Enhancing the Applicability of Randomized Response Techniques

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Fabiola Reiber
aus Freiburg im Breisgau

Tübingen
2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	03.11.2021
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Rolf Ulrich
2. Berichterstatter:	Prof. Dr. Edgar Erdfelder
3. Berichterstatter:	Prof. Reinhold Kliegl, PhD

Für Hauke

Contents

Summary

Surveys addressing sensitive research topics such as domestic violence or sexist attitudes are subject to self-protecting response biases. Randomized response techniques (RRTs) have been proposed to encourage honest responses to sensitive questions by guaranteeing privacy protection of survey respondents through randomization in the questioning design. Thereby, they aim to increase the validity of estimates of prevalences of sensitive attributes. However, the applicability of RRTs is impaired by a still less than ideal validity of prevalence estimates and high sample size requirements.

In this dissertation, I propose two approaches to enhance the applicability of RRTs. First, I present a testable model that incorporates a parameter measuring non-adherence to instructions in a common variant of the RRT. The results of an empirical study on intimate partner violence indicate that applying this extension enables a more valid description of the mechanisms underlying responses. Second, I propose incorporating RRTs into a sequential hypothesis testing framework using a curtailed sampling plan. Theoretical considerations and first empirical results show that following this approach the sample size requirements of RRTs can be substantially diminished while preserving an easy-to-conduct sampling procedure.

In summary, the proposed procedures can render applications of RRTs more feasible and, thereby, enable insightful future investigations of sensitive research questions.

Zusammenfassung

Umfragen zu sensiblen Themen, wie zum Beispiel häuslicher Gewalt oder sexistischen Einstellungen, unterliegen selbstschützenden Antworttendenzen. Randomized Response Techniken (RRTs) wurden entwickelt, um es den befragten Personen zu erleichtern, ehrlich auf sensible Fragen zu antworten, indem anhand einer Randomisierung im Befragungsdesign ihre Privatsphäre geschützt wird. Als Konsequenz werden validere Schätzungen zur Prävalenz sensibler Eigenschaften erwartet. Allerdings wird die Anwendbarkeit von RRTs davon beeinträchtigt, dass die Validität der Prävalenzschätzungen dennoch nicht optimal ist und sehr große Stichproben benötigt werden.

In dieser Dissertation schlage ich zwei Ansätze zur Verbesserung der Anwendbarkeit von RRTs vor. Als ersten Ansatz stelle ich ein testbares Modell zur Messung von Instruktionsmissachtungen in einer weit verbreiteten Variante der RRT vor. Die Ergebnisse einer empirischen Studie zu Gewalt in Partnerschaften zeigen, dass diese Erweiterung eine validere Beschreibung der Antwortmechanismen ermöglicht. Als zweiten Ansatz schlage ich vor, RRTs in einen Curtailed Sampling Plan zum sequenziellen Hypothesentesten einzubinden. Theoretische Überlegungen und erste empirische Ergebnisse zeigen, dass bei Anwendung dieses einfach durchzuführenden Erhebungsplans die benötigte Stichprobengröße stark reduziert werden kann.

Zusammenfassend können die vorgeschlagenen Verfahren Anwendungen von RRTs erleichtern und dadurch in Zukunft aufschlussreiche Untersuchungen zu sensiblen Forschungsfragen ermöglichen.

Articles

This dissertation is based on three articles, two of which have been published. The third is currently under review. Copies of the three articles are in Appendix B of this dissertation.

In the following, the articles are listed along with statements on the individual contributions of all contributing authors.

Article I

Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods and Research*. Advance online publication. doi: 10.1177/0049124120914919

Author contributions

Author	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Reiber, F.	45%	75%	50%	75%
Pope, H.	5%	0%	10%	5%
Ulrich, R.	50%	25%	40%	20%

Status in publication process: Advance online publication

Article II

Reiber, F., Bryce, D., & Ulrich, R. (in press). Self-protecting responses in randomized response designs: A survey on intimate partner violence during the COVID-19 pandemic. *Sociological Methods and Research*.

