) is problematic in that it might well be possible that the likelihood to disregard the RRT instructions is a function of the randomization probabilities applied (e.g., Scheers, 1992; Soeken & McReady, 1982). From a Bayesian perspective, the conditional probability of being identified as a respondent who has actually been engaged in the critical behavior given a “yes”-response to the sensitive question depends on the proportion of respondents who have been prompted to answer affirmatively irrespective of their true status: As the randomization probability declines, the likelihood that a “yes”-answer is associated with the critical behavior increases. Consequently, respondents may be more likely to disregard the RRT instructions in the

condition with a lower randomization probability than in the condition with a higher randomization probability. Under these circumstances, the assumption of equal proportions of cheaters across groups would be violated, which in turn might lead to biased parameter estimates. This would be especially critical if the violation of the assumption of equal proportions of cheaters across groups resulted in inflated estimates of the prevalence of the critical behavior. Unfortunately, it is unknown how the CDM performs in the presence of unequal proportions of cheaters across groups.

A related problem of the CDM is that the two independent proportions of "yes"-responses are just sufficient to estimate the parameters, that is, the CDM is saturated and will fit perfectly to the data, even when there are serious departures from equal proportions of cheaters across groups. Hence, it is not possible to get any clues about violations of assumptions and potentially biased parameter estimates, as the $\underline{G}^2$ fit statistic will be close to zero in any case. Given these concerns, we propose an enhancement of the cheating detection model (ECDM). The basic idea is to extend the CDM to the application on three different samples, each of which questioned with a different randomization probability, while keeping the assumption of equal parameters $\pi$, $\beta$, and $\gamma$ across groups. By this extension, the ECDM provides three independent proportions of "yes"-responses to estimate the two parameters $\pi$ and $\beta$ (with $\gamma = 1 - \pi - \beta$). Thus, the ECDM is overidentified and, thereby, provides a means for detecting violations of assumptions and model misfit in general. The multinomial representation of the ECDM is similar to the one shown in Figure 2 with the sole exception that this model is fit to the data in three groups with three different randomization probabilities $\underline{p}_1$, $\underline{p}_2$, and $\underline{p}_3$ (where $\underline{p}_1 \neq \underline{p}_2 \neq \underline{p}_3$). Although the assumption inherent in the CDM that the proportions of cheaters do not differ across conditions also applies to the ECDM, the latter has the advantage that it may be rejected on the basis of a significant $\underline{G}^2$ statistic with one degree of freedom. This feature is especially important, because applied researchers will

typically not be aware as to whether any assumption is violated, and therefore might run risk to obtain potentially severely biased parameter estimates without knowing.

The previous discussion highlights three points of critical importance for evaluating the usefulness of both the CDM and the ECDM in applied research settings. First, it is vital to examine the effect of violations of assumptions -- which are not detectable as far as the CDM is concerned -- on the estimate of the prevalence of the critical behavior and the proportion of cheaters, respectively. At the presence of unequal proportions of cheaters across groups, it is likely that the parameter estimates will be biased. It is not obvious, however, to what degree the parameter estimates will be biased, and, even more importantly, whether the prevalence of the critical behavior will be underestimated or overestimated. Whereas downward biased prevalence estimates, albeit clearly undesirable, would merely result in more conservative results, the possibility that prevalence might be overestimated effectively renders both the CDM and the ECDM useless. Second, it is important to examine the ability of the ECDM to detect violations of assumptions when they are actually present. As the CDM is saturated and will perfectly fit the data irrespective of model misspecifications, the advantage of the ECDM is that it is overidentified and may thereby be subject to empirical falsification. Yet, the statistical power for rejecting the null hypothesis that the ECDM fits the data is unknown. Finally, a major drawback of all techniques belonging to the RRT family is their poor statistical efficiency compared to direct questioning formats (e.g., Lensvelt-Mulders, Hox, & van der Heijden, 2005). The standard errors in randomized response models are a function of both sampling variation and variation due to the randomizing device, in turn resulting in reduced power for parameter restrictions and higher sample size requirements. In traditional single group randomized response models, it is clear that the statistical efficiency depends on the randomization probability, because more random noise is introduced into the data as the probability of being prompted to answer truthfully declines. Considering the CDM, the power estimates to detect the presence of cheating provided by Clark and Desharnais (1998) suggest

that power may be enhanced by choosing more diverging randomization probabilities. However, this relation does not necessarily hold for the ECDM, whose statistical efficiency remains to be determined. In pursuing these issues, the present study contributes both to the understanding of the statistical properties of the CDM and the ECDM as well as to the provision of recommendations for applied research.

## Method

A series of computer simulation studies were performed to (1) investigate the effects of violations of the assumption of equal proportions of cheaters across different groups on the accuracy of parameter estimates, (2) determine the statistical power to detect violations of assumptions using the ECDM, and (3) compare the CDM and the ECDM with respect to parameter standard errors and power for parameter restrictions. As elaborated in greater detail in the preceding sections, the CDM and the ECDM are special cases of the more general family of multinomial processing tree models. The model shown in Figure 2 can easily be reparameterized into an equivalent binary tree model for which model parameters can be estimated by the EM algorithm for general tree models (see Appendix for details of the reparameterization). In the present investigation, the asymptotically $\chi^2$ distributed loglikelihood ratio $\underline{PD}^{\lambda=0}$ (equivalent to the $\underline{G}^2$ statistic) power divergence statistic was used as discrepancy measure for both parameter estimation and power calculation.

### Violations of assumptions

Both the CDM and the ECDM assume the cross-group equality of the parameters $\pi$, $\beta$, and $\gamma$. Since there is no reason to assume that $\pi$ and $\beta$ systematically differ across groups over and above sampling fluctuations when participants are randomly assigned to conditions, we restricted the simulation studies to the effects of unequal proportions of cheaters ($\gamma$) across groups. Although, in principle, the proportion of cheaters in group $\underline{k}$ ($\gamma_k$) can take any value

between 0 and 1, we require the proportion of cheaters to vary symmetrically around a mean population value of $\gamma$. In order to quantify the magnitude of the violation of equal $\gamma$ parameters, we define $v$ as the absolute difference between the true proportions of cheaters in two groups:

$$v = | \gamma_k - \gamma_j |, \quad k \neq j$$

Hence, when the sample sizes are equal across groups (i.e., $n_k = n_j$), the true population value of $\gamma$ is the mean of its realizations in the $k$ groups and the maximum of $v$ equals twice the value of $\gamma$. The proportion of cheaters for the second group in the ECDM was always set equal to the mean population value of $\gamma$ to mimic the CDM as closely as possible. For example, given a mean population value of $\gamma = .2$ and a model violation of $v = 0.1$, the proportion of cheaters in group $k$ becomes $\gamma_1 = .15$, $\gamma_2 = .25$ in the CDM and $\gamma_1 = .15$, $\gamma_2 = .2$, $\gamma_3 = .25$ in the ECDM.

Randomization probabilities

The requirement of symmetric deviations of $\gamma_k$ around the mean population value of $\gamma$ results from the assumption of a linear relationship between a randomization probability and the likelihood to disregard the RRT instructions. Assuming that equal randomization probabilities $p_k$ lead to a specific proportion of cheaters, the extent to which the proportion of cheaters decreases at higher randomization probabilities should be the identical to the increment of cheating at lower randomization probabilities. For reasons of coherence and simplicity, the present studies used randomization probabilities that vary symmetrically around $p = .5$.

Data generation

Observed frequency counts for the response categories in group $k$ were generated based on combinations of the population parameters $\pi$ (which was always restricted to equality across groups), $\gamma_k$, and $p_k$. The resulting category probabilities were multiplied with

the number of observations $n_k$ to obtain observed frequency counts for the kth group. The total sample size N was always equally distributed across groups (i.e., $n_k = n_j$).

Study outcomes

The absolute bias of the γ parameter ($\Delta\gamma$) and the π parameter ($\Delta\pi$) was examined to investigate the effect of violations of assumption on the parameter estimates. The general equation for the absolute bias is:

$$\Delta\theta = \hat{\theta} - \theta$$

where θ is the true population value and $\hat{\theta}$ is the estimate of the respective parameter. A relative percentage bias ($RB_\theta$) may be computed by standardizing $\Delta\theta$:

$$RB_\theta = \Delta\theta / \theta$$

For the comparison of the efficiency of the two models, we examined the standard errors of π and γ. Hu and Batchfelder (1994) provided closed-form solutions for estimating the observed Fisher information matrix, which was used to obtain an estimate of the parameters' variance. Furthermore, the statistical power for rejecting the null hypothesis that π and γ, respectively, equals zero was examined. Power was calculated by evaluating the noncentral $\chi^2$ distribution at a significance level of $\alpha = .05$ with the difference of the $G^2$ fit-statistics of the restricted model and the unrestricted model as an estimate of the noncentrality parameter. The power to detect violations of assumptions (i.e., unequal γ parameters) using the ECDM was determined by estimating a model assuming equal γ parameters and utilizing the resulting discrepancy measure $G^2$ as an estimate of the noncentrality parameter. It should be noted that this computation merely gives the power to detect global model misfit. However, the only source of misfit within the present simulation studies constituted unequal γ parameters.

## Results

Effects of unequal cheating proportions on parameter estimates

In order to examine the effects of violations of assumptions on the estimates of $\pi$ and $\gamma$, a simulation study was conducted based on true $\gamma$ parameters between .05 and .5, $\pi$ parameters between .05 and 1 - $\gamma$, $\beta = 1 - \pi - \gamma$, and randomization probabilities of $\underline{p}_1 = .75$, $\underline{p}_2 = .25$ and $\underline{p}_1 = .75$, $\underline{p}_2 = .5$, $\underline{p}_3 = .25$ for the CDM and the ECDM, respectively. Since the EM algorithm internally operates with relative frequencies, parameter estimates are not affected by the sample size used. Therefore, the sample size was arbitrarily fixed at 1,000 observations. Violations of equal $\gamma$ parameters across groups were introduced by varying levels of $\underline{v}$. Table 1 summarizes the absolute and relative bias in $\gamma$ as a function of $\pi$ and the model violation $\underline{v}$. Because the absolute bias in $\gamma$ ($\Delta\gamma$) is independent of both the number of groups and the population value of $\gamma$, for the clarity of presentation only the results obtained using the CDM (i.e., two groups) at $\gamma = .2$ are reported.[1] Generally, $\gamma$ is consistently underestimated, with an increasing bias relative to an increase in $\underline{v}$ and $\pi$. Since $\Delta\gamma$ does not depend on the true value of $\gamma$, the relative bias increases with smaller values of $\gamma$. For instance, when the proportions of cheaters differ to $\underline{v} = 0.1$ and $\pi$ is .4, $\gamma$ is underestimated by 0.08, which constitutes a small bias at high proportions of cheaters, but a severe bias at small proportions of cheaters.

Considering the estimates for $\pi$, there is again no effect of the number of groups, so Table 2 only presents the absolute and relative bias in $\pi$ obtained by using the CDM. Similarly to the results reported for $\gamma$, $\pi$ is consistently underestimated relative to increasing model violation $\underline{v}$; however, the bias in $\pi$ is slightly lower than the bias in $\gamma$ at identical values of $\pi$, $\gamma$, and $\underline{v}$. Unlike $\Delta\gamma$, the absolute bias in $\pi$ depends on both $\gamma$ and $\pi$. At higher values of $\pi$, the absolute bias increases, whereas the relative bias decreases. Moreover, $\Delta\pi$ increases with an increasing proportion of cheaters. The effect of $\gamma$ is most pronounced when $\pi$ is high and model violation is severe, but diminishes when $\underline{v}$ is small.

Taken together, violations of the assumption of equal proportions of cheaters across groups lead to downward biased estimates of $\pi$ and $\gamma$ using both the CDM and the ECDM. The relative bias increases with smaller values of the corresponding parameter, but decreases with higher values of the second parameter, i.e., $RB_\pi$ is likely to be high when $\pi$ is small and $\gamma$ is high, and vice versa. However, $RB_\pi$ primarily depends on the value of $\pi$ rather than on the value of $\gamma$. The absolute bias in the parameter estimates is most likely to be high when $\gamma$, $\pi$, and $\underline{v}$ are high. Although the bias in parameter estimates is substantial even at low to moderate differences between $\gamma_k$, the models act conservatively by underestimating both the proportion of cheaters and the prevalence of the critical behavior.

### insert tables 1 and 2 about here ###

### insert figure 3 about here ###

Because the bias in $\pi$ and $\gamma$ also depends on the randomization probabilities, an additional simulation was conducted with $\gamma = .2$, $\pi = .2$, $\beta = .6$, various levels of $\underline{v}$, $\underline{N} = 1{,}000$, and three different randomization probabilities: (1) $\underline{p}_1 = .9$, $\underline{p}_2 = .1$; (2) $\underline{p}_1 = .75$, $\underline{p}_2 = .25$; (3) $\underline{p}_1 = .66$, $\underline{p}_2 = .33$ for the CDM and analogously for the ECDM. Results are similar for the two models with respect to both the bias in $\pi$ and $\gamma$, so we restrict ourselves to the effect of the randomization probabilities on the bias in $\pi$ using the CDM. Figure 3 shows that the absolute bias in $\pi$ may be dramatically reduced by choosing more diverging randomization probabilities. In fact, at $\underline{v} = 0.1$, $\Delta\pi$ is 4 times higher using randomization probabilities of $\underline{p}_1 = .66$, $\underline{p}_2 = .33$ ($\Delta\pi = -0.09$) when compared to randomization probabilities of $\underline{p}_1 = .9$, $\underline{p}_2 = .1$ ($\Delta\pi = -0.02$). Furthermore, there is a notable bend in the curve for randomization probabilities of $\underline{p}_1 = .66$, $\underline{p}_2 = .33$. At smaller $\underline{v}$, the bias in $\pi$ grows linearly, but when $\underline{v}$ exceeds 0.2, $\Delta\pi$ slowly approaches the maximum of -.2. This behavior is due to a concurrent underestimation

of $\gamma$ by nearly 100%, i.e., the maximum likelihood estimate of $\pi$ given a near zero estimate of $\gamma$ only changes marginally with increasing model violation $\underline{v}$. These results strongly suggest the use of diverging randomization probabilities to minimize the bias in parameter estimates.

### insert figures 4 – 6 about here ###

Power to detect unequal proportions of cheaters

The ECDM has the advantage that the model is overidentified and, thereby, provides a test of the null hypothesis that the proportions of cheaters are equal across groups. The power to detect violations of assumptions depends on a number of factors, including the degree of bias in the parameter estimates introduced by violations of the assumption of equal proportions of cheaters across groups. Given that the relative bias at a given level of $\underline{v}$ is likely to be severe when the population value of the respective parameter is small, power to detect violations of assumptions tends to be higher at smaller values of $\pi$ and $\gamma$. However, the effects of $\pi$ and $\gamma$ are rather small compared to the effects of degree of violation, sample size, and randomization probability. We therefore restricted the simulations on the latter factors with $\pi$ and $\gamma$ fixed at proportions likely to be encountered in practice. Thus, in order to determine the power to detect model violations, a simulation study was performed using $\gamma$ = .2, $\pi$ = .2, $\beta$ = .6, different levels of $\underline{v}$, sample sizes ranging from $\underline{N}$ = 250 to $\underline{N}$ = 10,000, and three different randomization probabilities: (1) $\underline{p}_1$ = .9, $\underline{p}_2$ = .5, $\underline{p}_3$ = .1; (2) $\underline{p}_1$ = .75, $\underline{p}_2$ = .5, $\underline{p}_3$ = .25; (3) $\underline{p}_1$ = .66, $\underline{p}_2$ = .5, $\underline{p}_3$ = .33. Figures 4-6 present the power curves for various sample sizes, with the total sample equally distributed across groups. Generally, the power to detect small to moderate violations is low even at large samples sizes. For example, the power to detect a cross group difference in $\gamma$ of $\underline{v}$ = 0.1 with $\underline{N}$ = 10,000 and randomization probabilities of $\underline{p}_1$ = .75, $\underline{p}_2$ = .5, $\underline{p}_3$ = .25 is merely 14.7% (Figure 5). The power to detect model violations, however, also varies with the randomization probabilities applied. When

choosing less diverging randomization probabilities (i.e., $p_1$ = .66, $p_2$ = .5, $p_3$ = .33; Figure 4), power is only acceptable at extremely large sample sizes when at least one parameter is underestimated by nearly 100% (which occurs when v exceeds 0.2). Conversely, power may be vastly enhanced by choosing randomization probabilities of $p_1$ = .9, $p_2$ = .5, $p_3$ = .1 (Figure 6). At identical levels of model violations, power at N = 1,000 using randomization probabilities of $p_1$=.9, $p_2$ = .5, $p_3$=.1 is even higher than power at N = 5,000 using randomization probabilities of $p_1$=.66, $p_2$ = .5, $p_3$=.33. Taking into account that more diverging randomization probabilities also tend to reduce the amount of bias in the parameter estimates, the power to detect model violations is higher for more diverging randomization probabilities although the parameter estimates are less biased. For example, $\pi$ is underestimated by 25% ($RB_\pi$ = -0.25) using randomization probabilities of $p_1$ = .66, $p_2$ = .5, $p_3$ = .33; $p_1$ = .75, $p_2$ = .5, $p_3$ = .25; and $p_1$ = .9, $p_2$ = .5, $p_3$ = .1 when v = .05, v = .09, and v = .18, respectively. Given a relative bias in $\pi$ of $RB_\pi$ = -0.25, the power to detect violations of assumptions with N = 5,000 is about 5.6%, 8.9%, and 48.3%, respectively. Thus, given a specific parameter bias, power to detect model violations is clearly higher at more diverging randomization probabilities

### insert table 3 about here ###

Statistical efficiency

In order to determine if the rather low power to detect model violations justifies the use of the ECDM, the effects of the number of groups on the efficiency of parameter estimates and the power for parameter restrictions are explored in the following. To this end, simulation studies were conducted based on true $\gamma$ parameters between .05 and .5, $\pi$ parameters between 0 and 1 - $\gamma$, $\beta = 1 - \pi - \gamma$, sample sizes ranging from N = 250 to N = 5,000, and randomization probabilities of $p_1$ = .9, $p_2$ = .1; $p_1$ = .75, $p_2$ = .25; and $p_1$ = .66, $p_2$ =

.33 for CDM and analogously for the ECDM. The γ parameters were restricted to equality across groups, that is, the assumption of equal proportion of cheaters across groups was not violated. Since results are similar for both π and γ with respect to parameter standard errors and power for parameter restrictions, only the results for π with γ fixed at .2 are reported. Table 3 summarizes the standard errors and the power for rejecting the null hypothesis that π equals zero as a function of the sample size, randomization probability, π, and the model used. Standard errors for π are higher in the ECDM than in the CDM with most pronounced differences at small sample sizes. As a consequence, power for rejecting the null hypothesis that π equals zero is also higher using the CDM; however, the differences between the power of the models rarely exceed 10%. Contrasting the fact that the differences between the CDM and the ECDM in the parameter standard errors are negligible when the sample is large, the power differences remain essentially unchanged with increasing sample sizes with the CDM consistently outperforming the ECDM. Consistent with the effect of the randomization probabilities on the power to detect model violations reported above, power for parameter restrictions is clearly higher at more diverging randomization probabilities. In fact, the sample size needs to be almost 10 times as large to achieve a similar power with randomization probabilities of $\underline{p}_1 = .66$, $\underline{p}_2 = .33$ compared to the power using randomization probabilities of $\underline{p}_1 = .9$, $\underline{p}_2 = .1$.

## Discussion

The present study evaluated the robustness of the cheating detection modification of the randomized response technique and proposed an enhancement that allows obtaining a testable model. The results demonstrate that violations of the assumption of equal proportions of cheaters across groups result in biased estimates using both the CDM and the ECDM. Although the bias in the parameter estimates is substantial even at minor violations of

assumptions, the models act conservatively by underestimating the prevalence of the critical behavior as well as the proportion of cheaters and, consequently, overestimating the proportion of compliant and honest "no"-respondents. Hence, confidence in the prevalence estimates obtained by these models may be placed in that they provide lower-bound estimates of the critical behavior, which may still be well above prevalence estimates obtained by using direct questioning formats and traditional randomized response models not considering cheating.

Still, the possibility that parameter estimates may be potentially severely biased due to violations of assumptions may be considered undesirable at best, and every model should provide a means to indicate whether it fits the data or not. Given that the CDM is saturated and will perfectly fit the data regardless of violations of assumptions, we proposed the enhanced cheating detection model. The advantage of the ECDM is that it is overidentified and thus may be empirically rejected on the basis of a significant $\underline{G}^2$ fit statistic with one degree of freedom. However, the power of the ECDM in detecting violations of assumptions was found to be rather low, unless violations are severe or the sample size is large, and the ECDM suffers a slight loss of efficiency. Nevertheless, the alternative in using a saturated model means that violations of assumptions and biased estimates will not be detectable at all, how large they may be. It seems more appropriate to accept a rather low power and a slight loss in efficiency to obtain a testable model that has the capability to indicate at least moderate misspecifications. Furthermore, power and efficiency may be greatly enhanced by choosing more diverging randomization probabilities.

The closer the randomization probabilities lie together (e.g., $\underline{p}_1 = .66$, $\underline{p}_2 = .33$), the larger the parameter standard errors, the stronger the bias in the parameter estimates when assumptions are violated, and the lower the power to detect violations of assumptions when they are present. Albeit these results suggest preferring randomization probabilities that lie further apart, it is obvious that the randomization probabilities cannot be varied indefinitely:

As the probability of being prompted to answer affirmatively approaches zero in one group, the privacy protection becomes nil and there would be no difference to traditional direct questioning formats. As noted by Clark and Desharnais (1998), exactly these conditions are likely to provoke the extent of cheating to differ as a function of the randomization probabilities, thereby introducing violations of the assumption of equal proportions of cheaters across conditions.

Notwithstanding, we recommend avoiding randomization probabilities that lie close together for a number of reasons. First, more diverging randomization probabilities vastly enhance statistical efficiency. In order to achieve an identical level of power for detecting that $\pi$ significantly differs from zero, the sample size at $\underline{p}_1 = .66$, $\underline{p}_2 = .33$ needs to be almost 10 times as large as the sample size at $\underline{p}_1 = .9$, $\underline{p}_2 = .1$. This is an important aspect since the (E)CDM is likely to be frequently used to estimate the prevalence of sensitive behaviors that are not very common in the population (e.g., drug use, illegal abortion, and rape). In the introductory example on cocaine use, the required sample size to reliably detect that 5% honestly admit to have ever used cocaine amounts to $\underline{N} = 10{,}000$ using less diverging randomization probabilities, but is merely $\underline{N} = 1{,}000$ using more diverging randomization probabilities. Hence, from the perspective of efficiency, more diverging randomization probabilities are clearly preferable.

Second, more diverging randomization probabilities substantially alleviate the effects of violations of assumptions on the bias in parameter estimates. As can be seen from Figure 3, the absolute bias in $\pi$ given a specific model violation is almost four times higher at less diverging than at more diverging randomization probabilities. Similarly, the parameter bias at $\underline{p}_1 = .9$, $\underline{p}_2 = .1$ is still less severe even when the difference of the proportion of cheaters across conditions at randomization probabilities of $\underline{p}_1 = .9$, $\underline{p}_2 = .1$ is three times the difference at randomization probabilities of $\underline{p}_1 = .66$, $\underline{p}_2 = .33$. The degree to which the extent of cheating depends on the randomization probabilities is unknown at present and finally is an empirical

question that remains to be investigated in future studies. However, unless randomization probabilities of $\underline{p}_1 = .9$, $\underline{p}_2 = .1$ triple the difference between the $\gamma$ parameters across conditions as compared to randomization probabilities of $\underline{p}_1 = .66$, $\underline{p}_2 = .33$, one fares better with more diverging randomization probabilities.

Finally, more diverging randomization probabilities greatly enhance the ability of the ECDM to detect violations of assumptions when they are actually present. Given that more diverging randomization probabilities also reduce the bias in the parameter estimates, this is an interesting result as it suggests that the power to detect violations does not primarily depend on the degree of bias in the parameter estimates. Power to detect violations of assumptions is higher for more diverging randomization probabilities, in spite of less biased parameters given particular model violations, again suggesting the use of more diverging randomization probabilities.

We have also shown how to represent both the CDM and the ECDM as special cases of the more general family of multinomial models. This representation is attractive in that it becomes unnecessary to perform by-hand calculations since established programs suitable for multinomial modeling, such as Appletree (Rothkegel, 1999), gpt (Hu, 1999), or HMMTree (Stahl & Klauer, 2007), may be used to estimate the parameters and to test the applicability of restrictions on them. Moreover, the multinomial modeling approach has the advantage to allow the formulation of more complex models with additional parameters representing, for example, different subgroups for which the parameters have to be estimated separately or direct questioning control conditions. As far as the ECDM is concerned, the possibility to include additional parameters also provides a means to deal with violations of assumptions when the likelihood ratio statistic indicates a significant departure from equality of the $\gamma$ parameters across groups. Apart from relaxing equality constraints on the $\gamma$ parameters, one may also wish to embed parameters that model the process of cheating. For example, assuming that the extent of cheating linearly depends on the randomization probabilities ($\gamma_k =$

$\underline{a}\ \underline{p}_k + \underline{b}$), it is possible to include the slope and the intercept of the regression equation in expanded multinomial models allowing for prevalence estimates while taking the inequality of the proportions of cheaters across conditions into account.

Implications for applied research

Based on the results of the present investigation, several recommendations can be provided for applied research. Generally, we suggest to prefer the ECDM over the CDM, as the ECDM is empirically falsifiable and thereby has the capability to indicate whether assumptions are violated and parameters may be biased. Unless statistical efficiency is a major concern, this advantage outweighs the slight loss in statistical efficiency compared to the CDM. To further improve statistical efficiency, we strongly suggest avoiding randomization probabilities that lie close to .5. The loss of power associated with less diverging randomization probabilities for both parameter restrictions and detecting model misfit is remarkable. Moreover, when violations of assumptions are present, more diverging randomization probabilities produce less biased parameter estimates. Thus, there is no compelling reason for using randomization probabilities that lie close together and we recommend choosing randomization probabilities that lie between $\underline{p}_1 = .75$, $\underline{p}_2 = .5$; $\underline{p}_3 = .25$ and $\underline{p}_1 = .9$, $\underline{p}_2 = .5$; $\underline{p}_3 = .1$. However, it is advisable to choose randomization probabilities contingent on the expected prevalence of the critical attitude or behavior and the available sample size. With respect to the latter, the question arises which sample may be regarded as sufficient. Although both the CDM and the ECDM require fairly large samples, the exact answer depends on a number of factors, including the expected prevalence of the critical behavior, the randomization probabilities, and the desired power. Figures 3-6 as well as Table 3 may serve as a reference for deciding on the sample size given these factors. Generally speaking, we would expect most investigations employing either the CDM or the ECDM to use a sample size equal to or greater than $\underline{N} = 1{,}000$. Given that typical applications of the

RRT merely contain one or two sensitive questions, obtaining such sample sizes does not seem as a principle limitation for using the (E)CDM.

Limitations and future research

Some limitations should be considered when interpreting the results of the present study. As with any simulation study, there are lots of conditions to manipulate and choices have to be made in order to keep the design manageable. We restricted the simulations studies to the effects of violations of the assumption of equal proportions of cheaters across groups and did not investigate the effect of violations of the assumption of cross-group equality of $\pi$ and $\beta$. Although the latter is not likely to occur when participants are randomly assigned to conditions, future research should determine the robustness of both the CDM and the ECDM to violations of these assumptions. Likewise, the forced response variant of the RRT (which underlies both the CDM and the ECDM) requires a-priory known randomization probabilities to estimate the parameters. There may be certain circumstances, where the randomization probabilities are unknown and need to be estimated, for example, when the participants' month of birth is used as randomization device. It is unknown how the (E)CDM performs when the actual randomization probabilities depart from the expected randomization probabilities. The basic idea in proposing the ECDM was to extend the CDM to three groups to gain one additional degree of freedom. In principle, it is also possible to use more than three groups, each of which questioned with different randomization probabilities. Based on the results of the present study, we hypothesize that the statistical efficiency will decline as the number of groups increases. However, we did not examine this issue, and future research should determine the benefits of this approach. Furthermore, in the present study, the total sample was always equally distributed across groups. From a practical perspective, it would be interesting to examine whether there is an optimal distribution of the total sample size on the experimental conditions with respect to changes in the statistical efficiency.

Conclusion

The cheating detection modification of the RRT provides a means to estimate the proportion of participants who fail to comply with the RRT instructions, but suffers from substantially biased prevalence estimates when assumptions are violated. Given that the CDM is saturated and will perfectly fit the data irrespective of violations of assumptions, we recommend employing the enhanced cheating detection model to obtain an empirically falsifiable model. We are confident that this model will be a valuable aid to researchers examining issues threatened by socially desirable responding.

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*, 97-145.

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

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

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.

Dawes, R. M., & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen [Traditional Guttman-scaling and randomized response]. In F. Peterman (Ed.), *Einstellungsmessung [Attitude measurement]* (pp. 117-133). Göttingen: Hogrefe.

Edgell, S. E., Duchan, K. L., & Himmelfarb, S. (1992). An empirical test of the unrelated question randomized response technique. *Bulletin of the Psychonomic Society, 30*, 153-156.

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

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

Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality & Quantity, 33*, 1–12.

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

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.

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*. Lengerich: Pabst.

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

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 31*, 696-700.

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

Shimizu, I. M., & Bonham, G. S. (1978). Randomized response technique in a national survey. *Journal of the American Statistical Association, 73*, 35-39.

Shotland, R. L., & Yankowski, L. D. (1982). The random response method: A valid and ethical indicator of the 'truth' in reactive situations. *Personality and Social Psychology Bulletin, 8*, 174-179.

Soeken, K. L., & Macready, G. B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin, 92*, 487-489.

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

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.

# Appendix

## Model estimation

Both the cheating detection model and the enhanced cheating detection model divide the responses to a sensitive question into three distinct and exhaustive groups: Compliant and honest "yes"-respondents ($\pi$), compliant and honest "no"-respondents ($\beta$), and non-compliant cheaters ($\gamma$). To estimate these parameters, the multinomial representation shown in Figure 2 needs to be reparameterized in order to satisfy the general form of binary processing tree models (Hu & Batchfelder, 1994). The probability for observing a "yes" (y) or "no" (n) response in group k is given by the sum of the probabilities of each branch i leading to that category:

$$p_{y,k} = \sum_{i=1}^{i} p_{i,y,k}$$

$$p_{n,k} = \sum_{i=1}^{i} p_{i,n,k}$$

In order to obtain parameter estimates and their associated standard errors, three different binary tree models need to be computed, each of which yielding an estimate of $\pi$, $\beta$, and $\gamma$ respectively. The branch probabilities $p_{i,j,k}$ for an estimate of $\pi$ are given by:

$$p_{1,y,k} = \pi_k$$

$$p_{2,y,k} = (1 - \pi_k)\,(1 - \eta_k)\,p_k$$

$$p_{1,n,k} = (1 - \pi_k)\,(1 - \eta_k)\,(1 - p_k)$$

$$p_{2,n,k} = (1 - \pi_k)\,\eta_k$$

where $p_k$ is the randomization probability (i.e., the probability of being prompted to answer in the affirmative) and $\eta_k$ reflects the proportion of non-compliant participants among those, who did not honestly admit to have been engaged in the critical behavior, such that $\beta_k = (1 - \pi_k)\,(1 - \eta_k)$ and $\gamma_k = (1 - \pi_k)\,\eta_k$. The utility of this particular representation is limited, since one

merely obtains conditional estimates of $\beta$ and $\gamma$. An alternative formulation is needed to obtain a direct estimate of $\beta$:

$$\underline{p_{1,y,k}} = \beta_k \, p_k$$

$$\underline{p_{2,y,k}} = (1 - \beta_k)(1 - \xi_k)$$

$$\underline{p_{1,n,k}} = \beta_k (1 - p_k)$$

$$\underline{p_{2,n}} = (1 - \beta_k) \, \xi_k$$

where the $\xi_k$ parameter reflects the proportion of non-compliant participants among those, who did not honestly report to have not been engaged in the critical behavior, such that $\pi_k = (1 - \beta_k)(1 - \xi_k)$ and $\gamma_k = (1 - \beta_k) \, \xi_k$. To obtain a direct estimate of $\gamma$ and associated standard errors, one may formulate the binary tree model as follows:

$$\underline{p_{1,y,k}} = (1 - \gamma_k) \, \mu_k$$

$$\underline{p_{2,y,k}} = (1 - \gamma_k)(1 - \mu_k) \, p_k$$

$$\underline{p_{1,n,k}} = (1 - \gamma_k)(1 - \mu_k)(1 - p_k)$$

$$\underline{p_{2,n}} = \gamma_k$$

where the $\mu_k$ parameter reflects the prevalence of the critical behavior among the compliant participants, such that $\pi_k = (1 - \gamma_k) \, \mu_k$ and $\beta_k = (1 - \gamma_k)(1 - \mu_k)$.

In typical applications, the model parameters, except for the randomization probability $\underline{p_k}$, will be restricted to equality across groups for the purpose of identification. It is assumed that the probability of being prompted to answer truthfully equals $1 - p_k$; however, asymmetric randomization probabilities may easily be embedded in these formulations by replacing the term $1 - p_k$ by the corresponding probability. The multinomial models may be estimated by suitable software packages such as Appletree (Rothkegel, 1999), gpt (Hu, 1999), or HMMTree (Stahl & Klauer, 2007).

## Footnotes

[1] Additional material may be obtained from the first author.

Author Note

Morten Moshagen, Jochen Musch, and Martin Ostapczuk, University of Duesseldorf, Germany; Robert Mischke and Arndt Bröder, University of Bonn, Germany; Edgar Erdfelder, University of Mannheim, Germany. This work was supported by a grant of the German Research Foundation (DFG, Grant No.: Mu 2674/1-1). Correspondence concerning this article should be addressed to Morten Moshagen, Institute for Experimental Psychology, University of Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany. E-Mail: morten.moshagen@uni-duesseldorf.de

# Tables

Table 1

Absolute and relative bias of $\gamma$^ as a function of $\pi$ and v at a population value of $\gamma = .2$

| | $\pi = .1$ | | $\pi = .2$ | | $\pi = .4$ | |
|---|---|---|---|---|---|---|
| v | $\Delta\gamma$ | $RB_\gamma$ | $\Delta\gamma$ | $RB_\gamma$ | $\Delta\gamma$ | $RB_\gamma$ |
| 0.05 | -0.03 | -0.17 | -0.04 | -0.18 | -0.04 | -0.20 |
| 0.10 | -0.07 | -0.34 | -0.07 | -0.36 | -0.08 | -0.41 |
| 0.15 | -0.10 | -0.50 | -0.11 | -0.54 | -0.12 | -0.61 |
| 0.20 | -0.13 | -0.67 | -0.14 | -0.72 | -0.16 | -0.81 |
| 0.25 | -0.160 | -0.80 | -0.18 | -0.90 | -0.20 | -1.00 |
| 0.30 | -0.18 | -0.90 | -0.20 | -1.00 | -0.20 | -1.00 |

Note. $\Delta\gamma$ = absolute bias, $RB_\gamma$ = relative bias. CDM; N = 1,000; $p_1 = .75$, $p_2 = .25$.

Table 2

Absolute and relative bias in π as function of γ, π, and v

| | | π = 0.1 | | π = 0.2 | | π = 0.4 | |
|---|---|---|---|---|---|---|---|
| γ | v | Δπ | $RB_\pi$ | Δπ | $RB_\pi$ | Δπ | $RB_\pi$ |
| 0.10 | 0.05 | -0.02 | -0.22 | -0.03 | -0.13 | -0.03 | -0.08 |
| | 0.10 | -0.04 | -0.44 | -0.05 | -0.26 | -0.07 | -0.16 |
| | 0.15 | -0.07 | -0.67 | -0.07 | -0.36 | -0.08 | -0.21 |
| | 0.20 | -0.07 | -0.68 | -0.08 | -0.37 | -0.09 | -0.21 |
| 0.20 | 0.05 | -0.02 | -0.23 | -0.03 | -0.13 | -0.03 | -0.09 |
| | 0.10 | -0.05 | -0.45 | -0.05 | -0.27 | -0.07 | -0.17 |
| | 0.15 | -0.07 | -0.68 | -0.08 | -0.40 | -0.10 | -0.26 |
| | 0.20 | -0.09 | -0.91 | -0.11 | -0.53 | -0.14 | -0.34 |
| | 0.25 | -0.10 | -1.00 | -0.13 | -0.66 | -0.17 | -0.42 |
| | 0.30 | -0.10 | -1.00 | -0.15 | -0.75 | -0.17 | -0.44 |
| 0.40 | 0.05 | -0.02 | -0.24 | -0.03 | -0.15 | -0.04 | -0.10 |
| | 0.10 | -0.05 | -0.48 | -0.06 | -0.29 | -0.08 | -0.20 |
| | 0.15 | -0.07 | -0.72 | -0.09 | -0.44 | -0.12 | -0.30 |
| | 0.20 | -0.10 | -0.96 | -0.12 | -0.58 | -0.16 | -0.40 |
| | 0.25 | -0.10 | -1.00 | -0.15 | -0.73 | -0.20 | -0.50 |
| | 0.30 | -0.10 | -1.00 | -0.18 | -0.88 | -0.24 | -0.59 |

Note. Δπ = absolute bias, $RB_\pi$ = relative bias. CDM; N = 1,000; $p_1$ = .75, $p_2$ = .25

Table 3
Standard errors of π and power for π = 0 as a function of the RRT model, sample size, randomization probability (p), and prevalence (π)

| | | CDM | | | | | | | | | | ECDM | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N = 250 | | N = 500 | | N = 1,000 | | N =2,500 | | N = 5,000 | | N = 250 | | N = 500 | | N = 1,000 | | N = 2,500 | | N = 5,000 | |
| $p_1$ | π | SE | 1 - β | SE | 1 - β | SE | 1 - β | SE | 1 - β | SE | 1 - β | SE | 1 - β | SE | 1 - β | SE | 1 - β | SE | 1 - β | SE | 1 - β |
| .9 | 0.025 | .031 | 0.14 | .022 | 0.23 | .015 | 0.41 | .010 | 0.78 | .007 | 0.97 | .037 | 0.11 | .026 | 0.17 | .019 | 0.30 | .012 | 0.62 | .008 | 0.89 |
| | 0.05 | .034 | 0.39 | .024 | 0.66 | .017 | 0.92 | .011 | 1.00 | .008 | 1.00 | .040 | 0.28 | .028 | 0.50 | .020 | 0.79 | .013 | 0.99 | .009 | 1.00 |
| | 0.075 | .036 | 0.68 | .025 | 0.93 | .018 | 1.00 | .011 | 1.00 | .008 | 1.00 | .043 | 0.53 | .030 | 0.82 | .021 | 0.98 | .014 | 1.00 | .010 | 1.00 |
| | 0.1 | .038 | 0.88 | .027 | 0.99 | .019 | 1.00 | .012 | 1.00 | .009 | 1.00 | .045 | 0.75 | .032 | 0.96 | .022 | 1.00 | .014 | 1.00 | .010 | 1.00 |
| .75 | 0.025 | .060 | 0.07 | .042 | 0.09 | .030 | 0.14 | .019 | 0.27 | .013 | 0.48 | .072 | 0.06 | .051 | 0.08 | .036 | 0.11 | .023 | 0.20 | .016 | 0.35 |
| | 0.05 | .061 | 0.13 | .043 | 0.22 | .031 | 0.39 | .019 | 0.76 | .014 | 0.96 | .073 | 0.18 | .052 | 0.17 | .037 | 0.29 | .023 | 0.60 | .016 | 0.88 |
| | 0.075 | .062 | 0.24 | .044 | 0.42 | .031 | 0.70 | .020 | 0.98 | .014 | 1.00 | .075 | 0.18 | .053 | 0.31 | .037 | 0.54 | .024 | 0.91 | .017 | 1.00 |
| | 0.1 | .064 | 0.38 | .045 | 0.64 | .032 | 0.91 | .020 | 1.00 | .014 | 1.00 | .076 | 0.28 | .054 | 0.49 | .038 | 0.78 | .024 | 0.99 | .017 | 1.00 |
| .66 | 0.025 | .092 | 0.06 | .065 | 0.07 | .046 | 0.08 | .029 | 0.14 | .021 | 0.23 | .112 | 0.06 | .079 | 0.06 | .056 | 0.07 | .035 | 0.11 | .025 | 0.17 |
| | 0.05 | .093 | 0.08 | .066 | 0.12 | .047 | 0.19 | .030 | 0.40 | .021 | 0.68 | .113 | 0.07 | .080 | 0.10 | .057 | 0.14 | .036 | 0.29 | .025 | 0.51 |
| | 0.075 | .094 | 0.13 | .067 | 0.21 | .047 | 0.36 | .030 | 0.72 | .021 | 0.95 | .114 | 0.10 | .081 | 0.16 | .057 | 0.26 | .036 | 0.56 | .026 | 0.84 |
| | 0.1 | .095 | 0.19 | .067 | 0.33 | .048 | 0.57 | .030 | 0.92 | .021 | 1.00 | .115 | 0.14 | .082 | 0.24 | .058 | 0.42 | .036 | 0.80 | .026 | 0.98 |

Note. CDM = cheating detection model; ECDM= enhanced cheating detection model; $p_1$= randomization probability in the first group; 1 - β = power. γ = .2.

Figure captions

Figure 1: Multinomial representation of the forced response variant of the RRT.

Figure 2: Multinomial representation of the cheating detection modification of the RRT.

Figure 3: Absolute bias in $\pi$ as a function of $\nu$ and the randomization probabilities using the CDM.

Figure 4: Power for detecting model violations at various sample sizes using the ECDM with randomization probabilities of $p_1 = .66$, $p_2 = .5$, $p_3 = .33$.

Figure 5: Power for detecting model violations at various sample sizes using the ECDM with randomization probabilities of $p_1 = .75$, $p_2 = .5$, $p_3 = .25$.

Figure 6: Power for detecting model violations at various sample sizes using the ECDM with randomization probabilities of $p_1 = .9$, $p_2 = .5$, $p_3 = .1$.

Figures

Figure 1

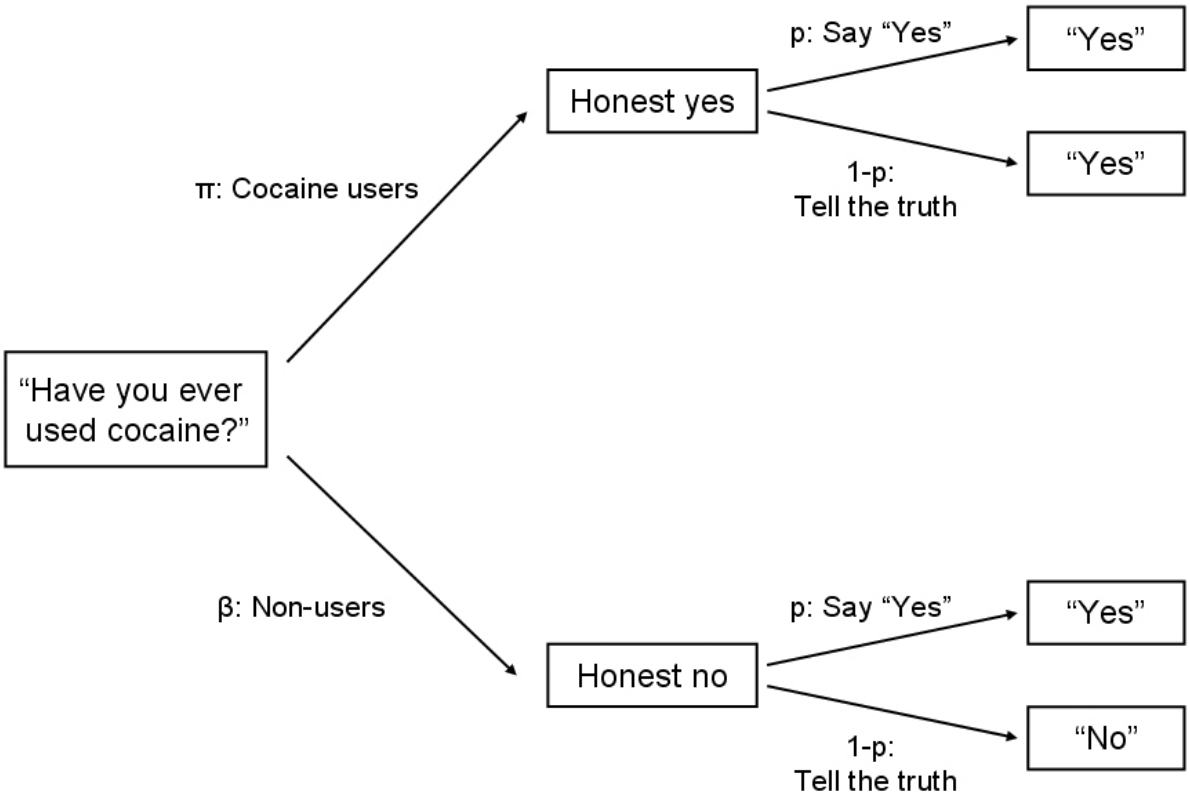


Figure 2

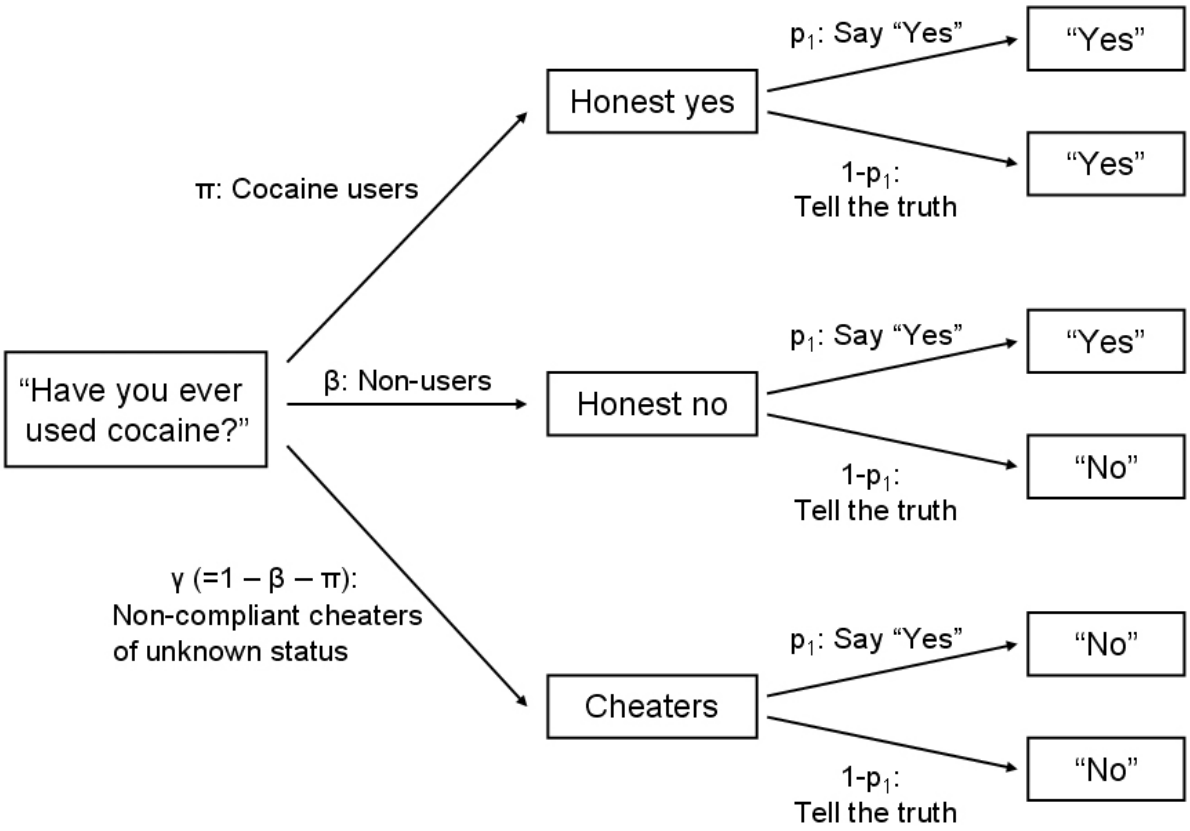

"Have you ever used cocaine?"
π: Cocaine users
β: Non-users
γ (=1 – β – π): Non-compliant cheaters of unknown status
Honest yes
Honest no
Cheaters
p1: Say "Yes"
1-p1: Tell the truth
"Yes"
"Yes"
p1: Say "Yes"
1-p1: Tell the truth
"Yes"
"No"
p1: Say "Yes"
1-p1: Tell the truth
"No"
"No"