Author contributions

Author	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Reiber, F.	33%	40%	70%	80%
Bryce, D.	33%	40%	15%	10%
Ulrich, R.	33%	20%	15%	10%

Status in publication process: Under review

Article III

Reiber, F., Schnuerch, M., & Ulrich, R. (2020). Improving the efficiency of surveys with randomized response models: A sequential approach based on curtailed sampling. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000353

Author contributions

Author	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Reiber, F.	40%	75%	60%	75%
Schnürch, M.	15%	20%	20%	15%
Ulrich, R.	45%	5%	20%	10%

Status in publication process: Advance online publication

1 Introduction

"Have you been physically assaulted by your partner? Do you believe men are better leaders? Have you made false statements on your tax return?" —

Like these, many research questions in the social sciences concern topics of sensitive nature. That is, they concern topics that are perceived as private or incriminating by society. Often, these topics are of high societal relevance, such as, for example, domestic violence, sexist attitudes, or tax fraud. For instance, in the context of the COVID-19 pandemic, research on domestic violence has been intensified, to validate the expectation that the impact of the pandemic and related containment measures would foster risk factors for domestic violence (Usher, Bhullar, Durkin, Gyamfi, & Jackson, 2020). Indeed, an increase in police records and helpline calls was registered for the year 2020 in many countries, leading to action appeals to policy makers (e.g., Bradbury-Jones & Isham, 2020; Jarnecke & Flanagan, 2020). However, the extent of the problem might still be underestimated by these numbers because there is expected to be a high number of unreported cases (i.e. a high *dark figure*; e.g., Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001; Gracia, 2004). To investigate this dark figure researchers rely on self-reports. Furthermore, there are sensitive research topics, such as sexist attitudes, for which there are no objective data sources at all and self-reports are the only available data source.

However, self-reports are subject to response biases and this problem is especially pronounced in the context of sensitive research questions (see Tourangeau & Yan, 2007). Sensitive research questions are defined by being intrusive, elicit threat of disclosure, or address socially undesirable characteristics. This can lead to decreased survey response rates, non-response to specific sensitive questions, or under- or overreporting in response to these sensitive questions. As a consequence, prevalence estimates of sensitive characteristics are biased. This means that societal problems, like for example domestic violence, are likely to be underestimated in self-report surveys.

However, the extent of response biases can be moderated by certain characteristics of the survey. For instance, Tourangeau and Yan (2007) discuss factors of the administration mode that reduce response biases, such as forgiving wording of the sensitive question, a sympathetic interviewer, self-administration of the questionnaire and privacy protection. In this vein, a group of questioning techniques was developed to guarantee the privacy protection of survey respondents, namely *indirect questioning techniques* (see, e.g., Fox,

2016; Chaudhuri & Christofides, 2013). Specifically, in indirect questioning techniqes, single responses are inconclusive with respect to the sensitive attribute, such that no inference about single respondents can be drawn. This way, the respondents' privacy protection is guaranteed. There are different types of indirect questioning techniques, a commonly applied subgroup of which are so-called *randomized response techniques*.

1.1 Randomized Response Techniques

The original randomized response technique (RRT) was developed in the 60s (Warner, 1965). In surveys applying this technique, respondents are randomly assigned to one of two questions using some type of randomization device, such as dice or a deck of cards. One asks whether they carry the sensitive attribute (e.g. "Have you been physically assaulted by your partner?") and the other whether they do *not* carry the sensitive attribute (e.g. "Have you not been physically assaulted by your partner?"). Importantly, the random assignment takes place covertly. Therefore, only the respondents themselves know the outcome of the randomization and, therefore, which question they are responding to. Consequently, a "Yes"-response can, for example, either mean "Yes, I have been physically assaulted by my partner" or "Yes, I have not been physically assaulted by my partner". This way, the single respondents' privacy is protected.

Following the original proposition of the RRT, many variants were devised. They differ in the concrete setup of the procedure, that is, the exact allocation to questions and types of alternatives (e.g., Boruch, 1971; Greenberg, Abul-Ela, Simmons, & Horvitz, 1969; for overviews, see, Fox, 2016; Chaudhuri & Christofides, 2013). For example, in the *unrelated question model* (UQM, Greenberg et al., 1969) version of the RRT, the alternative to the sensitive question S is not the reversed sensitive question ¬S but an unrelated, compelety neutral question N, such as, "Is your mother's birthday in the first half of the year?". Like in the original RRT, participants are allocated to one of the questions by a randomization procedure. For example, they are instructed to respond to the sensitive question S, if a die comes up one through four and to the neutral question N, if it comes up five or six. Importantly, again, the outcome of the randomization is only known to the respondents themselves. Therefore, a "Yes"-response can either mean "Yes, I have been physically assaulted by my partner" or "Yes, my mother's birthday is in the first half of the year". Like in the original RRT, single respondents' privacy is therefore protected. Furthermore, because the alternative question is unrelated to the sensitive attribute, it is straightforward that some responses have nothing to do with the sensitive attribute.