Figure 3

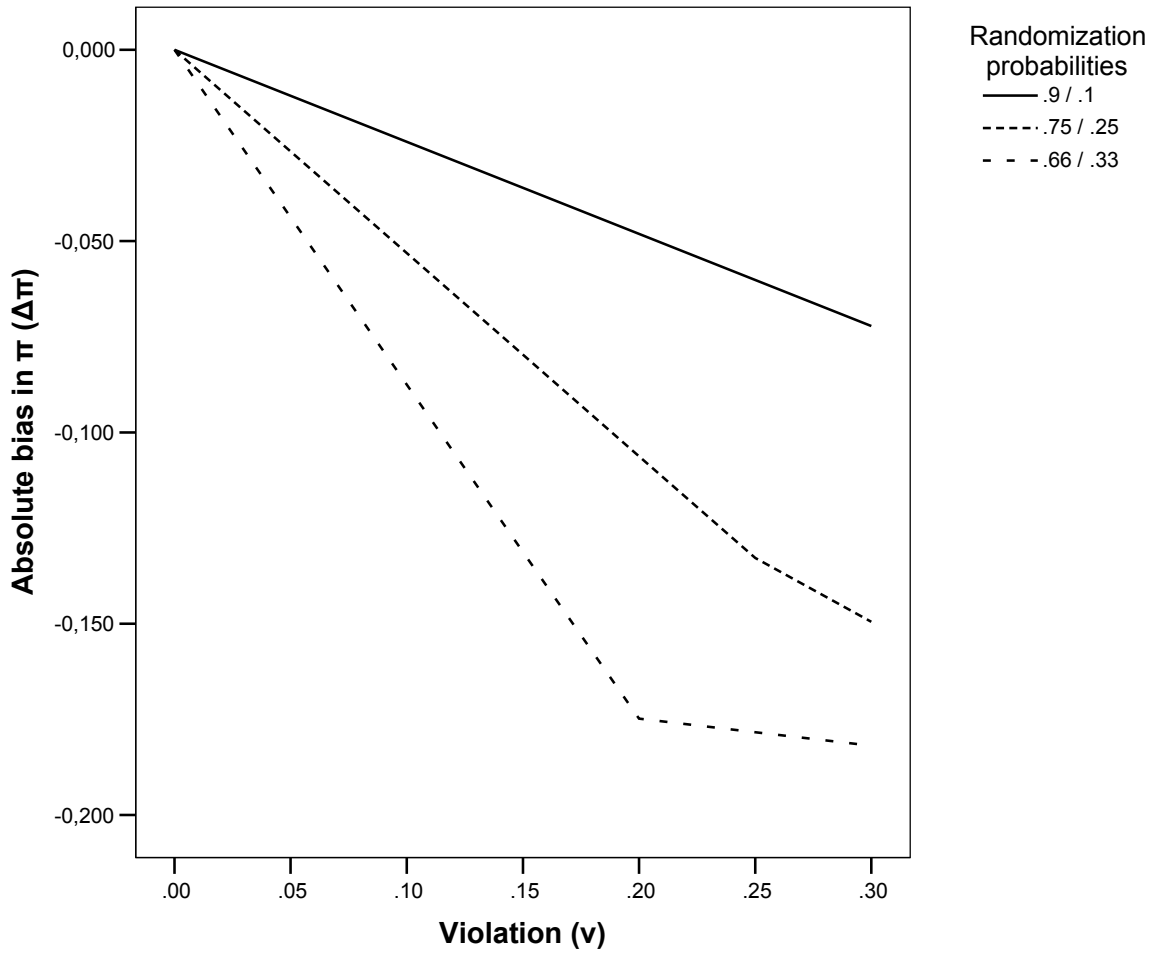

0,000
-0,050
-0,100
-0,150
-0,200
Absolute bias in π (Δπ)
.00
.05
.10
.15
.20
.25
.30
Violation (v)
Randomization probabilities
.9 / .1
.75 / .25
.66 / .33

Figure 4

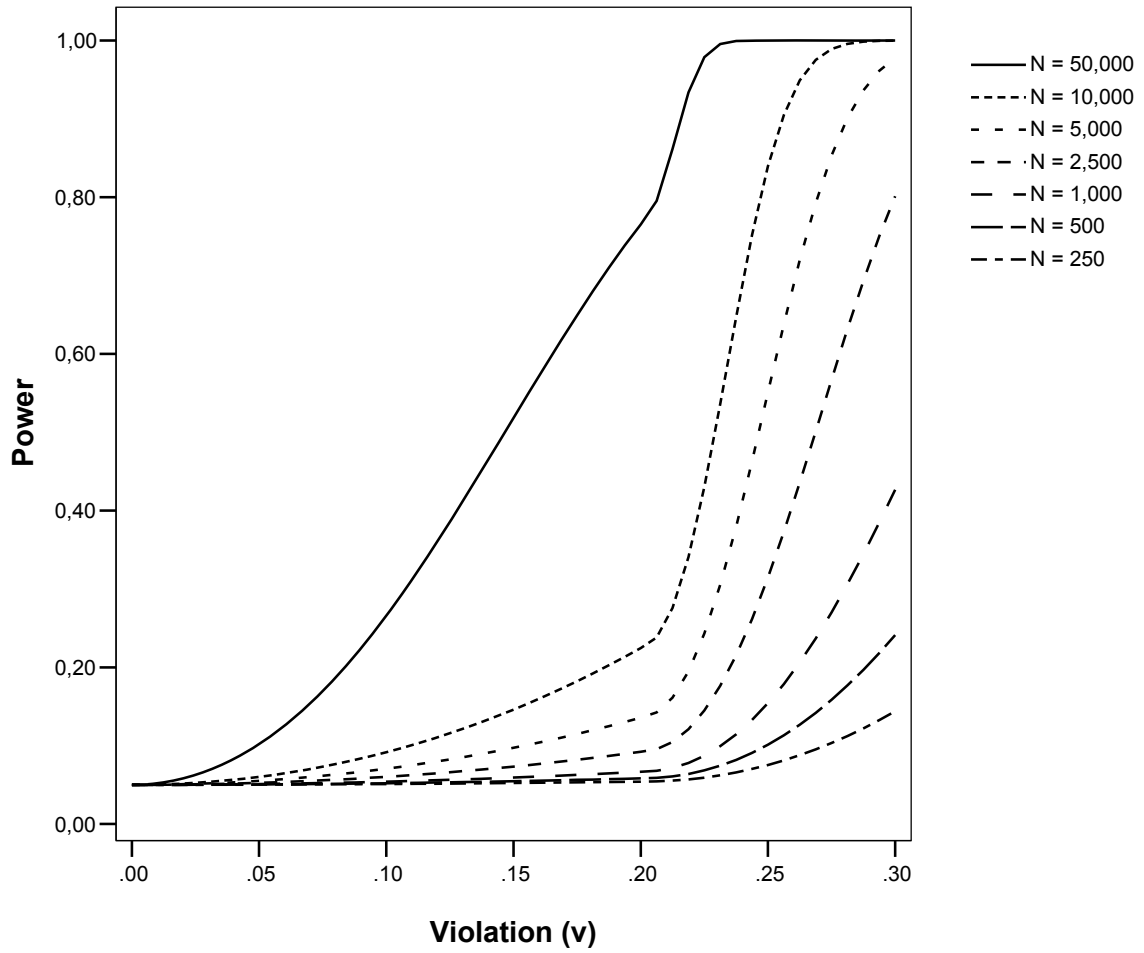

1,00
0,80
0,60
0,40
0,20
0,00
Power
.00
.05
.10
.15
.20
.25
.30
Violation (v)
N = 50,000
N = 10,000
N = 5,000
N = 2,500
N = 1,000
N = 500
N = 250

Figure 5

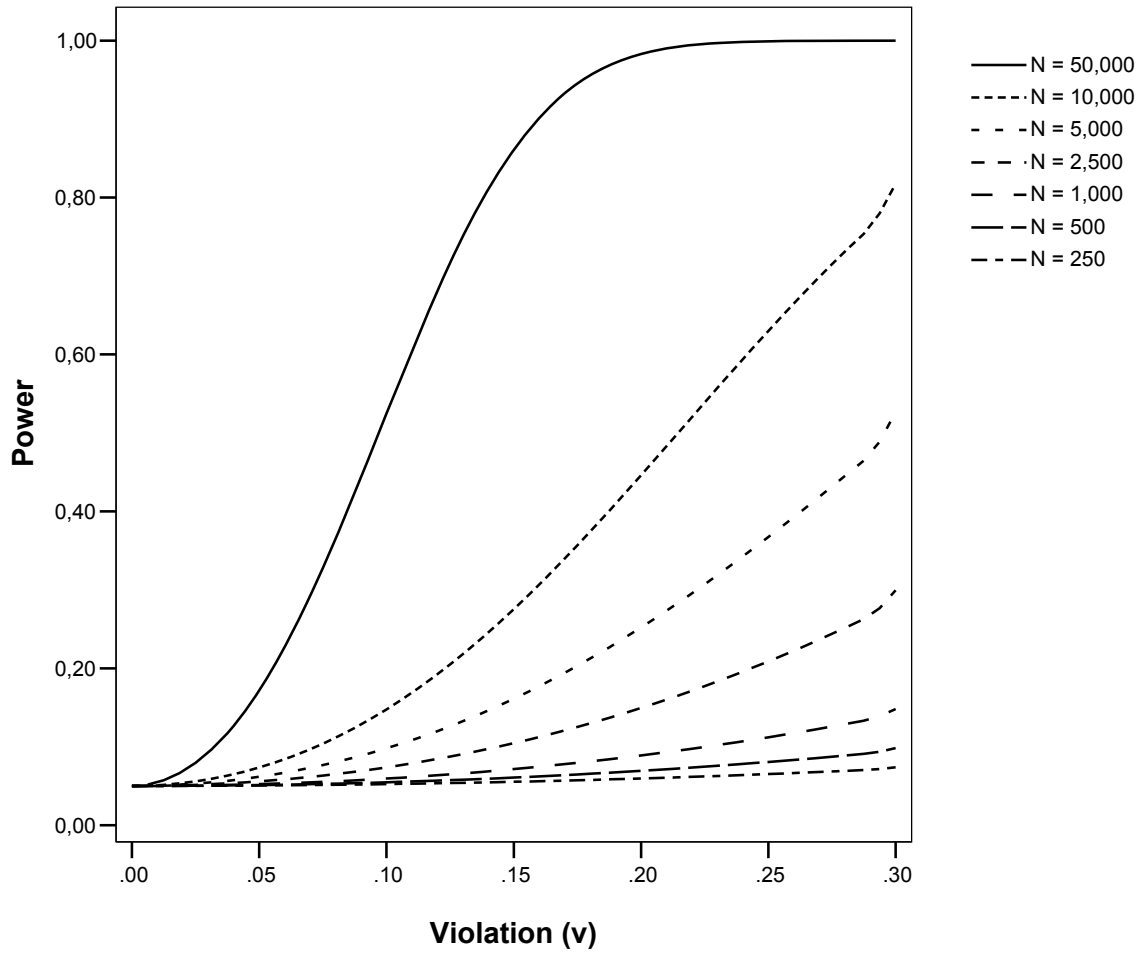
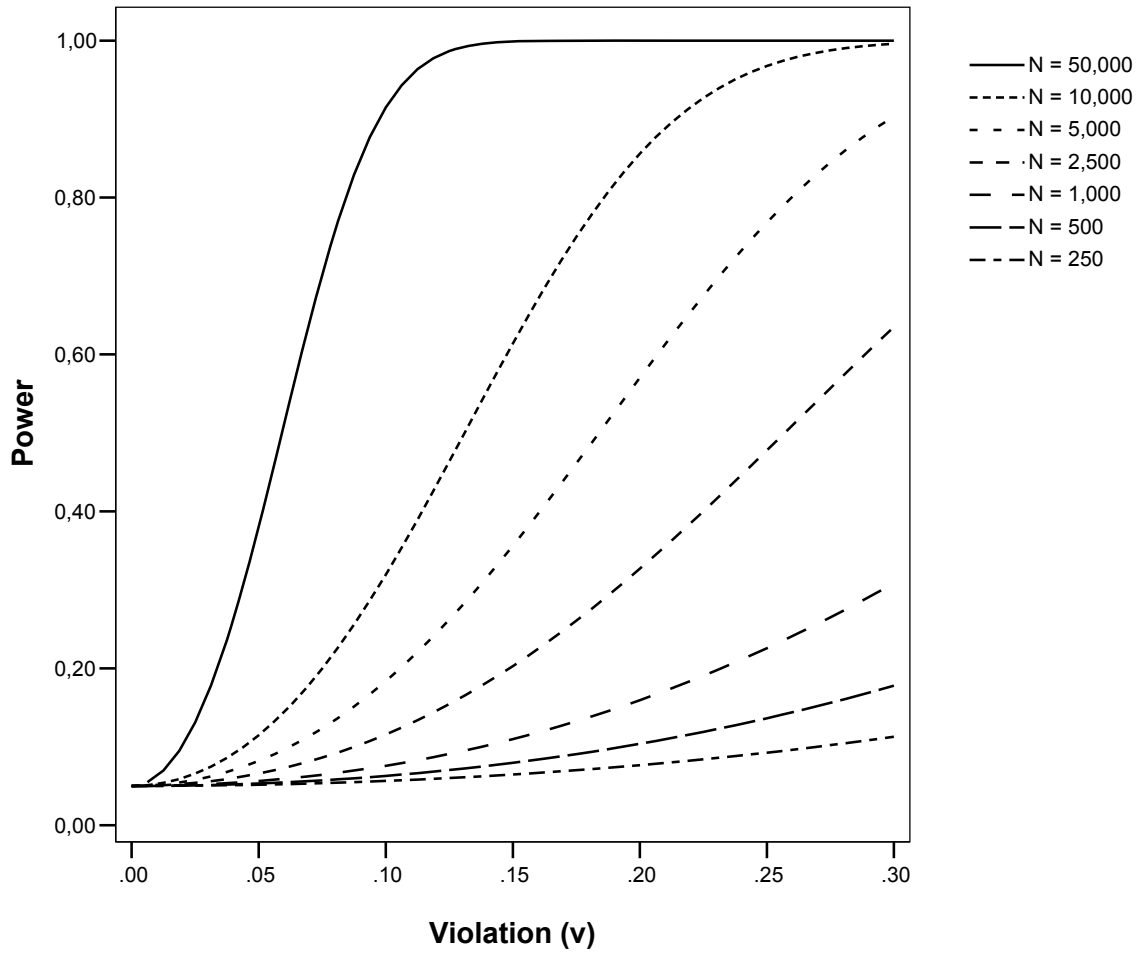

1,00
0,80
0,60
0,40
0,20
0,00
Power
.00
.05
.10
.15
.20
.25
.30
Violation (v)
N = 50,000
N = 10,000
N = 5,000
N = 2,500
N = 1,000
N = 500
N = 250

Figure 6


1,00
0,80
0,60
0,40
0,20
0,00
Power
.00
.05
.10
.15
.20
.25
.30
Violation (v)
N = 50,000
N = 10,000
N = 5,000
N = 2,500
N = 1,000
N = 500
N = 250

Running head: *Randomized response technique*

# Surveying multiple sensitive attitudes using a cheating detection extension of the randomized response technique

Morten Moshagen and Jochen Musch

University of Duesseldorf, Germany

Correspondence should be addressed to:

Morten Moshagen

Institute for Experimental Psychology

University of Duesseldorf

Universitaetsstr. 1

40225 Duesseldorf

Germany

Phone: +49 211 81 13494

Fax: +49 211 81 11753

Email: morten.moshagen@uni-duesseldorf.de

Abstract

The tendency to provide socially desirable responses to sensitive questions is a well known problem in attitude measurement. The randomized response technique (RRT) protects the privacy of respondents by adding random noise to their responses, such that there is no direct link between an individual's response and his or her true attitude. Although the RRT has repeatedly been shown to improve the validity of self-reports of behaviours, there are two reasons why the technique has rarely been used in research on attitudes. First, since the RRT was originally designed for posing single sensitive questions, it is difficult to survey attitudes on multiple issues simultaneously. Second, most RRT models do not take the problem of non-compliance to the RRT instructions into account. We describe a modification of the RRT that is capable of surveying multiple attitudes using a single randomization process, while at the same time controlling for non-compliance to the instructions. An empirical application demonstrates the superiority of this multiple issues cheating detection (MICD) model over both, simple direct questioning and the forced response variant of the RRT which does not take cheating into account. We recommend that researchers should routinely consider using the MICD variant of the RRT to obtain more valid prevalence estimates for sensitive attitudes.

# Surveying multiple sensitive attitudes using a cheating detection extension of the randomized response technique

Questionnaires and interviews are frequently used in social psychology to measure attitudes on sensitive topics. It is well known, however, that survey data do not necessarily reflect the respondents' actual opinion. The tendency to present themselves favourably systematically biases their responses to sensitive, incriminating, or illegal issues in the direction of their perception of what may be socially acceptable (Lee, 1993). As a consequence, self-report measures consistently underestimate the prevalence of socially undesirable attitudes, and overestimate the prevalence of socially desirable attitudes (Tourangou & Yan, 2007). Several methods have been proposed to reduce answering bias in self-reports of sensitive attitudes, including the bogus pipeline procedure (Jones & Sigall, 1971), implicit attitude measurement (Greenwald, McGhee, & Schwartz, 1998), and the use of scales measuring individual differences in the tendency to provide socially desirable responses (e.g., Paulhus, 1984).

Providing confidentiality and anonymity is a more simple but perhaps the most promising way to encourage truthful and honest responding. It has repeatedly been shown that providing questionnaires anonymously enhances the validity of responses as compared to more public modes of administration, such as face-to-face interviews. However, even in presumably anonymous surveys, participants may still fear that their responses may become known to the researcher, and may therefore decide to play safe and mask their true attitude by providing a response that is socially acceptable. In an attempt to maximize and guarantee the anonymity and confidentiality of responses, Warner (1965) therefore proposed the randomized response technique which adds random noise to the responses. Confidentiality is thus increased, and an individual's true attitude can no longer be determined on the basis of his or her response.

An example is helpful to illustrate how the RRT works. In the forced response variant of the RRT (Dawes & Moore, 1980), which is one of the most efficient RRT procedure available (Lensvelt-Mulders, Hox, & van der Heijden, 2005), each participant is confronted with the sensitive question. However, a randomization device is used to determine whether the participant is asked to respond truthfully, or whether he or she is prompted to provide a pre-specified response (e.g., "yes") irrespective of his or her true attitude. Because the outcome of the randomization process is known only to the participant, the investigator cannot determine whether a "yes"-response resulted from truthful answering, or from the randomization process. The proportion of "yes"-responses that have not been prompted by the randomization procedure can however be estimated by straightforward probability calculations based on the known probability distribution of the randomization device. For example, the participant's month of birth, unknown to the experimenter, may be used to determine whether participants are prompted to respond truthfully to a sensitive question (e.g., "Do you feel uneasy in the presence of people with disabilities?"; Ostapczuk & Musch, 2008). Depending on their month of birth, some participants are then asked to respond truthfully to this question, whereas others are prompted to reply "yes" irrespective of their true attitude. Using official birth statistics to determine the probability distribution of the randomization variable, it is possible to estimate the proportion of non-forced "yes"-responses (Musch, Bröder, & Klauer, 2001). Since the randomization procedure guarantees that a "yes"-response can no longer be unequivocally associated with a socially undesirable attribute and is therefore no longer stigmatizing, the RRT encourages more honest responding. Validation studies have repeatedly shown that this procedure results in more valid estimates of the prevalence of sensitive behaviours (for a recent review, see Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). Given the considerable number of successful applications in surveys of sensitive behaviour, it is rather surprising that the RRT has rarely been used in attitude

research (Antonak & Livneh, 1995). We argue that there are two likely reasons for this surprising lack of applications of the RRT to the measurement of attitudes.

First, it is often desirable to measure attitudes not only towards a single issue, but towards several issues simultaneously. The RRT, however, was originally developed for posing single questions based on the outcome of a single randomization process. In order to maintain the privacy protection offered by the RRT even when asking several questions, multiple randomization processes are needed (Tamhane, 1981). The necessity to use several randomization devices, or to use the same randomization device repeatedly to gather responses to more than one sensitive question, renders the administration of the RRT rather tedious and troublesome, however. Second, the RRT has been criticized as being susceptible to cheaters, that is, to respondents who do not comply with the instructions and deny replying as prompted by the outcome of the randomization device (Campbell, 1987). In the remainder of this paper, we will show how multiple sensitive attitudes may be assessed with just a single randomization process. The procedure we are proposing also allows us to estimate the proportion of non-compliant respondents for each question, and thus enables attitude researchers to compute an upper bound for the prevalence of several sensitive attitudes simultaneously.

## A Cheating Detection Extension of the RRT

A major concern with traditional randomized response models is that some participants may choose not to comply with the instructions by denying to reply as directed by the randomization device (Campbell, 1987). There is indeed evidence suggesting that such cheating occurs (Edgell, Duchan, & Himmelfarb, 1992; Lensvelt-Mulders & Boeije, 2007; Soeken & Macready, 1982). The RRT underestimates the prevalence of socially undesirable

attitudes to the extent that participants fail to comply with the instructions and reply “no” despite being prompted by the randomization device to answer in the affirmative regardless of question content.