Nevertheless, the prevalence of the sensitive attribute can be estimated from the proportion of "Yes"-responses and the known probabilities underlying the randomization procedure. Figure 1 depicts the probabilities to respond "Yes" or "No" in the UQM. A

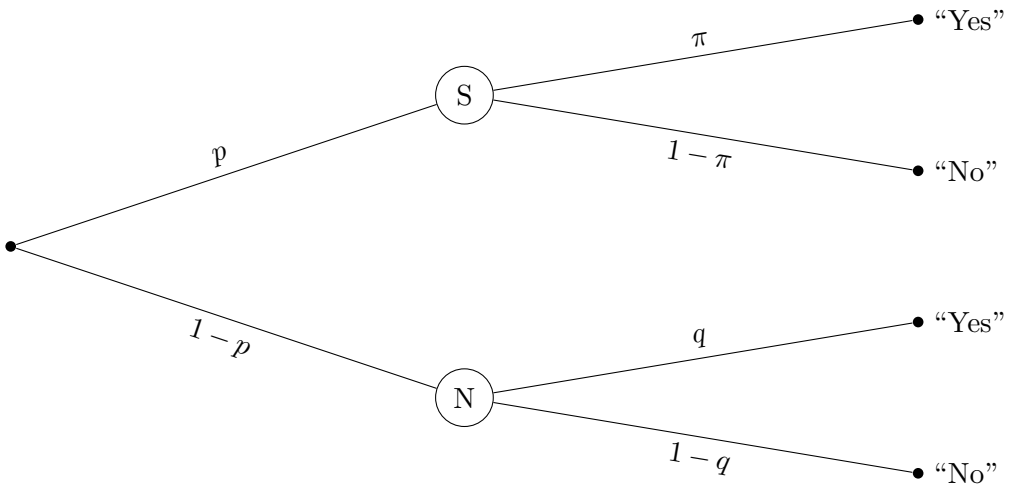


Figure 1: Probability tree of the UQM. Respondents are randomly allocated to respond to the sensitive question S or the neutral question N with probability p and $1 - p$, respectively. The probabilities of responding “Yes” and “No” to the neutral question N are q and $1 - q$ and the probabilities of responding “Yes” and “No” to the sensitive question S are π and $1 - \pi$. Adapted from “Cheater detection using the unrelated question model” by F. Reiber, H. Pope, and R. Ulrich, 2020, *Sociological Methods and Research*, advance online publication, p. 3, https://doi.org/10.1177/0049124120914919 published by SAGE Publishing under the terms of Creative Commons Attribution 4.0.

“Yes”-response can either come from a respondent who was instructed to respond to the sensitive question S with probability p and who carries the sensitive attribute with probabilty π, or a respondent who was instructed to respond to the neutral question N with probability $1 - p$ and who carries the neutral attribute with probability q. Therefore, the overall probability of a “Yes”-response is

$$\lambda_{\text{UQM}} = p \cdot \pi + (1 - p) \cdot q. \tag{1.1}$$

The parameters p and q are given by the questioning design and are therefore known. In the current example, p is the probability that a die comes up one through four, that is, .67. The prevalence of the neutral attribute q is the probability of a birthday being in the first half of a year, that is, about .50. The overall probability λ of a “Yes”-response can be estimated from the proportion of “Yes”-responses in a sufficiently large sample. Thus, Equation 1.1 can be rearranged for the prevalence π of the sensitive attribute of interest, yielding the estimate (see Greenberg et al., 1969)

$$\hat{\pi}_{\text{UQM}} = \frac{\hat{\lambda}_{\text{UQM}} - (1 - p) \cdot q}{p}. \tag{1.2}$$

Although different versions of the RRT differ in the concrete implementation, all follow the same general logic. Privacy protection is created using some sort of randomization