## insert figure 1 ###

However, advanced RRT models allow for estimating the extent of non-compliance to the RRT instructions. In what we will refer to as the cheating-detection model (Clark & Desharnais, 1998), it is explicitly assumed that a certain proportion of respondents may fail to comply with the instructions. Consequently, the cheating detection model shown in Figure 1 divides the population into three disjoint and exhaustive groups: The first group ($\pi$) represents the proportion of compliant and honest “yes”-respondents, that is, respondents who truthfully admit having a socially undesirable attitude. The second group ($\beta$) is the proportion of compliant and honest “no”-respondents, that is, respondents who truthfully deny having a socially undesirable attitude. The third group ($\gamma = 1 - \pi - \beta$) represents the proportion of non-compliant cheaters who do not conform to the instructions by denying to adopt a socially undesirable attitude irrespective of the outcome of the randomization process. It is important to note that nothing is, nor can be assumed about the true attitude of the non-compliant respondents. Conceivably, some respondents may deny the undesirable attitude although they are actually sharing it just to avoid to identify themselves as carriers of a socially undesirable attribute; however, it is also possible that respondents not sharing an undesirable attitude are deciding to play safe by answering “no” in an attempt to rule out even the slightest suspicion that they are holding an objectionable attitude. Thus, the true attitude of a respondent choosing not to follow the instructions necessarily remains unknown. As the three proportions $\pi$, $\beta$, and $\gamma$ add up to 1, the model contains two independent parameters which cannot be estimated on the basis of the one proportion of “yes"-responses that is provided by traditional

randomized response models. In order to obtain a sufficient data base, two independent samples of respondents with different randomization probabilities $p_1$ and $p_2$ have to be drawn (Clark & Desharnais, 1998). Figure 1 shows only the condition in which probability $p_1$ applies; the second condition could be represented by an identical figure with the sole exception that $p_1$ would be replaced with $p_2$. Assuming that $\pi$, $\beta$, and $\gamma$ are equal across groups when participants are randomly assigned to conditions, the two independent proportions of "yes"-responses are sufficient to estimate the two independent parameters $\pi$ and $\beta$ (with $\gamma = 1 - \pi - \beta$).

The cheating detection model offers a unique advantage over both, traditional direct questioning formats and previous randomized response models: If no cheating occurs, an assumption which can be tested within the model, the parameter $\pi$ provides an asymptotically unbiased estimate of the prevalence of the socially undesired attitude. To the best of our knowledge, there is no other technique that is capable of providing such an estimate. The model also allows us to estimate the proportion of non-compliant respondents. If this proportion differs significantly from zero, it should not be routinely assumed that of all them are actually carriers of the unwanted attribute (Clark & Desharnais, 1998); it is, however, possible to compute a lower and upper bound for the prevalence of an objectionable attitude by alternatively assuming that all non-compliant respondents either do, or do not hold the sensitive attitude (Musch et al., 2001).

## Using the RRT to Assess Multiple Attitudes

A second reason for the dearth of RRT studies in attitude measurement is the fact that multiple randomization processes are required to maintain the privacy protection offered by the RRT whenever a researcher wants to assess multiple attitudes within a single study

(Tamhane, 1981). The major strength of the RRT is that there is no direct link between a participant's attitude and the response he or she provides. If one were to use the outcome of a single randomization process repeatedly for posing multiple questions, a situation would arise in which some response patterns can only result from a particular outcome of the randomization device. It would then be possible to directly infer from a participant's response pattern whether he or she responded truthfully, or just responded affirmatively because being prompted by the randomization device. For example, consider an RRT study comprising three different questions using a randomization device with 12 possible outcomes such as the participants' month of birth (e.g, Ostapczuk, Moshagen, Zhao, & Musch, 2008). Participants born in January, February, or March might be prompted to respond affirmatively to these questions, whereas participants born in another month are asked to respond truthfully to each of the three questions (i.e., the randomization probability is $p = 3/12$). This randomization procedure would ask participants born in either January, February, or March to reply "yes" to each question. Therefore, if a participant denies at least one of these questions, it would become obvious that he or she was prompted to reply truthfully to each of the questions. Each "yes"-answer to one of the other questions would therefore expose his or her true attitude regarding this question. In such a situation, the RRT effectively offers no more privacy protection than a traditional direct questioning format. For this reason, the fixed outcome of a randomization process should not be used repeatedly for posing multiple questions in a RRT survey.

An obvious solution for this problem is to use a randomization device different from the participant's birth of month (e.g., a die), and to reinitialize the randomization process for each single question in the survey (e.g., Himmelfarb & Lickteig, 1982). However, this approach is not very attractive because it relies on the troublesome procedure of determining a new randomization outcome for each single question, for example by using a spinner over and

over again. Instead, we are proposing a method that relies on the use of just a single randomization process, but employs a carefully designed answering scheme that relates the participant's answers to the outcome of the randomization device in a different way for each question. This answering scheme makes sure that it is impossible to infer a respondent's true attitude from his or her response, even without a need to determine a randomization outcome for each question anew. The basic idea of this method is to take care that each possible response pattern (the vector of all yes/no-answers to all questions asked) may be the result of the randomization procedure entailed in the answering scheme. As an example, consider again a study using the participant's month of birth as a randomization device. Table 1 shows how the twelve possible outcomes of the randomization device (that is, January to December) can be used to simultaneously pose three questions such that each possible response may be the result of the randomization process. As a result, the privacy of the respondents is preserved. For example, a respondent may answer in the affirmative to the first two items, but may answer "no" to the third item. The response vector (Question 1: "yes", Question 2: "yes", Question 3: "no") of this respondent does not reveal anything about his or her true status with regard to the three sensitive questions, if the RRT questions follow the scheme shown in Table 1. According to this scheme, respondents born in January, February, April or May are prompted to provide the prespecified response "yes" to the first question; otherwise, they are asked to answer truthfully. With regard to the second question, participants are prompted to answer "yes" regardless of the question content if they are born in January, February, March, or June; otherwise, they are again asked to answer truthfully. Finally, with regard to the third question, participants are prompted to answer "yes" if they are born in January, March, April, or July, and to respond truthfully otherwise. Employing this answering scheme, there is no response pattern that allows inferring the true status of a respondent with regard to any of the questions. This is because each of the seven possible answering patterns to the three questions ("Yes/Yes/Yes"; "Yes/Yes/No"; "No/Yes/Yes"; "Yes/No/Yes"; "Yes/No/No"; "No/Yes/No";

and “No/No/Yes”) might be due to the respondent being born in a month that required him or her to respond with this particular pattern (that is, in either January, February, March, April, May, June, or July, respectively). Thus, the true status of a respondent with regard to the critical questions can never be determined on the basis of his or her responses. The only answer pattern that is not considered in Table 1 is the case of a respondent who is denying every question (“No/No/No”). However, this particular response pattern may be safely ignored anyway, as there is no reason for a respondent to feel embarrassed when all of his or her responses are in agreement with social norms.

To summarize, the procedure illustrated in Table 1 maintains the privacy protection offered by the randomization procedure and simultaneously allows to ask multiple questions using only a single randomization procedure. In the case of the present example, this method allows to pose three randomized questions using only one random variable, the participant’s month of birth.

## Please insert table 1 about here ##

The proposed method can be further enhanced to detect cheaters - that is, respondents who disobey the instructions - by sampling two groups for each question. In the first group, a randomization probability of $\underline{p_1}$ (for example, 4/12 if the four months of January, February, March and April are used to determine the outcome of the randomization process) has to be used, whereas in the second group, a randomization probability of $\underline{p_2} = 1 - \underline{p_1}$ is used by inverting the set of months that determines the outcome of the randomization process. Respondents in the first group may then be asked to answer “yes” regardless of the question content if they were born in either January, February, March, or April, and to respond truthfully otherwise; whereas respondents in the second group may be asked to answer “yes”

if they were born in neither January, February, March, nor April, and to respond truthfully otherwise. Thus, it is possible to pose three randomized questions using only one random variable, while simultaneously determining the proportion of cheaters for each of the three questions.

In what follows, we compare the performance of this multiple issues cheating detection (MICD) model against the traditional forced-response RRT model not considering cheating. Additionally, we include a direct questioning control condition to obtain an estimate of the extent to which the above two variants of the RRT are capable of reducing response bias.

## Methods

### Participants

One-thousand three-hundred and thirty-nine volunteers (515 female) participated in the study. Mean age was 29.58 years (SD=10.44). Participants were randomly assigned to one of three groups by a ratio of 2:2:1 to compensate for the loss of efficiency in parameter estimation associated with the use of the randomization procedure. The three groups employed the RRT with randomization probability $p_1$ ($N$ = 538), the RRT with randomization probability $p_2$ ($N$ = 515), or a direct questioning procedure ($N$ = 286). The two RRT conditions with different randomization probabilities $p_1$ and $p_2$ are needed in order to obtain a sufficient data base to estimate the two independent parameters $\pi$ and $\beta$ (with $\gamma = 1 - \pi - \beta$) included in the MICD model.

### Measures and Procedures

After completing demographic background information, participants received three sensitive questions relating to socially desirable attitudes in randomized order: (1) “I am of

the opinion that the Federal Republic of Germany should grant asylum to each civil war refugee"; (2) "I am of the opinion that except for their sexual preference, homosexuals are no different from other people"; (3) "I am willing to pay an additional amount of 100 € to obtain electricity from renewable sources". A pretest with 69 respondents confirmed that it is considered socially undesirable to deny these questions. Mean ratings on a nine-point scale ranging from (1) "It is socially desired to deny this question" to (9) "It is socially desired to answer this question affirmatively" for the three questions were (1) M = 6.70 (SD = 1.13), (2) M = 7.55 (SD = 1.21), and (3) M = 5.39 (SD = 1.10), respectively. The social desirability ratings for the questions differed significantly from each other: (1) vs. (2): $t(68)=3.98$, $p<.01$; (1) vs. (3): $t(68)=6.76$, $p<.01$; (2) vs. (3): $t(68)=10.48$, $p<.01$.

In the direct questioning control condition, participants were simply asked to answer the sensitive questions truthfully. In the two remaining conditions, instructions were given in RRT format, using the participants' month of birth as a randomization device according to the answering scheme shown in Table 1. Importantly, in the present study, the randomization procedure prompted participants for a "no"-answer, because it was socially desirable to respond to the critical attitude questions in the affirmative. Therefore, the anonymity of undesirable "no"-answers had to be protected by enforcing other participants to answer "no" via the randomization process. The usual rationale of the RRT, which asks participants to provide "yes"-answers in response to questions asking for socially undesirable behaviours, was thus inverted for the purpose of the present study. Consequently, the instructions given in the RRT high-probability condition for the first question read: "If you were born in January, February, April, or May, please reply 'no' to the following question independently of its content. If you were born in another month, please answer truthfully." The instructions given in the RRT low probability condition for the first question were: "If you were born in January, February, April, or May, please answer truthfully. If you were born in another month, please reply 'no' to the following question independently of its content." According to exact birth

statistics provided by the German Federal Agency for Statistics, the probability of being prompted to reply "no" to the first question in the high and low probability condition approximated $\underline{p_1}$ = . 34 and $\underline{p_2}$ = 1 - $\underline{p_1}$ = .66, respectively. The randomization probabilities for the second question approximated $\underline{p_1}$ = .34 and $\underline{p_2}$ = 1 - $\underline{p_1}$ = .66, and for the third question, $\underline{p_1}$ = .35 and $\underline{p_2}$ = 1 - $\underline{p_1}$ = .65. Detailed instructions explained how this randomization procedure guaranteed the confidentiality of responses, and that privacy protection was maintained in spite of the use of only a single randomization device due to the answering scheme governing the randomization process.

Statistical Analysis

Both the MICD and the traditional forced response model not considering cheating can be subsumed under the more general multinomial modelling framework (Batchelder & Riefer, 1999). Reparameterizing the model shown in Figure 1 into a statistically equivalent binary tree model (Moshagen et al., 2008) allowed us to utilize the EM-algorithm (Hu & Batchelder, 1994) implemented in the software program HMMTree (Stahl & Klauer, 2007) to compute maximum-likelihood estimates for the parameters $\pi$, $\beta$, and $\gamma$. The traditional forced response variant of the RRT is simply a restricted version of the MICD, constrained by the assumption that the proportion of cheaters equals zero ($\gamma$=0). The unconstrained MICD is saturated with zero degrees of freedom, as the two independent proportions of "yes" responses are just sufficient to estimate the two independent parameters $\pi$ and $\beta$ (with $\gamma = 1 - \pi - \beta$). In the following, we judged the justifiability of parameter restrictions by evaluating how much model fit worsened when imposing the respective restriction. The significance of a reduction in the fit of the model was assessed using the asymptotically $\chi^2$-distributed log-likelihood-ratio statistic $\Delta\underline{G^2}$ with one degree of freedom.

## please insert Table 2 here ##

## Results

As shown in Table 2, only 36.01 % (SE = 2.84) of the respondents in the DQ condition expressed the socially undesirable opinion that the Federal Republic of Germany should not grant asylum to each civil war refugee. When the MICD was employed, 49.19 % (SE = 4.99) admitted to share this attitude. Constraining the estimates of the DQ and the RRT conditions to be equal resulted in a significant loss of fit of the model [$\Delta G^2(\Delta df=1) = 4.95$, $p < .05$], indicating a significantly higher prevalence estimate using RRT, as compared to DQ. Employing the MICD also revealed that there was a significant proportion of non-compliant respondents who disobeyed the instructions by denying to adopt a socially undesirable attitude ($\gamma = 18.73$ %; SE = 4.86). This proportion of cheating respondents was not negligible and significantly higher than zero, $\Delta G^2(\Delta df=1)= 15.93$, $p <.01$. It is therefore inappropriate to use the traditional forced-response model not considering cheating. However, if the traditional forced variant had nevertheless been employed, the proportion of participants sharing the sensitive attitude would have been estimated at only 32.23 % (SE = 2.79), which is an even slightly – though not significantly – lower estimate than in the DQ condition [$\Delta G^2(\Delta df=1)= 0.93$, ns]. Thus, the traditional forced response RRT would have underestimated the proportion of respondents adopting a socially undesirable attitude. As explained above, no assumption can be made about the true status of the cheating participants. Alternately assuming in a worst- and best-case scenario that either none or all of the cheating participants actually adopted the sensitive attitude, we computed a lower and upper bound prevalence estimate of $\pi = 49.19$ % and $\pi + \gamma = 67.92$ %, respectively, for the first question.

A similar pattern of results emerged considering the second question ("I am of the opinion that except for their sexual preference homosexuals are no different from other

people"). Under DQ conditions, only 12.59 % (SE = 1.96) admitted holding a homophobic attitude. The MICD, however, estimated the proportion of homophobes at 30.48 % (SE = 5.26), which is significantly higher than the DQ-based estimate [$\Delta G^2(\Delta df=1) = 10.16$, $p < .01$]. Furthermore, significant non-compliance to the instructions occurred with the proportion of cheating participants estimated at 19.99% [SE = 5.23; $\Delta G^2(\Delta df=1) = 15.39$, $p < .01$]. Thus, the lower bound for the proportion of homophobes is $\pi = 30.48\%$, and the upper bound is $\pi + \gamma = 50.47$ %. Given that significant non-compliance to the RRT rules occurred, applying the traditional forced response variant not considering cheating may be misleading. Accordingly, the traditional forced response variant provides a prevalence estimate of $\pi = 13.31$ % (SE = 2.86), which is not significantly different from the DQ estimate [$\Delta G^2(\Delta df=1) = 0.45$, ns].

When questioned directly, 38.11 % (SE = 2.87) of respondents stated that they would not pay an additional annual fee to obtain electricity from renewable sources. Employing the MICD resulted in an estimate of 44.02 % (SE = 4.79); however, the difference between DQ-based estimate and RRT-based estimate failed to reach statistical significance [$\Delta G^2(\Delta df=1) = 0.40$, ns]. Moreover, there was a marginal proportion of cheating participants ($\gamma = 5.73$ %; SE = 4.46), which did not differ significantly from zero [$\Delta G^2(\Delta df=1) = 1.70$, ns]. Consequently, employing the traditional forced response model not considering cheating resulted in virtually identical prevalence estimates.

## Discussion

Although the superiority of the RRT over more traditional data collection techniques has been demonstrated repeatedly, the RRT has been rarely used in attitude research. Possible reasons for this lack of applications include the problem of non-compliance to the RRT instructions and the practical difficulties arising when multiple attitudes have to be assessed in a single session. In the present study, we demonstrated how to assess multiple sensitive attitudes

simultaneously using just a single randomization process, while at the same time employing Clark and Desharnais' (1998) cheating detection modification of the RRT to identify the proportion of respondents not following the instructions.

The results demonstrate that the multiple issues cheating detection (MICD) model allowed for obtaining higher, and thus presumably more valid, prevalence estimates for the three sensitive attitudes than both, the forced-response variant of the RRT not considering cheating and a traditional direct questioning format. The results also suggest that the forced response variant not considering cheating results in misleading prevalence estimates if a substantial proportion of the respondents fail to comply to the instructions and deny to respond to the sensitive question as directed by the randomization device. It is also interesting to note that the discrepancy in the estimates obtained by direct questioning and by employing the MICD increased with increasing sensitivity of the attitude under consideration. For the least sensitive question on renewable energy, the DQ estimate of 38.11% was quite close to the randomized response estimate of 44.02 %. Moreover, virtually no cheating was observed for this question ($\gamma = 5.73$ %). Regarding the most sensitive question on homophobia, however, the MICD estimates the prevalence of socially undesirable negative attitudes towards homosexuals at 30.48%, a significantly higher estimate than in the DQ condition (12.59%). Furthermore, a substantial amount of non-compliance to instructions was observed for this question ($\gamma = 19.99\%$). Assuming in a worst-case scenario that all of these non-compliant cheaters are in fact adopting a homophobic attitude, it is possible to compute an upper bound estimate of 30.48% + 19.99% = 50.47 % for the prevalence of this most socially undesirable attitude. This pattern of results suggests that the RRT is to be recommended only for sensitive issues that are severely threatened by socially desirable responding; with regard to the question on renewable energy, the small difference between the DQ and RRT estimate did not justify the additional cost (in terms of efficiency) that is associated with the use of the

RRT. Arguably, there are many socially acceptable reasons (e.g., financial constraints) for being unwilling or unable to pay more in order to obtain electricity from renewable sources.

Some limitations should be considered when interpreting the results of the present study, however. First, the proposed answering scheme is applicable only for use with up to three sensitive questions. This is because in the case of two response categories ("yes" and "no") and three different questions, there are only $2^3 = 8$ possible response patterns, whereas in the case of 4 questions, there are $2^4 = 16$ different response patterns, a figure that surpasses the number of months per year. Of course, to extend the present method to more than three sensitive questions, it is easily possible to use a randomization device different from the participant's month of birth yielding the required number of outcomes (e.g., the participant's day of birth, or another randomization device that provides a sufficiently large number of possible outcomes). Second, the RRT provides estimates at group level only, without revealing the true attitude of any individual. However, it is exactly this feature that enhances the confidentiality of responses and encourages the respondents to answer more honestly to a RRT than to a direct question. Finally, owing to the randomization employed, the RRT necessarily suffers a loss of efficiency as compared to a direct question. This loss of efficiency is only outweighed by a gain in precision if an attitude under question is of a sufficiently sensitive nature.

The above limitations notwithstanding, the present results show that the cheating detection model shows considerable promise as a means to measure attitudes threatened by socially desirable responding. Employing the answering scheme proposed here, it is also easily possible to survey more than one attitude in a single study without the need for the repeated use of a randomization device. Given the considerable discrepancy between the results obtained by direct questioning and by employing the cheating detection model, we

recommend that researchers should routinely consider using the cheating detection variant of the RRT in future surveys on sensitive attitudes.

References

Antonak, R. F., & Livneh, H. (1995). Randomized response technique: A review and proposed extension to disability attitude research. *Genetic, Social & General Psychology Monographs*, *121*, 97-145.

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

Campbell, 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., & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen [Traditional Guttman-scaling and randomized response]. In F. Petermann (Ed.), *Einstellungsmessung - Einstellungsforschung [Attitude measurement]* (pp. 117-133). Göttingen: Hogrefe.

Edgell, S. E., Duchan, K. L., & Himmelfarb, S. (1992). An empirical test of the unrelated question randomized response technique. *Bulletin of the Psychonomic Society*, *30*, 153-156.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.

Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology*, *43*, 710-717.

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

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

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

Moshagen, M., Ostapczuk, M., Musch, J., Mischke, R., Bröder, A., & Erdfelder, E. (2008). Making compliance testable: How to improve cheating detection in the randomized response technique. *Manuscript submitted for publication.*

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., Moshagen, M., Zhao, Z., & Musch, J. (2008). Assessing sensitive attributes using the randomized-response technique: Evidence for the importance of response symmetry. *Manuscript submitted for publication.*

Ostapczuk, M., & Musch, J. (2008). Projective questioning overestimates the prevalence of negative attitudes towards people with physical and mental disabilities. *Manuscript submitted for publication.*

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598-609.

Soeken, K. L., & Macready, G. B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, *92*, 487-489.

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

Tamhane, A. (1981). Randomized response techniques for multiple sensitive attributes. *Journal of the American Statistical Association*, *76*, 916-923.

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

Umesh, U., & Peterson, R. (1991). A critical evaluation of the randomized response method: Applications, validation and research agenda. *Sociological Methods & Research*, *20*, 104-138.

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

## Author Note

Morten Moshagen and Jochen Musch, University of Duesseldorf, Germany. This work was partly supported by a grant of the German Research Foundation (DFG, Mu 2674/1-1). The authors would like to thank Martin Ostapczuk and Jennifer Nicolai for their helpful comments on an earlier version of this article. Correspondence concerning this article should be addressed to Morten Moshagen (E-Mail: morten.moshagen@uni-duesseldorf.de) or Jochen Musch (E-Mail: jochen.musch@uni-duesseldorf.de), Institute for Experimental Psychology, University of Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany.

Tables

Table 1

Distribution of the outcomes of the participants' month of birth as a randomization device on three different questions

| | Month of birth | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sept | Oct | Nov | Dec |
| #1 | x | x | | x | x | | | | | | | |
| #2 | x | x | x | | | x | | | | | | |
| #3 | x | | x | x | | | x | | | | | |

Note. Participants born in a month marked with an 'x' are prompted to provide a prespecified response (e.g., "yes") to the particular item, the remaining participants are asked to answer truthfully. Each possible response pattern (except for denying every item) thus may be the result from either the randomization procedure or from truthful responding.

Table 2

Parameter estimates

| | | Question 1: Political Asylum | Question 2: Homosexuals | Question 3: Renewable energy |
|---|---|---|---|---|
| DQ | % "No" | 36.01** | 12.59** | 38.11** |
| RRT | Honest "No" ($\pi$) | 49.19** | 30.48** | 44.02** |
| | Honest "Yes" ($\beta$) | 32.08** | 49.53** | 50.25** |
| | Cheaters ($\gamma$) | 18.73** | 19.99** | 5.73 |

Note. DQ = direct questioning; RRT = randomized response. Question 1: "I am of the opinion that the Federal Republic of Germany should grant asylum to each civil war refugee"; Question 2: "I am of the opinion that except for their sexual preference, homosexuals are no different from other people"; Question 3: "I am willing to pay an additional amount of 100 € to obtain electricity from renewable sources".

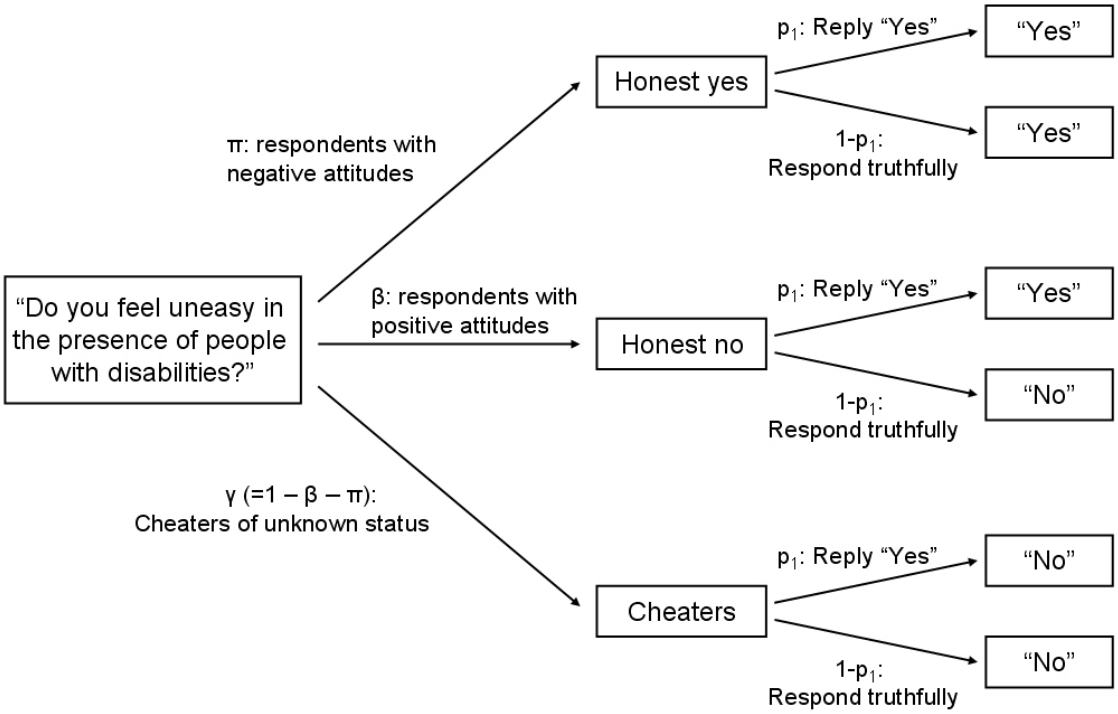
** $p < .01$ for the test that an estimated proportion is equal to zero.

Figure captions

Figure 1:

A multinomial representation of Clark & Desharnais' (1998) cheating detection modification of the randomized response technique.

Figure 1


"Do you feel uneasy in the presence of people with disabilities?"
π: respondents with negative attitudes
β: respondents with positive attitudes
γ (=1 – β – π): Cheaters of unknown status
Honest yes
Honest no
Cheaters
p1: Reply "Yes"
1-p1: Respond truthfully
"Yes"
"No"

Running Head: Randomized response technique

# A note on untruthful responding in randomized response surveys

Morten Moshagen & Jochen Musch

University of Duesseldorf, Germany

Correspondence should be addressed to:

Morten Moshagen

Institute for Experimental Psychology

Heinrich-Heine University

40225 Duesseldorf

Germany

Phone: +49-211-8113494

Fax: +49-211-8111753

Email: morten.moshagen@uni-duesseldorf.de

Abstract

In a randomized response survey (Warner, 1965), the participant's privacy is protected by adding random noise to their responses to prevent any direct link between their answers and their true status with regard to a sensitive attribute. Randomized response surveys provide distorted estimates, however, if some respondents do not answer truthfully in spite of the privacy that is guaranteed by this randomization process. We propose a modification of Mangat's (1994) variant of the randomized technique that allows to estimate the proportion of participants who are carriers of the sensitive attribute, but fail to respond truthfully.

Keywords: Randomized response technique, multinomial model, social desirability, untruthful responding

A note on untruthful responding in randomized response surveys

The randomized response technique (RRT; Warner, 1965) was developed as a means to reduce social desirability bias in surveys of sensitive behaviour. The rationale of the RRT is to add random noise to the participant's answers to ensure that an individual's true status cannot be determined on the basis of his or her response to a critical question. Although the randomization process protects the true status of any individual participating in the survey, elementary probability calculations allow us to compute estimates of the prevalence of a critical attribute at group level. In the original Warner (1965) model, respondents are to answer either the sensitive question (e.g., "Have you ever used cocaine?") with probability $p$, or the negation of the sensitive question ("Have you never used cocaine?") with probability $1-p$. Thus, the observed proportion of 'yes' responses is $\lambda = \pi p + (1-\pi)(1-p)$. It is easy to see that the proportion of respondents carrying the sensitive attribute ($\pi$) can be estimated by $\pi = [\hat{\lambda} + (p-1)]/(2p-1)$.

Since the seminal work of Warner (1965), several variants of the RRT have been proposed, including the unrelated question model (Bourke, 1984; Greenberg, Abul-Ela, Simmons, & Horvitz, 1969; Moors, 1971), forced response models (Boruch, 1971; Dawes & Moore, 1980), and random binary event models (Kuk, 1990). In an attempt to reduce the sampling variance, Mangat (1994) recommended an optimization of a model previously proposed by Mangat and Singh (1990). In Mangat's (1994) procedure, respondents actually carrying the sensitive attribute are requested to respond truthfully. Respondents who do not carry the sensitive attribute receive the sensitive question in the format proposed by Warner (1965) as explained above. Given that an individual "yes" response may then stem from either respondents that carry, or from respondents that do not carry the sensitive attribute, and given that the researcher cannot tell which of these two possibilities applies, the confidentiality of respondents answering in the affirmative is protected. Because only the "yes"-responses are

distorted and all “no” responses are true ”no’s”, the estimation of the proportion having the sensitive attribute is simplified, in turn resulting in a smaller variance of the estimates.

Although randomized response models have been repeatedly shown to yield more valid prevalence estimates of sensitive attributes (for a recent meta-analysis, see Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005), they have been criticized as being susceptible to respondents that fail to comply with the instructions by denying to reply as directed by the randomization device (Campbell, 1987). In the present paper, we address this concern by proposing a modification of Mangat’s (1994) procedure that allows for estimating the proportion of participants that fail to respond truthfully.

## The Proposed Method

We conceptually divide the population into three disjoint and exhaustive groups: Respondents who are carrying the sensitive attribute and reply truthfully ($\pi$), respondents who are carrying the sensitive attribute but fail to reply truthfully ($\gamma$), and respondents who are not carriers of the sensitive attribute and reply truthfully by denying it ($\beta$). By symmetry, it can be argued that there may also be respondents who are not carriers of the sensitive attribute but for some reason, nevertheless attest to it. Given that this is equivalent to associate oneself with a socially undesirable or prohibited behaviour without a need and without a reason, it can be argued that such a voluntary self-incriminating behaviour should be extremely rare. It is therefore assumed in the model that it can safely be ignored. In Mangat’s (1994) procedure, respondents belonging to $\pi$ would reply “yes”, respondents belonging to $\gamma$ would reply “no”, and respondents belonging to $\beta$ would reply “yes” if they received the negation of the sensitive question with probability $(1-p)$, and “no” otherwise. Thus, the proportion of “yes” responses in the $i$-th sample is $\lambda_i = \pi + \beta(1-p_i)$.

In order to estimate the two independent parameters (since $\pi + \beta + \gamma = 1$), it is necessary to draw two independent non-overlapping random samples, where the probability to

answer the sensitive question ( $p_i$ ) must differ across samples. Parameter estimates can then be obtained by utilizing the general family of multinomial processing tree models (Batchelder & Riefer, 1999; Hu & Batchelder, 1994). This modelling approach has several benefits, the main one being increased flexibility. For instance, it is easily possible to augment the proposed model with additional parameters representing, for example, different subgroups (e.g., males and females) for which parameters should be estimated separately. Since the definition of general processing tree models only covers the case of binary tree models, three different models need to be estimated in order to obtain unconditional estimates of the parameters of interest $\pi$, $\beta$, and $\gamma$ along with appropriate standard errors. Using a sensitive question on cocaine use as an example, the multinomial models of our proposed procedure are shown in Figures 1-3, where $\mu$, $\nu$, and $\xi$ are helper parameters with $\mu = \beta/(\beta + \gamma)$, $\nu = \pi/(\pi + \gamma)$, and $\xi = \pi/(\pi + \beta)$. Note that $\nu$ (often also denoted as $T$) represents the probability of answering truthfully for respondents who are actually carrying the sensitive attribute. The models may be estimated by employing the EM-algorithm (Dempster, Laird, & Rubin, 1977) adapted for binary tree models (Hu & Batchelder, 1994) as implemented in freely available software programs such as Appletree (Rothkegel, 1999), GPT (Hu & Phillips, 1999), and HMMTree (Stahl & Klauer, 2007).

### insert Figures 1-3 about here ###

## Statistical Efficiency

Table 1 shows the relative efficiency of the proposed method with respect to the Warner (1965) model as a function of $\pi$ and $\gamma$, where $N_{Warner} = n_1 + n_2$ and $n_1 = n_2$. In this table, the randomization probability is fixed at $p_1 = .67$ (with $p_2 = 1 - p_1$), corresponding to the mean $p$ found in a meta-analysis of 42 randomized response studies (Lensvelt-Mulders et al., 2005).

Generally, for values of $\pi$ and $\gamma$ most likely to be encountered in practice, our method suffers a slight loss in efficiency. It is interesting to note that this observation also holds true if $\gamma = 0$, that is, for completely truthful responding, unless $\pi$ exceeds .4.

## Conclusion

To the best of our knowledge, the present contribution shows for the first time how the proportion of participants actually carrying a sensitive attribute but failing to respond truthfully can be estimated. This is done by a small modification of Mangat's (1994) variant of the RRT. Because the possibility of noncompliant carriers of a critical attribute is perhaps the major obstacle to a more widespread use of the RRT in survey research, we believe that our model, which allows to determine the prevalence of such behaviour, will be a valuable aid for behavioural and survey research involving questions on sensitive issues threatened by socially desirable responding.

References

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

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist*, *6*, 308-311.

Bourke, P. D. (1984). Estimation of proportions using symmetric randomized response designs. *Psychological Bulletin*, *96*, 166-172. doi: 10.1037/0033-2909.96.1.166.

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

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

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

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

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21-47. doi: 10.1007/BF02294263.

Hu, X., & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments & Computers*, *31*, 220-234.

Kuk, A. (1990). Asking sensitive questions indirectly. *Biometrika*, *77*, 436-438.

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

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

Mangat, N. & Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, *77*, 439-442.

Moors, J. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association*, *66*, 627-629.

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments & Computers*, *31*, 696-700.

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

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

Figure Captions

Figure 1: A multinomial model for estimating the proportion of respondents who are carrying the sensitive attribute and respond truthfully ($\pi$). $\mu$ is a helper parameter, where $\mu = \beta/(\beta + \gamma)$. Two independent samples with different randomization probabilities $p_1$ and $p_2$ are needed to make the model identifiable.

Figure 2: A multinomial model for estimating the proportion of respondents who are not carrying the sensitive attribute ($\beta$). The helper parameter $v$ represents the probability of truthful answering for respondents who are carrying the sensitive attribute, that is, $\nu = \pi/(\pi + \gamma)$. Two independent samples with different randomization probabilities $p_1$ and $p_2$ are needed to make the model identifiable.

Figure 3: A multinomial model for estimating the proportion of respondents who are carrying the sensitive attribute, but fail to respond truthfully ($\gamma$). $\xi$ is a helper parameter, where $\xi = \pi/(\pi + \beta)$. Two independent samples with different randomization probabilities $p_1$ and $p_2$ are needed to make the model identifiable.

Figures

Figure 1

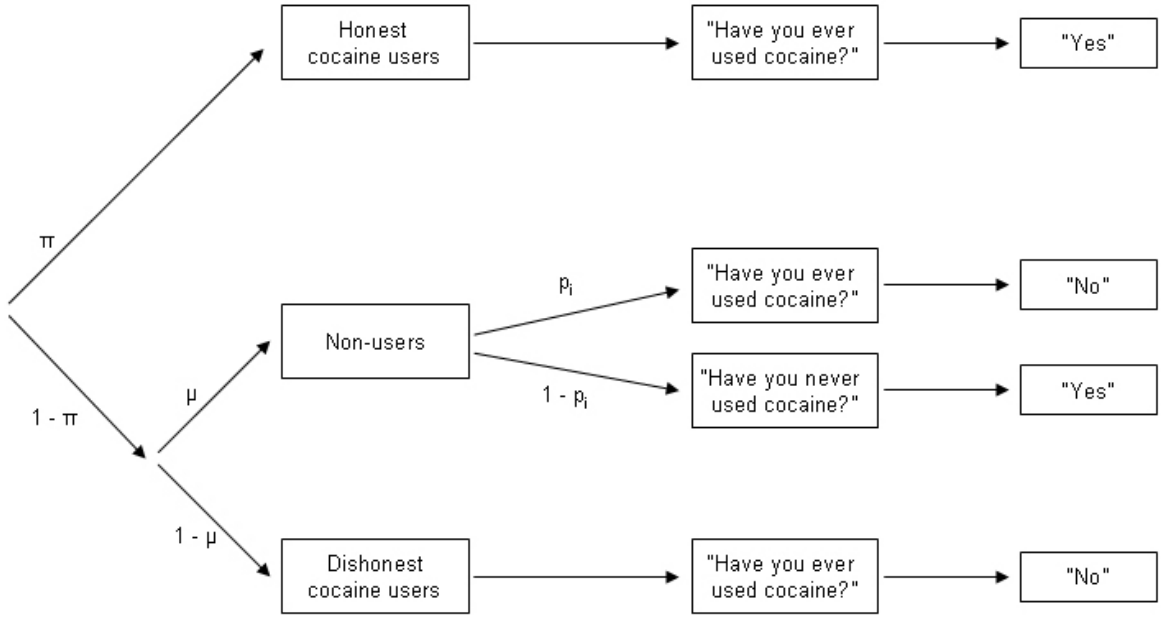


Figure 2

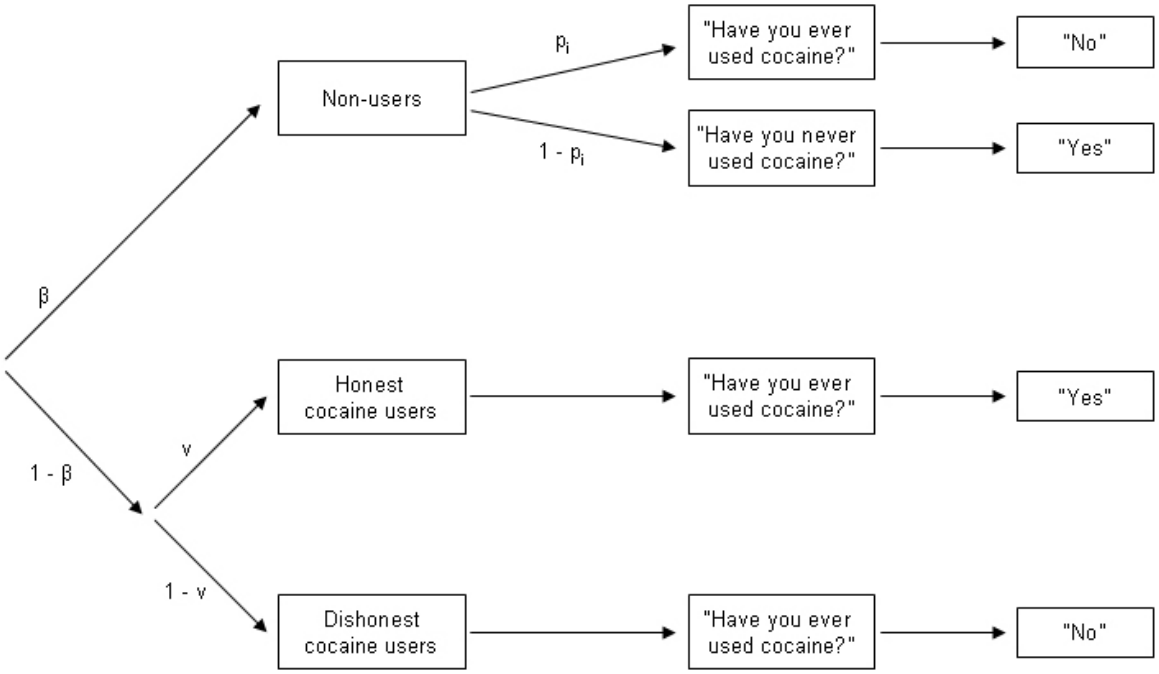
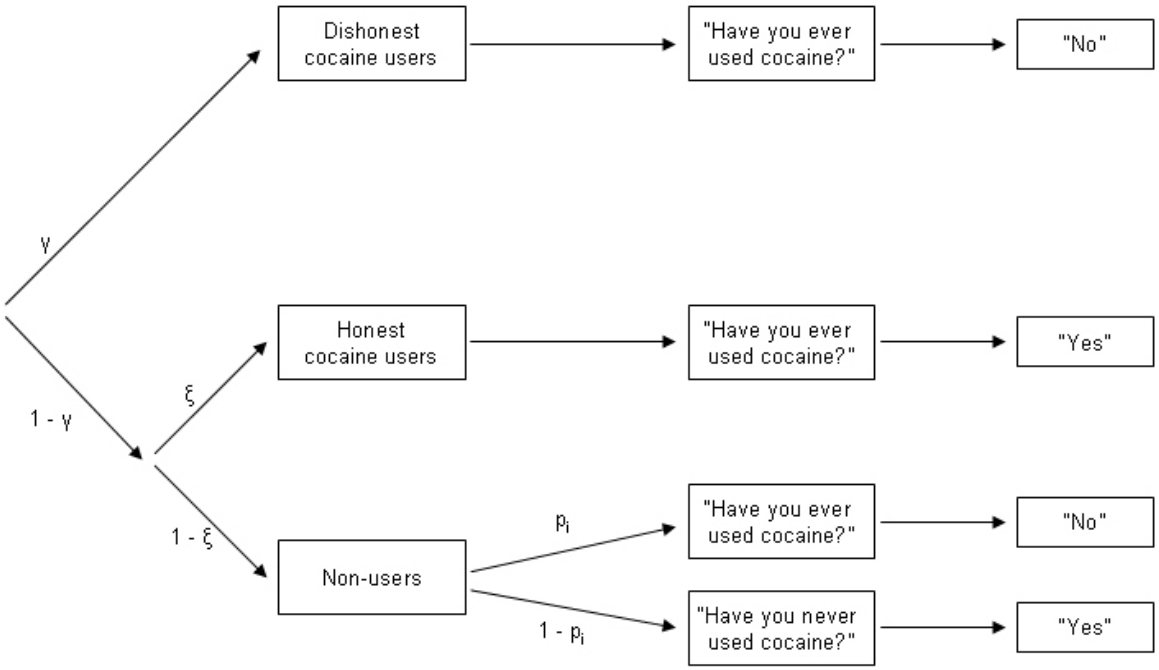

β
1 - β
Non-users
Honest cocaine users
Dishonest cocaine users
v
1 - v
$p_i$
1 - $p_i$
"Have you ever used cocaine?"
"Have you never used cocaine?"
"Have you ever used cocaine?"
"Have you ever used cocaine?"
"No"
"Yes"
"Yes"
"No"

Figure 3


Dishonest cocaine users
"Have you ever used cocaine?"
"No"
γ
Honest cocaine users
"Have you ever used cocaine?"
"Yes"
1 - γ
ξ
1 - ξ
Non-users
$p_i$
"Have you ever used cocaine?"
"No"
1 - $p_i$
"Have you never used cocaine?"
"Yes"

# Tables

Table 1. Relative efficiency of the proposed method with respect to the Warner model

| $\pi$ | $\gamma$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | .00 | .05 | .1 | .2 | .3 | .4 | .5 |
| .05 | 0.89 | 0.90 | 0.91 | 0.94 | 0.99 | 1.07 | 1.18 |
| .1 | 0.89 | 0.89 | 0.90 | 0.92 | 0.96 | 1.02 | 1.10 |
| .2 | 0.90 | 0.90 | 0.90 | 0.90 | 0.92 | 0.95 | 1.00 |
| .3 | 0.94 | 0.92 | 0.91 | 0.90 | 0.90 | 0.91 | 0.94 |
| .4 | 1.00 | 0.97 | 0.95 | 0.92 | 0.90 | 0.90 | 0.91 |
| .5 | 1.09 | 1.05 | 1.02 | 0.96 | 0.93 | 0.91 | 0.90 |

Running Head: Randomized response models

# Randomized response models: A review and a software program

Morten Moshagen & Jochen Musch

University of Duesseldorf, Germany

Correspondence should be addressed to:

Morten Moshagen

Institute for Experimental Psychology

Heinrich-Heine University

40225 Duesseldorf

Germany

Phone: +49-211-8113494

Fax: +49-211-8111753

## Abstract

Prevalence estimates of sensitive or incriminating issues are well known to be threatened by socially desirable responding. The randomized response technique (RRT) was developed as an attempt to reduce this bias by guaranteeing anonymity and confidentiality to respondents. Although the superiority of the RRT over traditional direct questioning formats has been repeatedly demonstrated, it has been rarely used in substantive research. Perhaps the major reason for the scarcity of RRT applications is the lack of an easy to use and freely available software program. In the present paper, we first review thirteen randomized response models and show how to adopt a multinomial representation of these. Second, we present a java-based software program allowing for the multinomial analysis of these models, including support for multiple-group analyses and power analyses.


Keywords: Randomized response technique, multinomial processing tree models, power analysis, computer program, social desirability

# Randomized response models: A review and a software program

In behavioural and survey research, it is often desired to estimate the population proportion that holds a certain attitude or has engaged in a certain behavior. To this end, researchers usually ask participants directly on the issue under consideration and utilize the resulting observed proportion of a particular response as an estimate of the prevalence of the respective attribute. It is well known, however, that survey responses merely reflect what participants tell investigators rather than their true status. As a consequence, prevalence estimates of sensitive, incriminating or illegal attributes are systematically biased toward respondents' perception of what is socially acceptable (e.g., Tourangou & Yan, 2007).

The randomized response technique (RRT; Warner, 1965) was developed as a means to overcome this response bias by providing a higher degree of anonymity and confidentiality than traditional direct questioning formats. The rationale of the RRT is to add random noise to the responses given by the participants such that the true status of an individual is not identifiable from his or her response. Although the randomization process renders it impossible to gather information about the status of individuals, elementary probability calculations allow us to yield group based prevalence estimates of the attribute in question.

Since the seminal work of Warner (1965), a variety of variants of the RRT have been proposed and successfully employed to obtain information about attitudes and behaviors as diverse as academic cheating (Scheers & Dayton, 1987), alcohol abuse (Volicer, B.J. & Volicer, L., 1982), employee theft (Wimbush & Dalton, 1997), doping in fitness sports (Simon, Striegel, Aust, Dietz, & Ulrich, 2006), hygiene practices (Moshagen, Ostapczuk, Zhao, & Musch, 2008), illegal substance use (Fisher, Kupferman, & Lesser, 1992), medication non-adherence (Ostapczuk, Musch, & Moshagen, 2008b), rape (Soeken & Damrosch, 1986), smuggle (Nordlund, Holme, & Tamsfoss, 1994), social security fraud (Lensvelt-Mulders, van der Heijden, & Laudy, 2006), and xenophobia (Ostapczuk, Musch, &

Moshagen, 2008a). A recent meta-analysis (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005) confirmed that the RRT generally yields more valid prevalence estimates of sensitive attributes than traditional data collection techniques. Lensvelt-Mulders, Hox, et al. (2005) concluded that "currently available research has not demonstrated the superiority of any [italics added] data collection method to RRT" (p.343).

Given that research repeatedly demonstrated the superiority of randomized response models over direct questioning formats, it is desirable that the RRT be routinely used in applied research on sensitive issues. However, as noted by Umesh and Peterson (1991), there is a mismatch between the theoretical development of the RRT and studies using this technique for substantive research questions (see also, Antonak & Livneh, 1995). Perhaps the major obstacle to a wider use of randomized response models is the lack of a freely available and easy to use software program. Therefore, the present paper presents a software program tailored for the needs of a wider audience wishing to use randomized response models in applied research settings. In the remainder of this paper, we first review thirteen randomized response models and show how to adopt a multinomial representation of these. Second, we present a software program called "RRTM" (Randomized Response Tree Modelling) which is suitable for the analysis of the randomized response models in single and multiple groups and also includes the possibility to perform power analyses.