Table 1: Exemplary RRT applications in psychology and related fields

Topic	Study	N
Induced abortion	Abernathy, Greenberg, & Horvitz, 1970	2,871
Rape victimization	Soeken & Damrosch, 1986	368*
Employee theft	Wimbush & Dalton, 1997	196
Job applicant faking	Donovan, Dwight, & Hurtz, 2003	221
Xenophobia	Ostapczuk, Musch, & Moshagen, 2009	606
Corruption	Gingerich, 2010	2,859
Dental hygiene	Moshagen, Musch, Ostapczuk, & Zhao, 2010	2,254
Poaching	Razafimanahaka et al., 2012	1,851
Cognitive enhancement	Dietz et al., 2013	2,557
Academic misconduct	Hejri, Zendehdel, Asghari, Fotouhi, & Rashidian, 2013	144
Organized crime	Wolter & Preisendörfer, 2013	333
Physical doping	Ulrich et al., 2018	2,168*
Prejudice against women leaders	Hoffmann & Musch, 2019	721

Note. This table contains exemplary studies applying RRTs to investigate various sensitive topics. It serves to demonstrate the application range and does not comprise an exhaustive literature review. *N*: Total size of the sample administered for the respective question using the RRT. * These samples consist of subsamples that were analyzed separately. From "Improving the efficiency of surveys with randomized response models: A sequential approach using curtailed sampling." by F. Reiber, M. Schnuerch, and R. Ulrich, 2020, *Psychological methods*, advance online publication, p. 8, https://doi.org/10.1037/met0000353. Copyright 2020 by the American Psychological Association. Adapted with permission.

in the questioning design. Therefore, respondents have less reason to give self-protecting responses and, consequently, they respond more honestly. This, in turn, leads to more valid prevalence estimates. In fact, a meta-analysis (G. J. Lensvelt-Mulders, Hox, Van Der Heijden, & Maas, 2005) of validations studies showed that RRTs elicit estimates that are both less socially desirable and closer to known true prevalences. RRTs have been applied to various topics. An excerpt of applications is presented in Table 1.

However, despite the theoretical effort put into model development and the general empirical efficacy, RRT applications are rather scarce (Blair, Imai, & Zhou, 2015). This is not too surprising because the validity of RRTs is less than ideal. The beforementioned meta-analysis showed that RRTs yield more valid estimates only in certain cases (G. J. Lensvelt-Mulders et al., 2005). Moreover, there is evidence that RRTs, too, are subject to serious response biases (see John, Loewenstein, Acquisti, & Vosgerau, 2018). In other words, prevalence estimates from surveys applying RRTs can be and often are biased due to instruction non-adherence. However, RRT applications are motivated by the aim to elicit honest responses to sensitive questions. Because there are reasons to doubt this characteristic, researchers can be discouraged to invest the extra effort applying the more cumbersome RRT.

This is especially relevant in light of the fact that RRT applications are associated with high costs. The random noise, which creates privacy protection, induces uncertainty in the estimates and, to compensate for that, very large sample sizes are required (Ulrich, Schröter, Striegel, & Simon, 2012). In combination with the doubts concerning instruction adherence it is to be expected that researchers often do not want to invest in RRT applications.

To summarize, by guaranteeing the privacy protection of respondents, RRTs have a high potential to elicit valid prevalence estimates in investigations of sensitive research questions. However, their applicability is impaired by certain restrictions.

1.2 Objective

Therefore, the aim of this dissertation was to increase the applicability of RRTs following two routes. First, to increase the validity of RRT estimates, a model that makes non-adherence to instructions measurable was developed. Second, the RRT was combined with a sequential sampling approach to decrease sample size requirements. In the following, both appoaches are described in more detail. The theoretical foundations and first empirical results are reported.

2 Problem I: Instruction Non-adherence

As mentioned above, although RRT estimates have been shown to often be more valid than estimates from direct questioning, their validity is still less than ideal (see John et al., 2018). Prevalences are still underestimated, corroborating the assumption that there is instruction non-adherence in RRTs. This instruction non-adherence is possibly due to the complicated instructions of RRTs, which lead to impaired understanding of the procedure and therefore a lack of trust in its mechanism to provide privacy (Landsheer, Van Der Heijden, & Van Gils, 1999; Hoffmann, Waubert de Puiseau, Schmidt, & Musch, 2017). John et al. (2018) argue that respondents are afraid that certain responses will be interpreted as admissions to carrying the sensitive attribute, despite the fact that there is no definitive link between the response and the sensitive attribute. Krumpal and Voss (2020) even propose that this is rational because the conditional probabilities of being a carrier given a specific response differ between response options. For example, in the UQM, the conditional probability of being a carrier is lower given a "No"-response than given a "Yes"-response.[1] The authors conclude that giving self-protecting responses can be seen as rational behavior even in RRTs. Problematically, such self-protecting responses distort the resulting RRT prevalence estimates.