## Randomized Response Models

This section is devoted to the review of thirteen randomized response models. Extending the taxonomy proposed by Antonak and Livneh (1995), we classify randomized response models into related question models, unrelated question models, forced response models (also known as directed-answer models), random binary event models, and variants of these models that are capable of detecting untruthful answering.

## Related Question Models

### Warner Model

In the historically first randomized response model propounded by Warner (1965), respondents are to answer either the sensitive question (e.g., "Have you ever used cocaine?") with probability $p$ (where $p \neq .5$) or its negation ("Have you never used cocaine?") with probability $1 - p$ (Figure 1). Because it is not known to the interviewer which of these questions was answered, a "yes" answer is no longer stigmatizing to the participants, thus increasing the truthfulness of responses. Generally, a "yes" answer may indicate that a respondent who has used cocaine answered the sensitive question with probability $p$ or that a respondent who has not used cocaine answered the negation of the sensitive question with the complementary probability $1 - p$. Hence, the proportion of "yes" responses ($\lambda$) is

$$\lambda = \pi p + (1 - \pi)(1 - p). \tag{1}$$

A simple algebraic rearrangement yields a maximum likelihood estimator of the proportion of respondents actually having used cocaine, that is, the prevalence of the sensitive attribute ($\pi$):

$$\hat{\pi} = \frac{\hat{\lambda} + (p - 1)}{(2p - 1)}, \tag{2}$$

with variance

$$\text{var}(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} + \frac{p(1 - p)}{n(2p - 1)^2}. \tag{3}$$

The variance estimator includes two terms: the first term is the usual sampling variance of proportions; the second term represents the variance added by the randomization procedure. Because the second term is always greater than zero, the RRT suffers a considerable loss of efficiency compared to a direct question (e.g., Lensvelt-Mulders, Hox, van der Heijden, 2005). Accordingly, many efforts of developing variants of the Warner model were directed to reduce the variance added by the randomization procedure.

### please insert figure 1 about here ###

Mangat's Two-Step Procedure

In Mangat's (1994) variant of the Warner model (Figure 2), each respondent actually having the sensitive attribute is asked to answer the sensitive question truthfully. Respondents who do not have the sensitive attribute receive the sensitive question with probability $p$ or its negation with probability $1-p$. Consequently, only the "yes" responses are distorted with

$$\lambda = \pi + (1-\pi)(1-p). \tag{4}$$

The prevalence of the sensitive attribute can be estimated by

$$\hat{\pi} = \frac{\hat{\lambda} - 1 + p}{p}. \tag{5}$$

Given that only the "yes" responses are distorted and all "no" responses are true "no's", the variance of $\pi$ is smaller in Mangat's procedure than in the Warner model:

$$\text{var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + (1-\pi)\frac{(1-p)}{np}. \tag{6}$$

### please insert figure 2 about here ###

Unrelated Question Models

As an attempt to increase the truthfulness of responses, Horvitz, Shah, and Simmons (1967) proposed a modification of the Warner model known as unrelated question technique (UQT). The rationale of the UQT is that respondents should have more confidence toward the privacy protection feature of the RRT when one question pertains to a completely innocuous attribute. Although the primary motivation of this approach was to enhance respondents' cooperation, it has the further advantage of reducing the variance of the estimates (Greenberg, Abul-Ela, Simmons, & Horvitz, 1969). As shown in Figure 3, respondents either receive the sensitive

question with probability $p$ or an unrelated innocuous question with probability $1 - p$. Thus, the proportion of "yes" responses in the *i*-th sample equals the sum of respondents who have the sensitive attribute and are prompted to answer the sensitive question and respondents who have the unrelated attribute and are prompted to answer the innocuous question:

$$\lambda_i = p_i\pi + (1 - p_i 0)\pi_B, \tag{7}$$

where $\pi_B$ is the prevalence of the unrelated attribute. Depending on whether the prevalence of the unrelated attribute is known beforehand, one may distinguish unrelated question models with known and unknown prevalence.

### please insert figure 3 about here ###

UQT with Known Prevalence of the Unrelated Attribute

If the prevalence of the second attribute is known beforehand, only one sample is needed to estimate the prevalence of the sensitive attribute:

$$\hat{\pi} = \frac{\hat{\lambda} - \pi_b(1 - \hat{\lambda})}{p}. \tag{8}$$

The sampling variance of $\pi$ is

$$\text{var}(\hat{\pi}) = \frac{\hat{\lambda}(1 - \hat{\lambda})}{np^2}. \tag{9}$$

In practice, the unrelated question is often constructed from demographic information such as "Is the last digit of your telephone number odd?" or "Were you born in April?" (Scheers, 1992). When researchers want to ask a series of sensitive questions, however, respondents may feel that they are identifiable from their pattern of responses to the unrelated question (Tracy & Fox, 1981). Thus, there may be circumstances when it is desirable to use an innocuous question with unknown prevalence of the unrelated attribute.

UQT with Unknown Prevalence of the Unrelated Attribute

If the prevalence of the unrelated innocuous attribute is not known beforehand, it is possible to estimate $\pi_B$ from the data by drawing two samples. In the classic UQT with unknown prevalence of the unrelated attribute (Greenberg, Abul-Ela, Simmons, & Horvitz, 1969), two independent random samples (henceforth called conditions) are questioned with different probabilities of receiving the sensitive question (conventionally, $p_1 + p_2 = 1$, where $p_i \neq .5$). The prevalence of the sensitive attribute can then be estimated by

$$\hat{\pi} = \frac{\hat{\lambda}_1(1-p_2) - \hat{\lambda}_2(1-p_1)}{(p_1 - p_2)}, \tag{10}$$

where $\hat{\lambda}_1$ = proportion of "yes" responses in the first condition, $\hat{\lambda}_2$ = proportion of "yes" responses in the second condition, $p_1$ = probability of receiving the sensitive question in the first condition, and $p_2$ = probability of receiving the sensitive question in the second condition. The variance of $\pi$ is given by

$$\text{var}(\hat{\pi}) = \frac{1}{(p_1 - p_2)^2}\left[\frac{\hat{\lambda}_1(1-\hat{\lambda}_1)(1-p_2)^2}{n_1} + \frac{\hat{\lambda}_2(1-\hat{\lambda}_2)(1-p_1)^2}{n_2}\right] \tag{11}$$

with $n_1$=number of respondents in the first condition and $n_2$=number of respondents in the second condition. The multinomial model depicted in Figure 3 shows only one condition; however, the second condition could be represented by an identical figure with the sole exception that $p_1$ would be replaced by $p_2$.

### please insert figure 4 about here ###

Moors' Procedure

Moors (1971) developed an optimization of the UQT with unknown prevalence of the unrelated attribute by proposing to fix the probability of answering the sensitive question at

zero in the second condition; that is, a second condition sans randomization is used to estimate the prevalence of the unrelated attribute (Figure 4). Consequently, the proportion of observed "yes" responses in the second condition ( $\hat{\lambda}_2 = n_{y2} / n_2$ ) also is the estimate of the prevalence of the unrelated attribute. The prevalence of the sensitive attribute can then be computed by

$$\hat{\pi} = \frac{\hat{\lambda}_1 - \hat{\lambda}_2 (1 - p_1)}{p_1} \tag{12}$$

with a variance of

$$\text{var}(\hat{\pi}) = \left[ \frac{(1 - p_1)\sqrt{\hat{\lambda}_2 (1 - \hat{\lambda}_2)}\sqrt{\hat{\lambda}_1 (1 - \hat{\lambda}_1)}}{p_1 n^2} \right]^2 . \tag{13}$$

### please insert figure 5 about here ###

Bourke's Symmetric UQT

Bourke (1984) called attention to the fact, that unrelated question models offer less protection to respondents actually carrying the sensitive attribute than the Warner model. The Warner model is symmetric in that both "yes" and "no" responses may be associated with having the sensitive attribute. In unrelated question models, however, a "yes" response indicates that a respondent has either the sensitive attribute or the unrelated attribute, whereas a "no" response conveys the information that a respondent has neither of these attributes. Therefore, a "no" response is safer, for it is never associated with the sensitive attribute. Accordingly, asymmetric designs are more likely to provoke respondent's non-cooperation.

In order to improve the protection offered by unrelated question models, Bourke (1984) proposed a symmetric model by combining the Warner model with the UQT with known prevalence of the unrelated attribute (Figure 5). Participants randomly get one of three cards containing two mutually exclusive questions each (e.g., 1. "I have used cocaine." and 2.

"I have never used cocaine") and are prompted to give as response the number of the question describing their true status. The first two cards contain the sensitive question and its negation (the Warner format) in balanced order, and the third card contains the unrelated question and its negation. If $p_a$ is the probability of answering the sensitive question in order A, $p_b$ ($\neq p_a$) is the probability of answering the sensitive question in order B, and $p_u$ ($=1-p_a-p_b$) is the probability of answering the unrelated question, the observed proportion of "number 1" responses ($\lambda_1$) is given by

$$\lambda_1 = \pi p_a + (1-\pi)p_b + \pi_b p_u . \tag{14}$$

Since the intuitive estimator of $\hat{\lambda}_1$ ($=n_{y1}/n_1$) may lie outside the admissible region, $\hat{\lambda}_1$ is truncated to fall in the region of $\min(p_a, p_b) > n_{y1}/n_1 > \max(p_a, p_b)$. The maximum likelihood estimator for the prevalence of the sensitive attribute is

$$\hat{\pi} = \frac{\hat{\lambda}_1 - \pi_b p_u - p_b}{(p_a - p_b)} \tag{15}$$

with a variance of

$$\text{var}(\hat{\pi}) = \frac{\hat{\lambda}_1(1-\hat{\lambda}_1)}{n(p_a - p_b)^2} . \tag{16}$$

Using the multinomial modeling framework, it is also possible to extend Bourke's symmetric UQT with known prevalence to situations where the prevalence of the unrelated attribute is not known beforehand. The model shown in Figure 5 can be extended to two conditions with different randomization probabilities $p_{ai}$, $p_{bi}$, and $p_{ui}$. When $\pi$ and $\pi_B$ are restricted to equality across conditions, the two independent proportions of observed "number 1" responses allow for estimating the prevalence of both the sensitive and the unrelated attribute.

### please insert figure 6 about here ###

## Forced Response Models

Forced response (or directed-answer) models (Dawes & Moore, 1980; Greenberg et al., 1969) were developed to eliminate the problem in unrelated question models to either know or determine the prevalence of the unrelated attribute. In forced response models, only one question is asked, but a certain proportion of participants is prompted to disregard the question entirely and to provide a prespecified response. Depending on the outcome of the randomization device, respondents are prompted to reply "yes" with probability $p_y$ or "no" with probability $p_n$ independently of the content of the question, or are asked to respond truthfully with probability $1 - p_y - p_n$ (Figure 6). The proportion of "yes" responses equals the sum of those respondents who were prompted to reply "yes" irrespectively of the content of the question and those respondents who have the sensitive attribute and were prompted to answer truthfully:

$$\lambda = \pi(1 - p_y - p_n) + p_y . \tag{17}$$

The prevalence of the sensitive attribute can be estimated by

$$\hat{\pi} = \frac{\lambda - p_y}{(1 - p_y - p_n)} \tag{18}$$

with a variance

$$\text{var}(\hat{\pi}) = \frac{\hat{\lambda}\ (1 - \hat{\lambda}\ )}{n(1 - p_y - p_n)^2} . \tag{19}$$

In the asymmetric forced response variant (Dawes & Moore, 1980), $p_n$ is set equal to zero. The maximum likelihood estimator and its variance remain identical to the symmetric forced response variant (with $p_n = 0$), however, the proportion of "yes" responses becomes

$$\lambda = \pi + (1 - \pi) p_y , \tag{20}$$

since respondents who have the sensitive attribute will always reply "yes", and respondents who do not have the sensitive attribute will only reply "yes" if prompted by the outcome of the randomization device.

### please insert figure 7 about here ###

## Random Binary Outcome Models

The basic idea of random binary outcome models is to let participants generate two binary outcomes and report either the first or the second outcome instead of a "yes" or a "no" answer. In Kuk's (1990) playing card method (Figure 7), the binary outcomes are generated by two card decks containing a different proportion of red cards ( $p_1$ and $p_2$, where $p_1 \neq p_2$). Respondents receive the sensitive question and are prompted to provide the first and second outcome if they have and do not have the sensitive attribute, respectively. The proportion of reported red cards is

$$\lambda = \pi p_1 + (1-\pi) p_2 . \tag{21}$$

The maximum likelihood estimator of the prevalence of the sensitive attribute is

$$\hat{\pi} = \frac{\hat{\lambda} - p_2}{(p_1 - p_2)} \tag{22}$$

with variance

$$\text{var}(\hat{\pi}) = \frac{\hat{\lambda}(1-\hat{\lambda})}{n(p_1 - p_2)^2} . \tag{23}$$

## Detecting Untruthful Answering in Randomized Response Models

Although is has been repeatedly demonstrated that RRT based prevalence estimates are less biased than prevalence estimates based on more conventional data collection techniques, it is clear that randomized response models rather reduce than eliminate the tendency to provide

socially acceptable responses (Campbell, 1987; Edgell, Duchan, & Himmelfarb, 1992; Landsheer, van der Heijden, & van Gils, 1999; Lensvelt-Mulders, Hox, et al., 2005; Soeken & McReady, 1982). Accordingly, more recent developments attempt to estimate proportion of participants that fail to respond truthfully.

### please insert figure 8 about here ###

Cheating Detection Model

The cheating detection model (Clark & Desharnais, 1998) extends the asymmetric forced response model (Dawes & Moore, 1980) to two conditions with different probabilities of being prompted to reply "yes" $p_{y1}$ and $p_{y2}$. It is assumed that a certain proportion of respondents (hereafter called cheaters) fails to comply with instructions and replies "no" although being asked by the outcome of the randomization device to answer affirmatively. Two types of cheating may occur in the asymmetric forced response model (Antonak & Livneh, 1995; Lensvelt-Mulders & Boeije, 2007): First, respondents actually carrying the sensitive attribute may reply "no" despite being asked to answer truthfully (with $1 - p_{yi}$). Second, both guilty and innocent respondents may decide to reply "no" although being prompted to answer affirmatively (with $p_{yi}$). Because both guilty and innocent respondents may decide to disregard the instructions, no assumption is made about the true status of the cheating participants. As shown in Figure 8, the cheating detection model therefore contains three parameters representing the proportion of non-compliant cheaters ($\gamma$), the proportion of participants that have ($\pi$), and the proportion of participants that do not have ($\beta$) the sensitive attribute (where $\gamma + \pi + \beta = 1$). Although the true status of the non-compliant cheaters cannot be identified, it is still possible to compute upper and lower bound estimates of the prevalence of the sensitive attribute by alternately assuming that all non-compliant

participants either have ( $\hat{\pi}+\hat{\gamma}$ ) or do not have ( $\hat{\pi}$ ) the sensitive attribute (Moshagen, Ostapczuk, et al., 2008). Under the assumption that the parameters $\pi$ , $\beta$ , and $\gamma$ are equal across conditions, the observed proportion of "yes" responses in condition *i* is

$$\lambda_i = \pi + p_{yi}\beta \tag{24}$$

and the maximum likelihood estimators for the three parameters are

$$\hat{\pi} = \frac{p_{y2}\hat{\lambda}_1 - p_{y1}\hat{\lambda}_2}{(p_{y2} - p_{y1})}, \tag{25}$$

$$\hat{\beta} = \frac{\hat{\lambda}_2 - \hat{\lambda}_1}{(p_{y2} - p_{y1})}, \tag{26}$$

and

$$\hat{\gamma} = 1 - \hat{\pi} - \hat{\beta}\,. \tag{27}$$

The asymptotic variance of $\pi$ and $\beta$ are given by

$$\text{var}(\hat{\pi}) = \frac{1}{(p_{y1} - p_{y2})^2}\left[\frac{{p_{y1}}^2 n_{y2} n_{n2}}{{N_2}^3} + \frac{{p_{y2}}^2 n_{y1} n_{n1}}{{N_1}^3}\right] \tag{28}$$

and

$$\text{var}(\hat{\beta}) = \frac{1}{(p_{y1} - p_{y2})^2}\left[\frac{n_{y2} n_{n2}}{{N_2}^3} + \frac{n_{y1} n_{n1}}{{N_1}^3}\right], \tag{29}$$

where $n_{yi}$ and $n_{ni}$ represent the observed frequency of "yes" and "no" responses in the *i*-th condition, respectively.

Since the asymmetric forced response model underlying Clark and Desharnais' (1998) cheating detection model is not optimal with respect to the protection of respondents' privacy (Bourke, 1984), Ostapczuk, Moshagen, Zhao, and Musch (2008) proposed a symmetric variant of the cheating detection model, in which participants are prompted either to provide the prespecified answers "yes" with probability $p_{yi}$ or "no" with probability $p_{ni}$ , or to respond honestly with probability $1 - p_{yi} - p_{ni}$ . The parameters of the symmetric cheating

detection model can be estimated by using the multinomial modelling framework as described in greater detail below.

### please insert figure 9 about here ###

Chang and Huang's Two-Step Procedure

In Chang and Huang's (2001) procedure (Figure 9), each participant receives a sensitive question in a direct questioning format at the first step. If a participant denies the direct question, the sensitive question is asked again in the Warner format at the second step. Using two conditions with different randomization probabilities, it is possible to estimate the probability that participants who have the sensitive attribute reply truthfully when questioned directly (denoted as $T$). The proportion of "yes" responses in condition $i$ thus equals the sum of participants that have the sensitive attribute and reply truthfully, participants that have the sensitive attribute and reply "yes" when asked the sensitive question in the Warner format, and participants that do not have the sensitive attribute and reply "yes" to the negation of the sensitive question when asked in the Warner format:

$$\lambda_i = \pi T + \pi(1-T)p_i + (1-\pi)(1-p_i)\,. \tag{30}$$

The maximum likelihood estimator of the prevalence of the sensitive attribute and its variance are identical to the unrelated question model with unknown prevalence (Equations 10 and 11). An estimator of $T$ is given by

$$\hat{T} = \frac{(1-2p_2)\hat{\lambda}_1 - (1-2p_1)\hat{\lambda}_2 - (p_1-p_2)}{(1-p_2)\hat{\lambda}_1 - (1-p_1)\hat{\lambda}_2} \tag{31}$$

with variance

$$\operatorname{var}(\hat{T}) = \frac{1}{(p_1-p_2)^2\pi^2}\left\{\frac{\left[\hat{T}(1-p_2)+(2p_2-1)\right]^2\hat{\lambda}_1(1-\hat{\lambda}_1)}{n_1} + \frac{\left[\hat{T}(1-p_1)+(2p_1-1)\right]^2\lambda_2(1-\hat{\lambda}_2)}{n_2}\right\} \tag{32}$$

It is important to note that the parameter *T* differs substantially from the proportion of cheating participants ($\gamma$) in the cheating detection models (Clark & Desharnais, 1998; Ostapczuk, Moshagen, et al., 2008): Chang and Huang's (2001) two-step procedure assumes that participants may not respond truthfully when asked directly, but are completely honest when asked in the Warner format. The cheating detection models, however, do not assume that participants are completely honest when being asked in a randomized response format, but still may decide to cheat. Consequently, the $\gamma$ parameter directly translates to an estimate of the proportion of non-compliant participants and is of substantive interest (e.g., for constructing upper-bound prevalence estimates; Moshagen, Ostapczuk, et al., 2008), whereas the parameter *T* is an estimate of the probability of truthful responding given a direct question and may primarily be used as a measure of sensitivity for deciding whether to employ a randomized response survey.

## Multinomial Processing Tree Models

It is possible to subsume the randomized response models reviewed above under the more general family of multinomial processing tree models (Batchelder & Riefer, 1999; Hu & Batchelder, 1994). This representation has several advantages: First, it is easily possible to place constraints on certain parameters, for example, to test whether an RRT-based prevalence estimate of a sensitive attribute differs significantly from the estimate obtained in a conventional direct questioning condition. Second, the models can be extended to the simultaneous analysis of multiple groups with and without cross-group equality constraints on the parameters. Third, more complex models involving additional parameters may be formulated, permitting the estimation of models for which there are currently no closed-form solutions available, such as the extension of Bourke's symmetric UQT to situations when the prevalence of the second attribute is unknown and the symmetric variant of Clark and

Desharnais' (1998) cheating detection model (Ostapczuk, Moshagen, et al., 2008). Finally, it is easily possible to perform power analyses for each of these models in order to determine the required sample size to reliably reject a certain null hypothesis if it is in fact false.

Definition of Multinomial Processing Tree Models

Multinomial processing tree models attempt to estimate latent parameters from observed category frequency counts. A multinomial processing tree model comprises a set of branches that lead to observable response categories. Each branch consists of a series of conditional link probabilities of one stage to another. The probability of a branch *i* leading to category *j* equals the product of the corresponding link probabilities

$$p_{ij}(\Theta) = c_{ij} \prod_{s=1}^{S} \theta_s^{a_{ijs}} (1-\theta_s)^{b_{ijs}} , \tag{33}$$

where $\Theta = (\theta_1,...,\theta_s)$ is a vector of the parameters representing the conditional link probabilities, $a_{ijs}$ and $b_{ijs}$ are nonnegative integer structure coefficients, and *c* is a nonnegative real representing the product of constants on the links. A particular response may be reached by more than one branch, thus, the probability of observing a response in category *j* equals the sum of the branch probabilities leading to that category

$$p_j(\Theta) = \sum_{i=1}^{Ij} p_{ij}(\Theta) . \tag{34}$$

To illustrate the definition of multinomial tree models consider as an example the asymmetric forced response model, which can be derived from the symmetric forced response model shown in Figure 6 by setting $p_n$ equal to zero. The reduced model consists of four branches leading to two response categories ("yes" and "no") and comprises two independent parameters ($\pi$ and $p_y$). Participants who have the sensitive attribute ($\pi$) are either prompted to reply "yes" with probability $p_y$ or to answer truthfully with probability $p_t = 1 - p_y$. In either way, these participants will reply "yes". Participants who do not have the sensitive

attribute ($1-\pi$) are also prompted to reply "yes" with probability $p_y$ or to answer truthfully with probability $p_t = 1 - p_y$. These participants will reply "yes" if asked by the outcome of the randomization device to answer affirmatively, but will answer "no" if asked to answer truthfully. Thus, the probabilities of the three branches leading to a "yes" response are

$$p_{1y} = \pi p_y, \tag{35}$$

$$p_{2y} = \pi(1 - p_y), \tag{36}$$

$$p_{3y} = (1-\pi)p_y \tag{37}$$

and the probability of the branch leading to a "no" response is

$$p_{1n} = (1-\pi)(1-p_y). \tag{38}$$

The joint probability of observing a "yes" and "no" response, respectively, is simply the sum of the probabilities of corresponding branches:

$$p_{yes} = \sum_{i=1}^{3} p_{iy} = \pi p_y + \pi(1-p_y) + (1-\pi)p_y \tag{39}$$

and

$$p_{no} = \sum_{i=1}^{1} p_{in} = (1-\pi)(1-p_y). \tag{40}$$

Note that $p_{yes}$ is equivalent to the expected proportion of observed "yes" responses ($\lambda$) of the asymmetric forced response model (Equation 20).