To address self-protecting responses within RRTs, extensions measuring the extent of such behavior have been developed. These models include the *cheater detection model* (CDM, Clark & Desharnais, 1998), the *stochastic lie detector* (Moshagen, Musch, & Erdfelder, 2012), and the *extended crosswise model* (Heck, Hoffmann, & Moshagen, 2018). Of these, the CDM has been applied most frequently (e.g., Elbe & Pitsch, 2018; Moshagen et al., 2010; Ostapczuk, Musch, & Moshagen, 2011; Ostapczuk, Moshagen, Zhao, & Musch, 2009; Pitsch, Emrich, & Klein, 2007; Schröter et al., 2016). It is based on the *forced response method* variant of the RRT (Boruch, 1971), in which respondents are either instructed to respond honestly to a sensitive question or simply respond "Yes" depending on the outcome of a randomization procedure. The main assumption of the CDM is that only part of the respondents adhere to these instructions and some respondents instead always give a self-protecting "No"-response to rule out being perceived as a carrier of the sensitive attribute. In the above mentioned applications of the CDM, substantial proportions of respondents of the latter group, termed *cheaters*, were observed (Elbe & Pitsch,

[1] Of course, it is unclear whether respondents are aware of these conditional probabilities and base their response behavior on them.

2018; Moshagen et al., 2010; Ostapczuk et al., 2011; Ostapczuk, Moshagen, et al., 2009; Pitsch et al., 2007; Schröter et al., 2016).

However, the forced response method has been found to elicit less valid estimates compared to other RRTs and evoke response reluctance (Coutts & Jann, 2011; Höglinger, Jann, & Diekmann, 2016; G. J. L. M. Lensvelt-Mulders & Boeije, 2007). Therefore, we proposed to transfer the CDM's concept of cheating to another, more valid RRT, and developed the *unrelated question model - cheating extension* (UQMC, Reiber, Pope, & Ulrich, 2020).

2.1 Unrelated Question Model - Cheating Extension

Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods and Research*. Advance online publication. doi: 10.1177/0049124120914919

The UQMC is based on the standard design of the UQM but incorporates the cheating concept of the CDM. As such, a part of the respondents is expected to respond honestly to the UQM's instructions and another part, the cheaters, is expected to always respond with a self-protecting "No". Figure 2 depicts the probabilities underlying responses in the UQMC. Some respondents cheat, with probability γ, and always respond "No" irrespective of the question they are allocated to and of whether they carry the respective attribute or not. The rest of the respondents responds according to the UQM's instructions with probability $1 - \gamma$. If there is substantial cheating and the standard UQM is applied for estimation, the prevalence of the sensitive attribute is underestimated.

Using two independent samples with varying randomization probabilities p_i, the prevalence of cheating γ can be estimated in addition to the prevalence of the sensitive attribute. Importantly, following the logic of the CDM, the overall prevalence of the sensitive attribute cannot be estimated because the true status of cheaters cannot be inferred. Instead, the prevalence of honest carriers π_{UQMC}, that is, the joint probability of not being a cheater $1 - \gamma$ and of carrying the sensitive attribute ϵ is estimated. Using the two estimates $\hat{\gamma}$ and $\hat{\pi}_{UQMC}$, a lower and upper bound to the estimate of the prevalence of the sensitive attribute can be determined. The lower bound, that is, the estimate if none of the cheaters were carriers is denoted by $\hat{\pi}_{UQMC}$. The upper bound, that is, the estimate if all cheaters were carriers is denoted by $\hat{\pi}_{UQMC} + \hat{\gamma}$. This range provides information about some of the uncertainty in the data, which is ignored in the standard UQM.

However, the UQMC still makes quite strong assumptions about response behavior. For instance, it assumes that the different randomization probabilites p_i do not influence

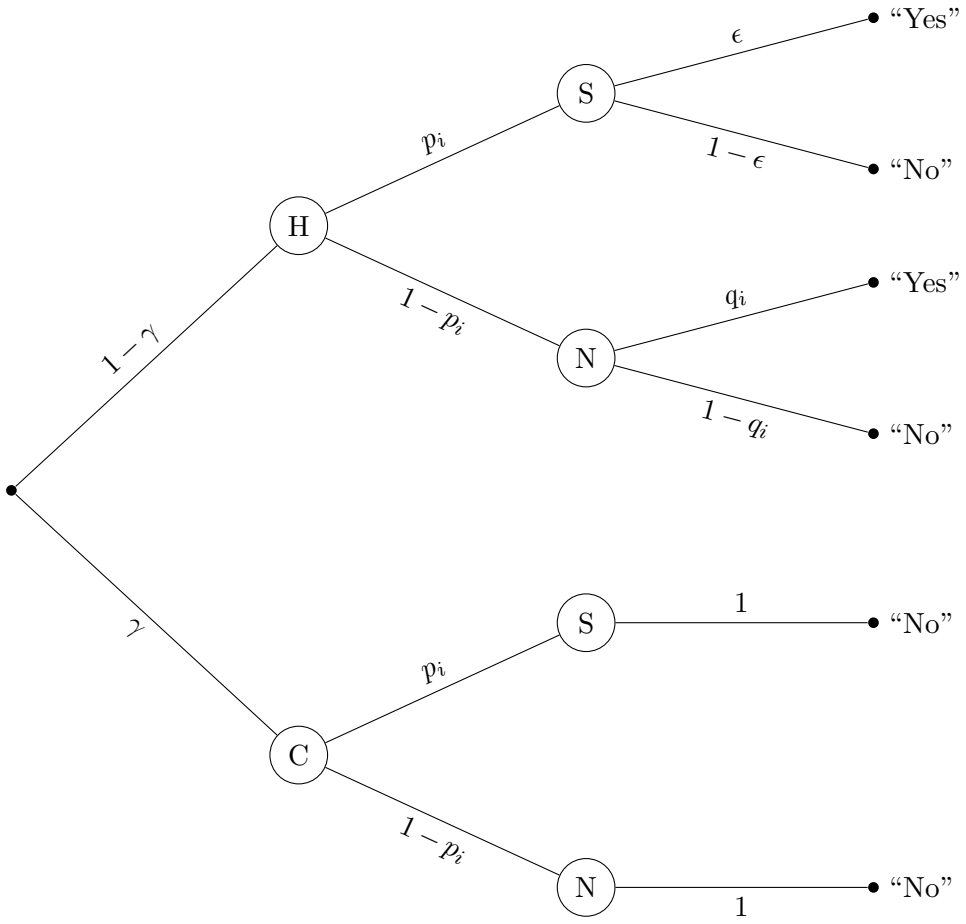


Figure 2: Probability tree of the UQMC. The prevalence of cheaters C is γ and the prevalence of honest participants H is $1 - \gamma$. Both types of respondents are allocated to respond to the sensitive question S and the neutral question N with probability p_i and $1 - p_i$, respectively. The model assumes that cheaters always respond "No" regardless of the question received. Honest participants respond "Yes" with probability q_i and "No" with probability $1 - q_i$ if instructed to answer the neutral question N. They answer "Yes" with probability ϵ and "No" with probability $1 - \epsilon$, if instructed to answer the sensitive question S. Thus, there are three groups of participants: (a) honest participants who are carriers of the sensitive attribute, who will respond "Yes" with probability $(1 - \gamma) \cdot \epsilon = \pi$ if they are allocated to S; (b) honest non-carriers of this attribute who will respond "No" with probability $(1 - \gamma) \cdot (1 - \epsilon)$ if they are allocated to S; and (c) cheaters, who will respond "No" with probability γ regardless of whether they are allocated to S or N. Adapted from "Cheater detection using the unrelated question model" by F. Reiber, H. Pope, and R. Ulrich, 2020, *Sociological Methods and Research*, advance online publication, p. 8, https://doi.org/10.1177/0049124120914919 published by SAGE Publishing under the terms of Creative Commons Attribution 4.0.

response behavior.[2] Therefore, it makes sense to test the assumptions of the UQMC empirically. It is possible to test the model fit of the UQMC using a four-sample extension. Specifically, by varying the prevalence of the neutral attribute q_i in addition to the randomization probability p_i, four indepent samples can be assessed. Consequently, there are four independent response categories instead of two and the resulting extra degrees of freedom enable testing the model fit. This testable version of the UQMC was applied in an empirical study to validate the model and its assumptions.