As the definition of multinomial processing tree models given above only covers the case of binary tree models, a slight complication arises for randomized response models that require more than two branches at a given stage of the model. For example, in the symmetric forced response model (Figure 6), respondents are prompted to reply "yes", "no", or truthfully, resulting in three branches. However, it is easily possible to reparametrize this model into a statistically equivalent binary tree model: Because the randomization probabilities are constrained to add up to one ($p_y + p_n + p_t = 1$) and two of three branches

lead to the same response category, the probabilities $p_y$ and $p_t$ can be replaced by $1 - p_n$ in the branch for the cocaine-users and the probabilities $p_n$ and $p_t$ can be replaced by $1 - p_y$ in the branches for the non-users. A similar reparametrization can be applied to Bourke's (1984) variant of the UQT and to the cheating detection models (Clark & Desharnais, 1998; Ostapczuk, Moshagen, et al., 2008). The latter models, however, additionally require estimating a particular model for each of the three parameters $\pi$, $\beta$, and $\gamma$ in order to obtain appropriate parameter standard errors (see Moshagen, Musch, et al., 2008, for details).

Parameter Estimation

Parameter estimation proceeds by employing the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) adapted for binary tree models (Hu, 1999; Hu & Batchelder, 1994). The algorithm attempts to obtain a set of parameters $\Theta = (\theta_1,...,\theta_s)$ that minimize the distance between observed and expected frequencies. Multinomial processing tree models utilize distance measures that can be characterized as a power divergence family (Read & Cressie, 1988). $PD^{\lambda}$ defines an asymptotically $\chi^2$ distributed family of distance measures depending on the value of $\lambda$

$$PD^{\lambda}(R, P(\Theta)) = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{j} n_j \left[ \left( \frac{n_j}{NP_j(\Theta)} \right)^{\lambda} - 1 \right], \tag{41}$$

where $R = (n_1,...,n_j)$ is a vector of observed frequencies and $P(\Theta) = [NP_1(\Theta)\ ,...,NP_j(\Theta)]$ is a vector of expected frequencies given parameter $\Theta$. The Pearson $\chi^2$ distance measure is a special case with $\lambda = 1$. For $\lambda = 0$, $PD^{\lambda}$ reduces to the likelihood ratio statistic $G^2$.

The EM algorithm is an iteration of trials consisting of an expectation (E) and a maximization (M) step. In the E-step of the EM algorithm, expected frequencies $m_{ij}$ for each

branch are generated given the parameter vector $\Theta$ from the previous trial (respectively, the initial start values)

$$m_{ij}(\Theta) = \frac{n_j p_{ij}(\Theta)}{p_j(\Theta)}. \tag{42}$$

The M-step calculates revised parameter estimates $\Phi = (\phi_1,...,\phi_s) = M(\Theta)$ given the expected frequencies of the E-step:

$$\phi_s{}^{\lambda}(\Theta) = \frac{\sum_{j=1}^{J}\left[\left(\frac{n_j}{np_j(\Theta)}\right)^{\lambda}\sum_{i=1}^{I_j}\left(\frac{n_j np_{ij}(\Theta)}{p_j(\Theta)}a_{ijs}\right)\right]}{\sum_{j=1}^{J}\left[\left(\frac{n_j}{np_j(\Theta)}\right)^{\lambda}\sum_{i=1}^{I_j}\left(\frac{n_j np_{ij}(\Theta)}{p_j(\Theta)}(a_{ijs}+b_{ijs})\right)\right]}. \tag{43}$$

The final parameter estimates after each iteration depend on the value of the step-width parameter $\varepsilon$

$$\Theta^{(n)} = \Theta^{(n-1)} - \varepsilon[\Theta^{(n-1)} - M_{\lambda}(\Theta^{(n-1)})]. \tag{44}$$

Variances (and hence standard errors) for the parameter estimates may be obtained by computing the inverse of the observed Fisher information matrix which is an estimate of the variance-covariance matrix of the estimators (see Hu & Batchelder, 1994, for details). As the maximum likelihood estimates of the parameters are asymptotically normally distributed, confidence intervals can be computed according to $\theta_s \pm z_{\alpha} SE_{\theta_s}$, where $z_{\alpha}$ is the tail of the standard normal distribution corresponding to the desired $\alpha$ level.

RRTM: A Software Program to Multinomially Analyze Randomized Response Models

RRTM is a Java-based software program for the multinomial analysis of the randomized response models reviewed in the preceding sections. Given a particular model, design parameters, and a set of frequencies, RRTM computes maximum likelihood estimates of the parameters (using $PD^{\lambda=0} = G^2$ as distance measure) along with standard errors, 95%-confidence intervals, and significance levels. Owing to the multinomial modeling approach, a

model can be easily extended to estimate the parameters simultaneously for multiple groups. Note that a distinction is made between the number of conditions that are necessary to make a particular model identifiable and the number of groups for which the parameters should be estimated separately. For example, employing Clark and Desharnais' (1998) cheating detection model requires two conditions with different randomization probabilities for the purpose of identification. In contrast, groups always comprise as many conditions as necessary to obtain an identified model. Groups may represent different subgroups such as gender groups, but may also represent different randomized response questions such as hierarchically ordered questions on a quantity of a sensitive attribute (e.g., "Were you ever involved in a theft from your employer of cash worth from 5$-10$/10$-50$/50$ and more?"). The number of conditions is determined from the selected model, whereas the number of groups should be chosen according to the particular research question. If more than one group was specified, RRTM provides parameter estimates both with and without cross-group equality restrictions on the parameters along with $\Delta G^2$ statistics indicating the applicability of these constraints.

RRTM is not limited to the estimation of the parameters of a model given observed frequencies, but is also capable of performing a-priori and post hoc power analyses. Statistical power of a test is defined as the complement of the $\beta$-error probability of falsely retaining an incorrect null-hypothesis (Cohen, 1988; Faul, Erdfelder, Lang, & Buchner, 2007). Generally, the power of a test is a function of the $\alpha$-error probability (the significance level), the sample size, and the degree of deviation between null ($H_0$) and alternative hypothesis ($H_1$). In the context of multinomial models, the latter can be defined as the difference between the parameters of a more restricted $H_0$ model and a less restricted $H_1$ model (Erdfelder, 2000). For instance, one may be interested in the required sample size to reliably detect that a certain proportion carrying a sensitive attribute (e.g., $\pi$=.05) differs significantly from zero. The $H_1$ model would place no constraints on the parameter representing the proportion carrying a

sensitive attribute ($\pi$). In the $H_0$ model, however, $\pi$ would be restricted to be equal to zero. Power can then be calculated by evaluating the non-central $\chi^2$ distribution at a given $\alpha$ with the difference of the $G^2$ fit-statistics of the more restricted $H_0$ model and the less restricted $H_1$ model as an estimate of the non-centrality parameter. Two types of power-analysis are implemented in RRTM: First, in a-priori power analyses (Cohen, 1988), the required sample size to reject a $H_0$ if it is in fact false is computed given a significance level $\alpha$ and the desired power. Second, in post-hoc power analyses (Cohen, 1988), the achieved power to reject a $H_0$ if it is in fact false is computed given a significance level $\alpha$ and a prespecified sample size.

## Program Handling

The following sections describe how to use RRTM to estimate the parameters of a model and how to perform power analyses. Finally, an overview of the save and export features of RRTM are given.

### Estimating the Parameters of a Randomized Response Model

Using RRTM to estimate the parameters of a particular model involves three steps: (1) select and specify a model, (2) enter data, and (3) perform analysis to view the results. Right after starting RRTM, a dialog appears prompting to select a model, to specify the number of groups, and to indicate whether the design includes a direct questioning control condition. After selecting and specifying a model, the observed frequencies and the randomization probabilities need to be entered. Parameter estimation may then be started by selecting "Run Analysis" from the menu, pressing ALT+R, or clicking the green arrow in the toolbar. The output is organized as follows: After repeating a summary of the input, the actual output starts with the goodness of fit statistic ($G^2$) of the estimated model.[1] Next, the output shows the parameter estimates for $\pi$ and $\beta$ (and eventually additional parameter estimates depending

on the selected model) along with associated standard errors, 95% confidence intervals, and $\Delta G^2$ statistics for the test that a parameter does not differ significantly from zero ($\theta_s = 0$). If a direct questioning control condition was specified, the resulting prevalence estimate with standard error and 95% confidence interval is presented below; including the $\Delta G^2$ statistic for the test that the randomized response prevalence estimate does not differ from the direct questioning prevalence estimate ($\theta_{RRT} = \theta_{DQ}$). Finally, if the design comprises multiple groups, cross-group equality constrained parameter estimates with standard errors and 95% confidence intervals as well as the $\Delta G^2$ statistics for the test that the parameters do not differ significantly across groups are given.

Performing Power Analyses

If it is desired to perform a power analysis, it is necessary to select and specify a particular model and mark the checkbox labeled "Power analysis" in the model selection window. Performing a power analysis involves four steps: (1) select the type of power analysis, (2) specify the population model, (3) specify the $H_0$ model, and (4) start the analysis to view the results. The first step is to select the type of power analysis (a-priori versus post-hoc) and, optionally, to enter the $\alpha$ level and the desired power. The population model needs to be specified in the second step. This involves setting population values for the parameters, the randomization probabilities and, if a post-hoc power analysis is to be performed, specifying sample sizes. The population model is used to generate observed frequency counts given the parameter specifications. It is not necessary to explicitly specify a $H_1$ model, since RRTM estimates an unrestricted $H_1$ model as a baseline model. The third step is to specify a $H_0$ model reflecting the particular parameter constraint of interest. The parameters of a model may be freely estimated, constrained to a constant value or restricted to be equal to another parameter. To replace a parameter with a constant, the drop-down box next to the parameter

of interest is switched from "free" to "constant" and the associated text-field is changed to hold the desired constant value. In order to set a parameter A to be equal to another parameter B, the name of the parameter B is selected from drop-down box next to parameter A. It is also possible to combine several constraints, for example, restricting two parameters to equality and assign a constant value to one of these parameters. After these steps are completed, the power analysis may finally be started in the same way as stated above.

To illustrate how a power analysis would actually be performed, suppose a researcher using the UQT with known prevalence of the unrelated attribute is interested in the required sample size to detect a difference of 10% between the $\pi$ parameters of two groups ($\pi_1$ and $\pi_2$) with a power of .8 at an $\alpha$ error probability of .05. The prevalence of the unrelated attribute is known to be 20%, and the probability of receiving the sensitive question is *p*=.8 for both groups. After choosing UQT with known prevalence, selecting two groups, and marking the checkbox labeled "Power analysis" in the model selection window, the first step is to change the type of power analysis to "Compute required sample size given power". The population model is specified in the second step. Suppose that the prevalence of the sensitive attribute is hypothesized to be 10% for the first group ($\pi_1 = .1$) and 20% for the second group ($\pi_2 = .2$). The randomization probabilities and the prevalence of the unrelated attribute are equal for both groups, thus we enter $p_{crit1} = p_{crit2} = .8$ and $\pi_{unrelated1} = \pi_{unrelated2} = .2$. Since we are interested in computing the required sample size, the fields labeled sample size may remain empty. The third step is to specify the $H_0$ model. In the present example, the $H_0$ model contains two parameters $\pi_1$ and $\pi_2$ which are freely estimated by default. In order to test whether these parameters can be restricted to equality given the implied frequencies of the population model, the parameters are constrained to be equal by selecting "= Pi Group 2" from the drop-down menu next to $\pi_1$. Finally, the power analysis can be started. The results

show that we would need 327 participants in each group to detect a difference between $\pi_1$ and $\pi_2$ of 10% with a power of .8 when $\alpha$ is .05.

Save and Export

RRTM offers several save and export features: The current state of the program (specified model, entered data, and displayed output) can be saved in a binary file format with file extension ".rrt". Furthermore, there is an option to save the output as a regular ASCII formatted text file. It is also possible to export equation (".eqn") and data files (".mdt") suitable to be used as input files for general purpose multinomial processing tree programs such as AppleTree (Rothkegel, 1999), gpt (Hu & Phillips, 1999), and HMMTree (Stahl & Klauer, 2007).

## System Requirements and Program Availability

RRTM is a Java-based program running under Linux, MacOS, and Windows operation systems provided that at least version 1.5.0 of the Java runtime environment is installed on the target machine. The Java runtime environment may be freely downloaded from http://java.sun.com. RRTM itself requires about 5 MB free disk space. Processing speed varies considerably with the model estimated; however, current machines estimate even very large models including multiple groups and conditions within seconds.

Linux, MacOS, and Windows versions of RRTM can be downloaded from http://www.psycho.uni-duesseldorf.de/abteilungen/ddp/RRTM free of charge for academic and personal use. Users who wish to distribute RRTM in another way need to ask the first author for permission. The charge for commercial applications is US $300. Although considerable effort has been put into program development and evaluation, there is no warranty whatsoever.

References

Antonak, R. F., & Livneh, H. (1995). Randomized response technique: A review and proposed extension to disability attitude research. *Genetic, Social & General Psychology Monographs*, 121, 97-145.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process 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. (1987). Randomized response technique. *Science*, 236, 1049.

Chang, H., & Huang, K. (2001). Estimation of proportion and sensitivity of a qualitative character. *Metrika*, 53, 269-280.

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

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

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

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Edgell, S. E., Duchan, K. L., & Himmelfarb, S. (1992). An empirical test of the unrelated question randomized response technique. *Bulletin of the Psychonomic Society*, 30, 153-156.

Erdfelder, E. (2000). *Multinomiale Modelle in der kognitiven Psychologie [Multinomial models in cognitive psychology]*. Bonn: Philosophische Fakultät der Rheinischen Friedrich-Wilhelms-Universität in Bonn.

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

Fisher, M., Kupferman, L. B., & Lesser, M. (1992). Substance use in a school-based clinic population: Use of the randomized response technique to estimate prevalence. *Journal of Adolescent Health*, 13, 281-285.

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

Horvitz, D., Shah, B., & Simmons, W. (1967). The unrelated question randomized response model. In: *Proceedings of the Social Statistics Section, American Statistical Association* (pp. 65-72).

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

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

Hu, X., & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments & Computers*, 31, 220-234.

Kuk, A. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436-438.

Landsheer, J. A., van der Heijden, P. G. M., & van Gils, G. (1997). Trust and understanding: Two psychological aspects of randomized response. *Quality & Quantity*, 33, 1-12.

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

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, Series A*, 169, 305-318.

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

Moors, J. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association*, 66, 627-629.

Moshagen, M., Musch, J., Ostapczuk, M., Mischke, R., Bröder, A., & Erdfelder, E. (2008). *Extending the cheating detection modification of the randomized response technique*. Manuscript submitted for publication.

Moshagen, M., Ostapczuk, M., Zhao, Z., & Musch, J. (2008). *Reducing socially desirable responding in epidemiological surveys using a cheating detection extension of the randomized-response-technique*. Manuscript submitted for publication.

Nordlund, S., Holme, I., & Tamsfoss, S. (1994). Randomized response estimates for the purchase of smuggled liquor in Norway. *Addiction*, 89(4), 401-405.

Ostapczuk, M., Moshagen, M., Zhao, Z., & Musch, J. (2008). *Assessing sensitive attributes using the randomized-response-technique: Evidence for the importance of response symmetry*. Manuscript submitted for publication.

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

Ostapczuk, M., Musch, J., & Moshagen, M. (2008b). *Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique*. Manuscript submitted for publication.

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

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments & Computers*, 31, 696-700.

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

Scheers, N. J., & Dayton, C. M. (1987). Improved estimation of academic cheating behavior using the randomized response technique. *Research in Higher Education*, 26, 61-69.

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

Soeken, K. L., & Damrosch, S. P. (1986). Randomized response technique: Applications to research on rape. *Psychology of Women Quarterly*, 10, 119-125.

Soeken, K. L., & Macready, G. B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92, 487-489.

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

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

Tracy, P., & Fox, J. (1981). The validity of randomized-response for sensitive measurements. *American Sociological Review*, 46, 187-200.

Umesh, U., & Peterson, R. (1991). A critical evaluation of the randomized response method: Applications, validation and research agenda. *Sociological Methods & Research*, 20, 104-138.

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.

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

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

## Footnotes

[1] Although each of the models reviewed is just-identified and, consequently, is likely to show a perfect fit to the data, the $G^2$ statistic may become greater than zero if the assumptions of a particular model are seriously violated. For example, the forced response model may yield a $G^2$ statistic greater than zero if the observed proportion of "yes" responses is smaller than would be expected solely on the basis of prompted "yes" responses (i.e., if $\lambda < p_y$).

# Authors Note

This work was supported by a grant of the Deutsche Forschungsgesellschaft (Mu 2674/1-1). Correspondence concerning this article should be addressed to Morten Moshagen (morten.moshagen@uni-duesseldorf.de) or Jochen Musch (jochen.musch@uni-duesseldorf.de), University of Duesseldorf, Institute for Experimental Psychology, Universitaetsstr. 1, 40225 Duesseldorf, Germany.

Figure captions

Figure 1:

A multinomial representation of the Warner model.

Figure 2:

A multinomial representation of Mangat's two-step procedure.

Figure 3:

A multinomial representation of the unrelated questions technique. If the prevalence of the unrelated attribute ($\pi_B$) is known beforehand, one condition is sufficient to estimate the parameters. If the prevalence of the unrelated attribute is unknown, two conditions are needed with different probabilities ($p_1$ and $p_2$) and $\pi$ and $\pi_B$ constrained to equality across conditions.

Figure 4:

A multinomial representation of Moors' variant of the unrelated questions technique.

Figure 5:

A multinomial representation of Bourke's variant of the unrelated questions technique. One condition is sufficient to estimate the prevalence of the sensitive attribute ($\pi$) if the prevalence of the unrelated attribute ($\pi_B$) is known beforehand. If the prevalence of the unrelated attribute is unknown, two conditions with different randomization probabilities ($p_{a1}$, $p_{a2}$; $p_{b1}$, $p_{b2}$; $p_{u1}$, $p_{u2}$) are needed with $\pi$ and $\pi_B$ constrained to equality across conditions.

Figure 6:

A multinomial representation of the symmetric forced response variant. In asymmetric forced response models, $p_n$ is equal to zero.

Figure 7:

A multinomial representation of Kuk's playing card method.

Figure 8:

A multinomial representation of the symmetric cheating detection model. Two conditions with different randomization probabilities ( $p_{y1}, p_{y2}; p_{n1}, p_{n2}; p_{t1}, p_{t2}$ ) are needed to make the model identifiable, where the parameters $\pi$ , $\beta$ , and $\gamma$ are constrained to be equal across conditions. In the asymmetric cheating detection model, $p_{ni}$ is equal to zero.

Figure 9:

A multinomial representation of Chang and Huang's two-step procedure. For the purpose of identification, two conditions with different probabilities of receiving the sensitive question ( $p_1$ and $p_2$ ) are needed, where $\pi$ and T are constrained to equality across conditions.

Figure 1

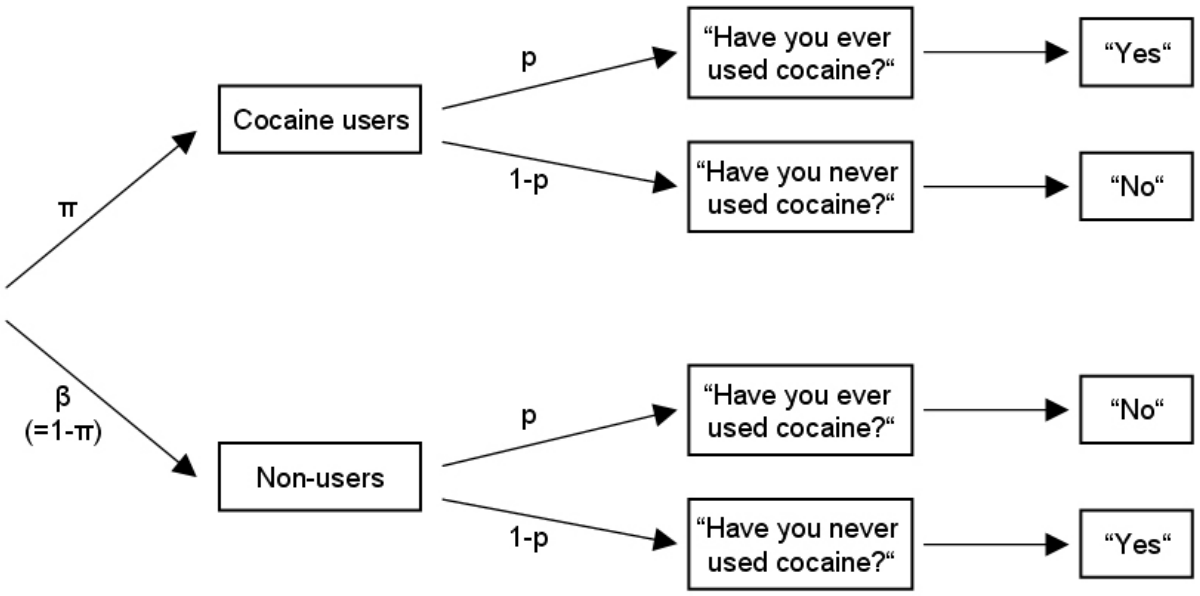

π
β
(=1-π)
Cocaine users
Non-users
p
1-p
p
1-p
"Have you ever used cocaine?"
"Have you never used cocaine?"
"Have you ever used cocaine?"
"Have you never used cocaine?"
"Yes"
"No"
"No"
"Yes"

Figure 2

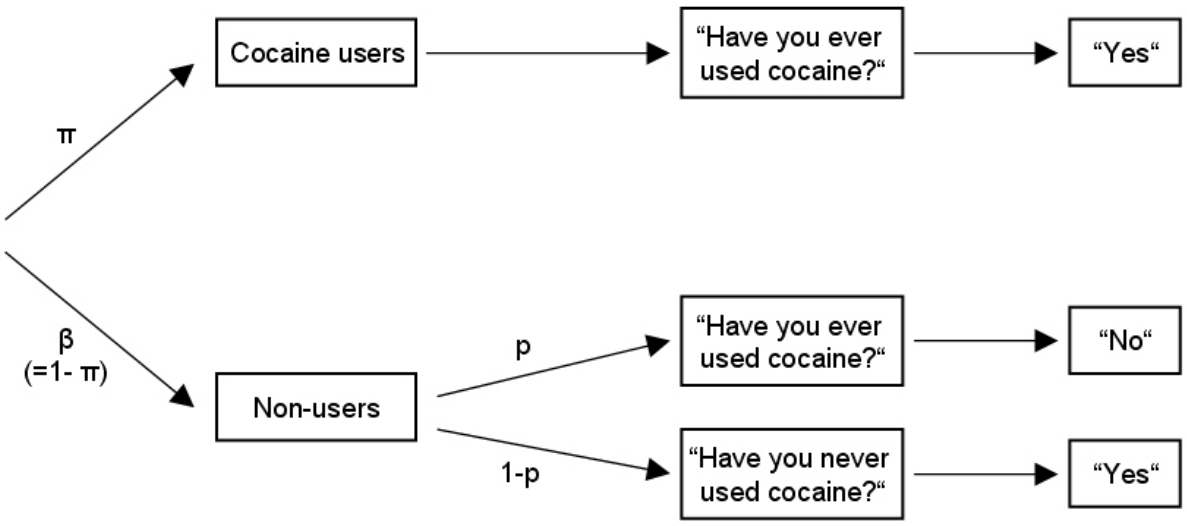

Cocaine users
"Have you ever used cocaine?"
"Yes"
π
β
(=1- π)
Non-users
p
"Have you ever used cocaine?"
"No"
1-p
"Have you never used cocaine?"
"Yes"