2.2 Validation of the Unrelated Question Model - Cheating Extension

Reiber, F., Bryce, D., & Ulrich, R. (in press). Self-protecting responses in randomized response designs: A survey on intimate partner violence during the COVID-19 pandemic. *Sociological Methods and Research*.

The validation study was conducted in the context of a large scale online survey on the prevalence of physical intimate partner violence (IPV) during the first contact restrictions due to the COVID-19 pandemic in Germany in spring and early summer 2020. To test the UQMC's assumptions, the four sample version was applied and, additionally, the question sensitivity was manipulated between two conditions.

Physical IPV, that is, behaviors such as hitting, slapping, or shoving a current or former romantic partner, is a highly stigmatized topic (Birkel & Guzy, 2015; Franke, Seifert, Anders, Schröer, & Heinemann, 2004). Therefore, questions about experiencing IPV are sensitive questions making the application of an RRT recommendable. We additionally manipulated the question sensitivity by varying the queried role between participants. They were either queried about victimization or perpetration of IPV. Because perpetration of IPV can have legal consequences and it has been found to have an even stronger association with social desirability (Sugarman & Hotaling, 1997), we expected this question to be the more sensitive one and therefore elicit more cheating. Because participation was restricted to persons in a relationship with one partner, the true prevalence of IPV perpetration and victimization was assumed to be the same. Therefore, any differences in the estimates of the honest carrier prevalence were expected to result from complementary differences in cheating.

The fit test indicated an overall good model fit of the UQMC to the data. Importantly, this did not hold for the standard UQM, which does not account for cheating. Thus, including a cheating parameter enabled a better description of the data. However, contrary

[2] Dietz et al. (2018) found a non-significant difference between UQM estimates from conditions applying different randomization probabilities. This, however, might be due to a lack of power.

to our expectation, cheating was estimated to be higher in the victimization condition and the honest carrier and cheater prevalences were not complementary. Therefore, apparently, there were more factors that influenced estimation but were not accounted for by the UQMC. Possible influencing factors are selective sampling, differences in the perception of violent events between perpetrators and victims of IPV (see Follingstad & Rogers, 2013), or additional types of instruction non-adherence.

2.3 Discussion

The UQMC was developed to account for instruction non-adherence within the UQM. Indeed, it yielded better interpretable results in the validation study than the standard UQM. Still, the experimental manipulation of the question sensitivity disclosed inconsistencies. Specifically, parts of the results were not explainable by the model. It is important to mention that the UQMC only accounts for one specific type of instruction non-adherence, that is, always responding "No" irrespective of the question one is assigned to and the carrier status. However, there are other conceivable types of non-adherence. For example, in Reiber, Pope, and Ulrich (2020) we discuss the possibility of so-called *partial cheaters*, who would respond honestly if they were allocated to the neutral question but cheat if they were allocated to the sensitive question. Such a response style is not detectable in a fit test, because it is mathematically consistent with the UQMC. However, it would influence the interpretation of the estimates. It could, in theory, explain the unexpected data pattern in the validation study. Thus, the UQMC does potentially not offer an exhaustive description of all possible response styles. However, this is true for all models, which are simplifications of the more complex subject. Therefore, the application of the UQMC is nevertheless recommendable, because it accounts at least for one prevalent type of instruction non-adherence.

Despite offering a more refined desciption of the data, models accounting for instruction non-adherence have one general disadvantage compared to conventional RRTs. They incorporate even higher sample size requirements. The validation study required responses from about 3 000 participants after data exclusion. This is a sample size that can often not be accomplished. In other words, the extra information yielded by the cheating parameter comes at a cost in terms of sample size.

3 Problem II: Sample Size Requirements

The fact that extra information comes at cost in terms of sample size is true for RRTs in general. The privacy protection, which is meant to increase data quality, has to be compensated with sample size (Ulrich et al., 2012). Specifically, the randomization, which induces privacy, adds random noise to the data. Therefore, this crucial element of RRTs decreases sampling efficiency. To counterplay this drawback, which is design inherent, huge samples are required. As a consequence, RRT applications are very cost intensive and this arguably discourages their realization.