Figure 3

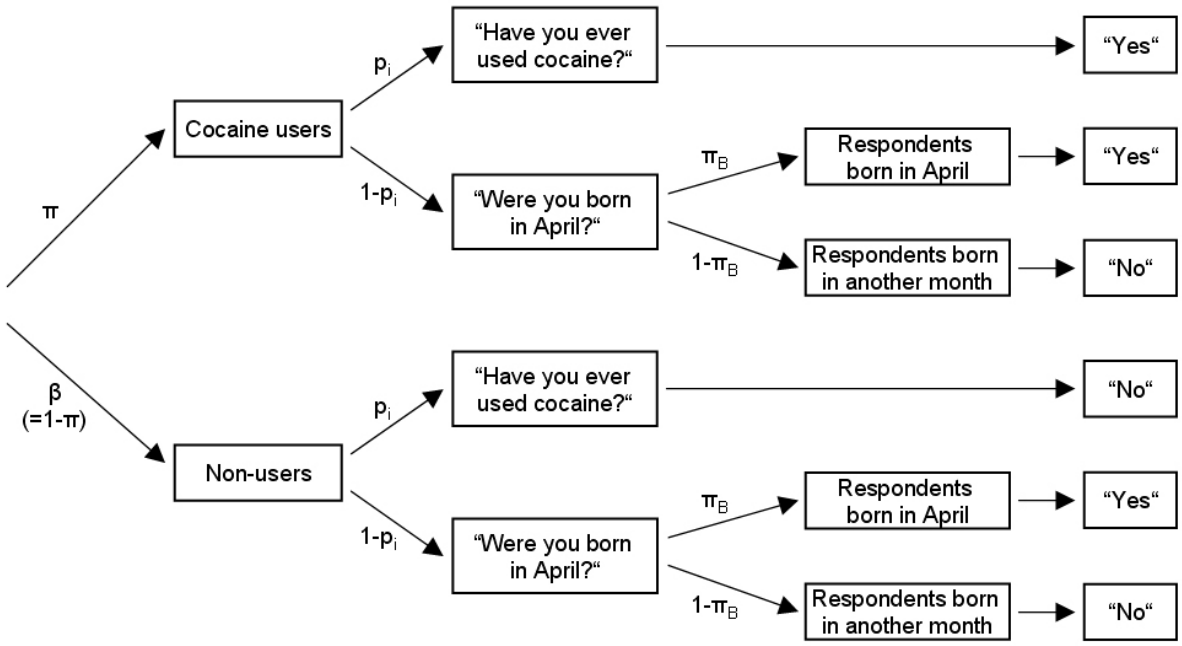

π
β
(=1-π)
Cocaine users
Non-users
$p_i$
1-$p_i$
"Have you ever used cocaine?"
"Were you born in April?"
$\pi_B$
1-$\pi_B$
Respondents born in April
Respondents born in another month
"Yes"
"No"

Figure 4

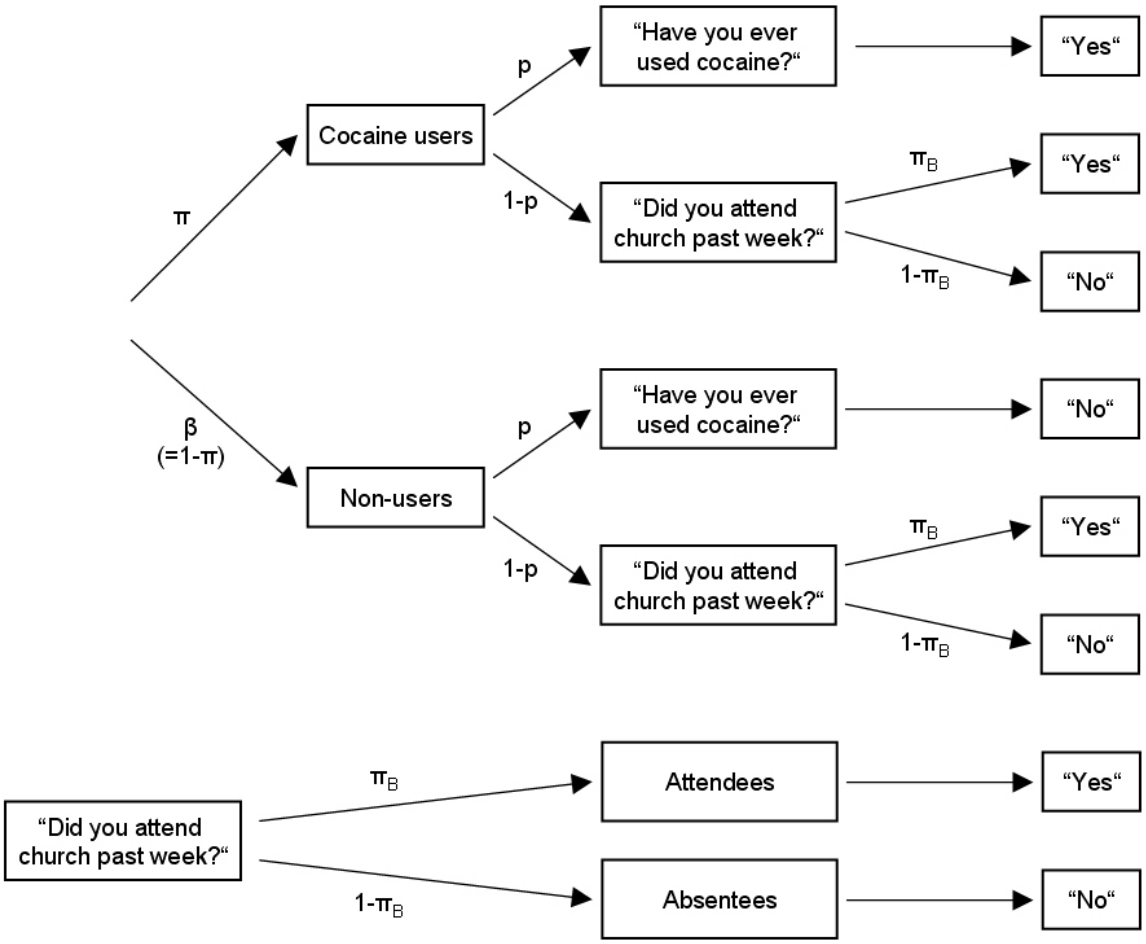

"Have you ever used cocaine?"
"Yes"
p
Cocaine users
$\pi_B$
"Yes"
1-p
"Did you attend church past week?"
π
$1-\pi_B$
"No"
β
(=1-π)
"Have you ever used cocaine?"
"No"
p
Non-users
$\pi_B$
"Yes"
1-p
"Did you attend church past week?"
$1-\pi_B$
"No"
"Did you attend church past week?"
$\pi_B$
Attendees
"Yes"
$1-\pi_B$
Absentees
"No"

Figure 5

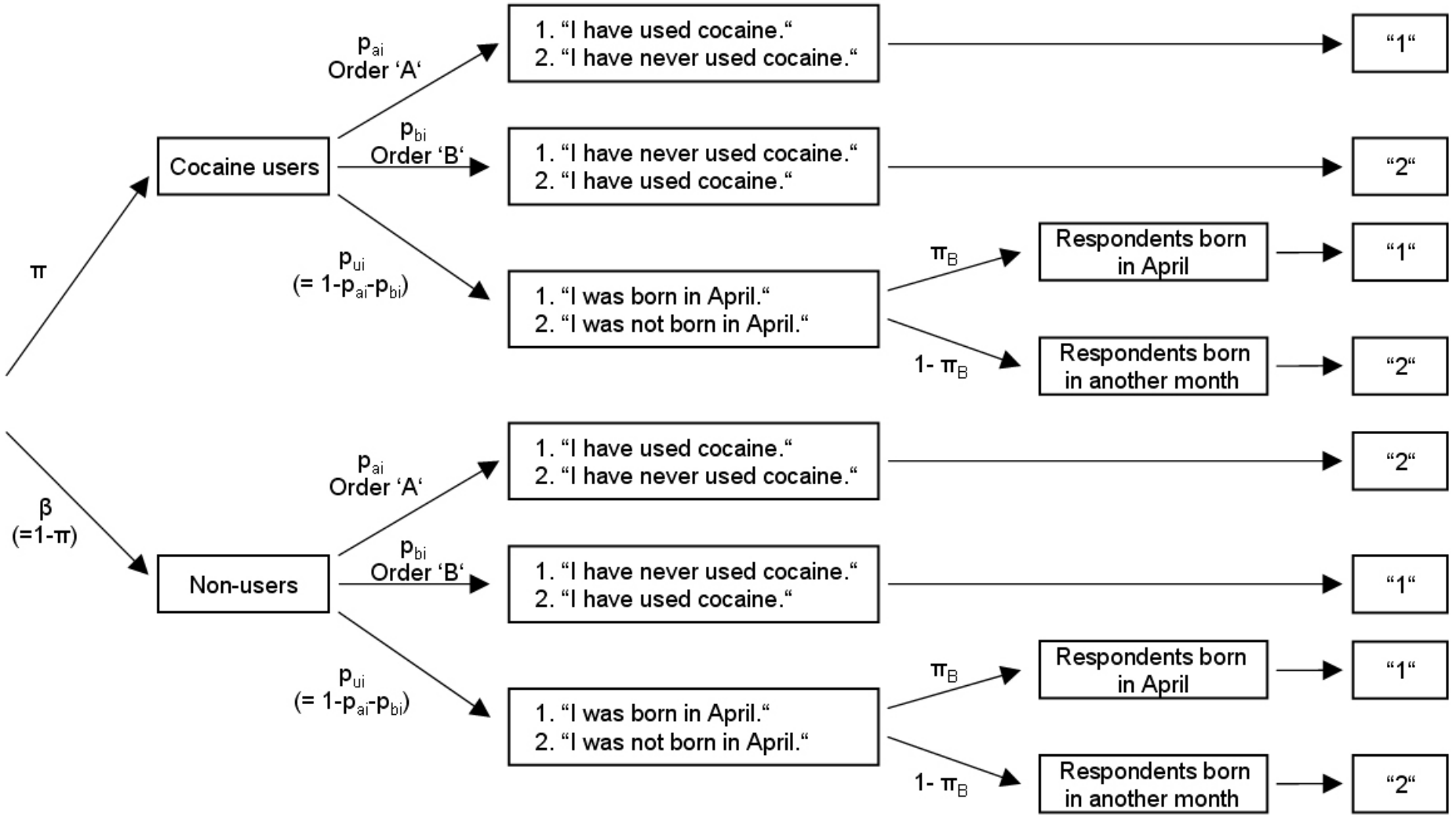

π
β
(=1-π)
Cocaine users
Non-users
p_ai
Order 'A'
p_bi
Order 'B'
p_ui
(= 1-p_ai-p_bi)
1. "I have used cocaine."
2. "I have never used cocaine."
1. "I have never used cocaine."
2. "I have used cocaine."
1. "I was born in April."
2. "I was not born in April."
π_B
1- π_B
Respondents born in April
Respondents born in another month
"1"
"2"

Figure 6

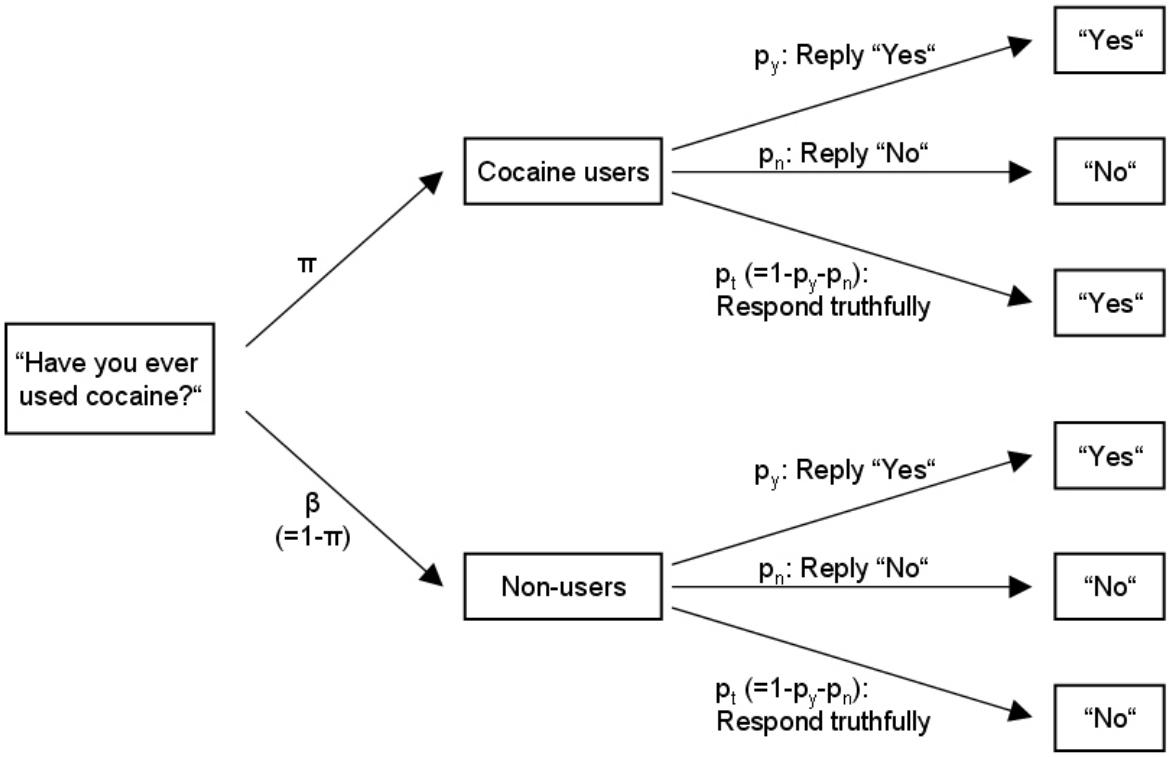

"Have you ever used cocaine?"
π
β
(=1-π)
Cocaine users
Non-users
$p_y$: Reply "Yes"
$p_n$: Reply "No"
$p_t$ (=1-$p_y$-$p_n$): Respond truthfully
$p_y$: Reply "Yes"
$p_n$: Reply "No"
$p_t$ (=1-$p_y$-$p_n$): Respond truthfully
"Yes"
"No"
"Yes"
"Yes"
"No"
"No"

Figure 7

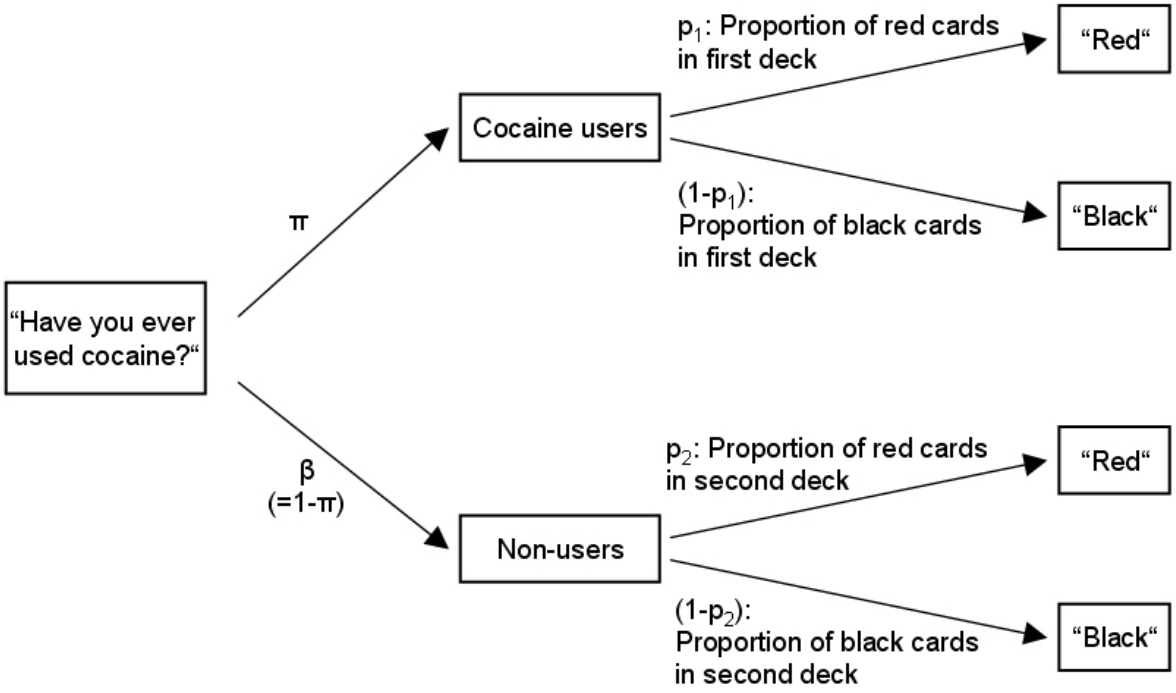

"Have you ever used cocaine?"
π
β
(=1-π)
Cocaine users
Non-users
$p_1$: Proportion of red cards in first deck
(1-$p_1$): Proportion of black cards in first deck
$p_2$: Proportion of red cards in second deck
(1-$p_2$): Proportion of black cards in second deck
"Red"
"Black"
"Red"
"Black"

Figure 8

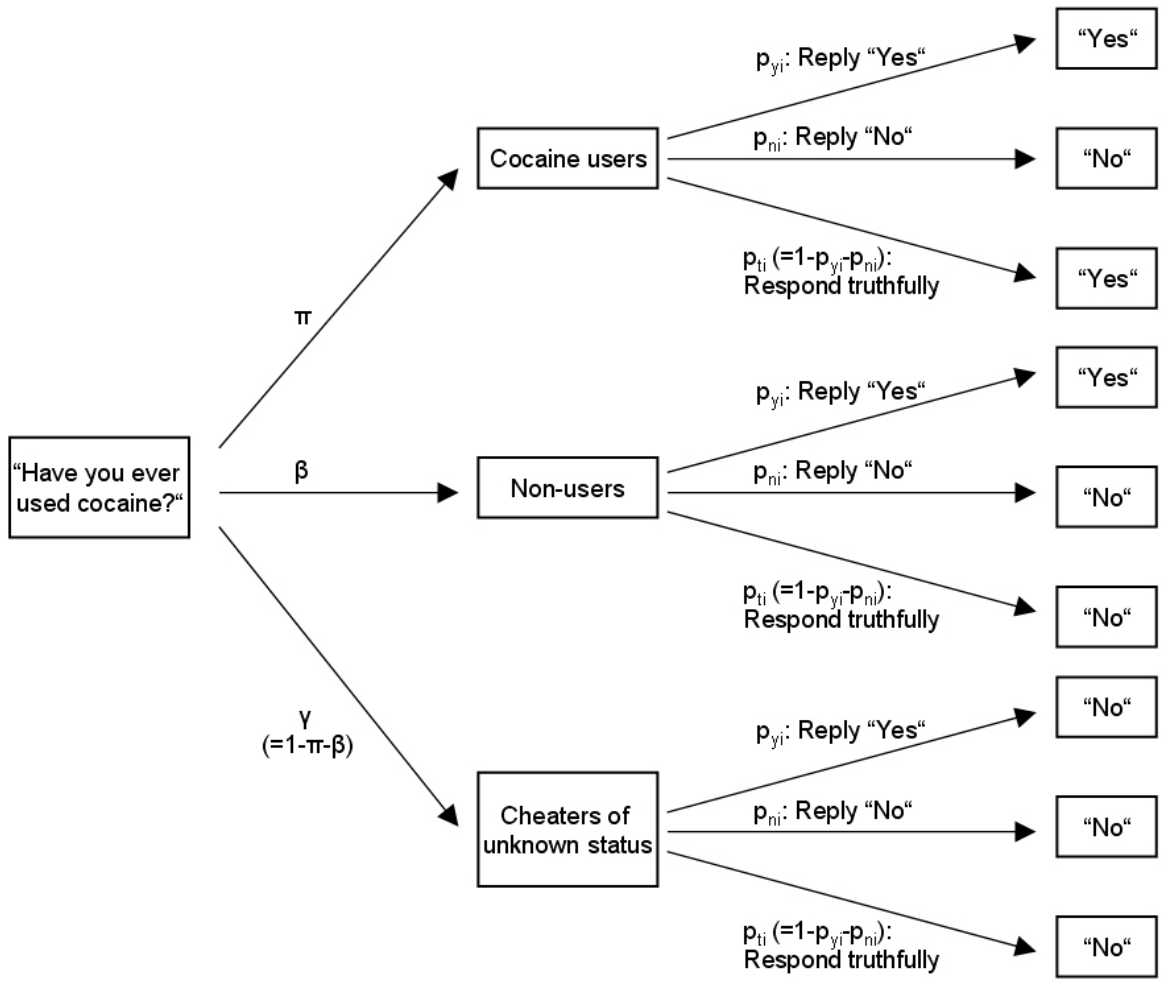
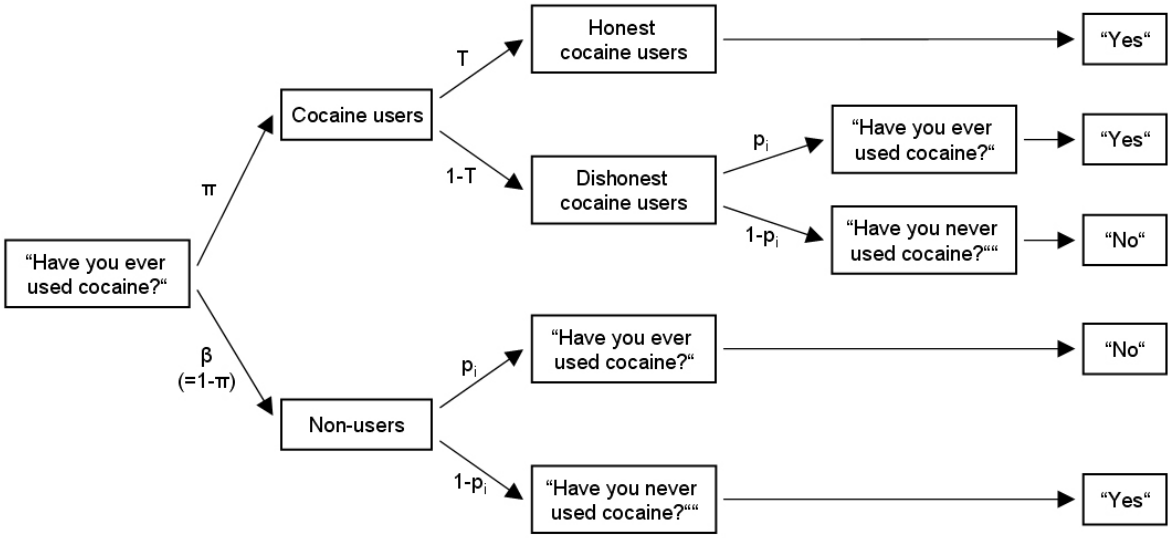

"Have you ever used cocaine?"
π
β
γ (=1-π-β)
Cocaine users
Non-users
Cheaters of unknown status
$p_{yi}$: Reply "Yes"
$p_{ni}$: Reply "No"
$p_{ti}$ (=1-$p_{yi}$-$p_{ni}$): Respond truthfully
"Yes"
"No"
"Yes"
"No"
"No"
"No"
"No"
"No"

Figure 9


"Have you ever used cocaine?"
π
β (=1-π)
Cocaine users
Non-users
T
1-T
$p_i$
1-$p_i$
Honest cocaine users
Dishonest cocaine users
"Have you ever used cocaine?"
"Have you never used cocaine?""
"Have you ever used cocaine?"
"Have you never used cocaine?""
"Yes"
"Yes"
"No"
"No"
"Yes"

# Erklärung

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

Düsseldorf, 13.06.2008 Morten Moshagen