This holds true for research aiming at precise prevalence estimates of sensitive attributes as well as studies testing hypotheses about these prevalences (Ulrich et al., 2012). Although most studies applying the RRT entail prevalence estimation, often the underlying research questions call for hypothesis testing. For instance, many validation studies rely on the *more-is-better* validation criterion (see G. J. Lensvelt-Mulders et al., 2005). This criterion is based on the assumption that the prevalence estimate of a socially undesirable attribute is more valid if it is higher. Thus, RRTs are concluded to be more valid if RRT estimates exceed estimates from direct questioning (e.g. Nordlund, Holme, & Tamsfoss, 1994; Wimbush & Dalton, 1997; Wolter & Preisendörfer, 2013). The straightforward approach to this research question would be a hypothesis test (as in Hoffmann & Musch, 2016). However, also for studies applying the RRT to test hypotheses, power analyses indicate very high sample size requirements (Ulrich et al., 2012).

A general approach to decrease sample size requirements in any type of study is *sequential sampling*. The logic underlying all sequential sampling schemes is to not sample a pre-specified number of observations but to stop sampling as soon as sufficient information is available (Wetherill, 1975). In case of hypothesis testing, sufficient information can mean sufficienty small long term error rates (Neyman & Pearson, 1933). Specifically, in a classical Neyman-Pearson hypothesis test for binomial data, such as "Yes" vs. "No" responses, the number of collected responses N is pre-specified based on the desired long term error rates. After collecting all data, the number of successes is compared to a criterion c. Based on whether this criterion is reached, a decision with respect to the hypotheses is made with control over the long term error rates. In contrast, in sequential sampling, the sample size is not pre-specified. There are several sequential sampling schemes for this purpose (see Wetherill, 1975). A very simple one of these is *curtailed sampling*.

3.1 Curtailed Sampling for RRTs

Reiber, F., Schnuerch, M., & Ulrich, R. (2020). Improving the efficiency of surveys with randomized response models: A sequential approach based on curtailed sampling. *Psychological Methods*. Advance online publication. doi: 10.101037/met0000353

The logic of curtailed sampling is very close to that of a classical Neyman-Pearson test. The same parameters, N and c, are determined before data collection. However, instead of always collecting N responses, sampling can be stopped earlier according to stopping rules. Specifically, sampling is stopped as soon as (a) c successes are observed or (b) $N - c + 1$ failures are observed because at this point c successes can no longer be observed within N responses (see Wetherill, 1975). Thus, the number of responses becomes a random variable with a maximum of N responses but an expectation below N.

It is straightforward to combine this simple sequential sampling plan with RRT applications. In RRT applications, successes are "Yes"-responses and failures are "No"-responses. Importantly, however, due to the randomization, the hypotheses concerning the prevalence of the sensitive attribute are not directly linked to the responses. Therefore, the hypothesized prevalence values need to be transformed to probabilities of "Yes"-responses using the known randomization probabilities. In case of an application of the UQM, for instance, this is done using Equation 1.1.

To demonstrate, the solid curve in Figure 3 depicts the probability of accepting the null hypothesis as a function of the true prevalence π for a curtailed sampling plan testing the following hypotheses:

$$H_0 : \pi \leq \pi_0 = .05$$
$$H_1 : \pi \geq \pi_1 = .15$$

The error probabilities α and β are .05, such that, when $\pi = \pi_0$, the probability to accept H_0 is .95 and when $\pi = \pi_1$, the probability to accept H_0 is .05. In a direct question survey, the sampling plan is based directly on these hypotheses.

In an RRT survey, however, the hypotheses on the prevalence need to be transformed to hypotheses on the probability of a "Yes"-response first. Inserting π_0 and π_1 into Equation 1.1 yields:[3]

$$H_0 : \lambda \leq \lambda_0 = .25$$
$$H_1 : \lambda \geq \lambda_1 = .29$$

The dotted curve in Figure 3 depicts the probability to accept the null hypothesis as a function of the probability of a "Yes"-response λ. Because λ_0 and λ_1 are closer together

[3]For this example, standard design parameters $p = .75$ and $q = .70$ were used.

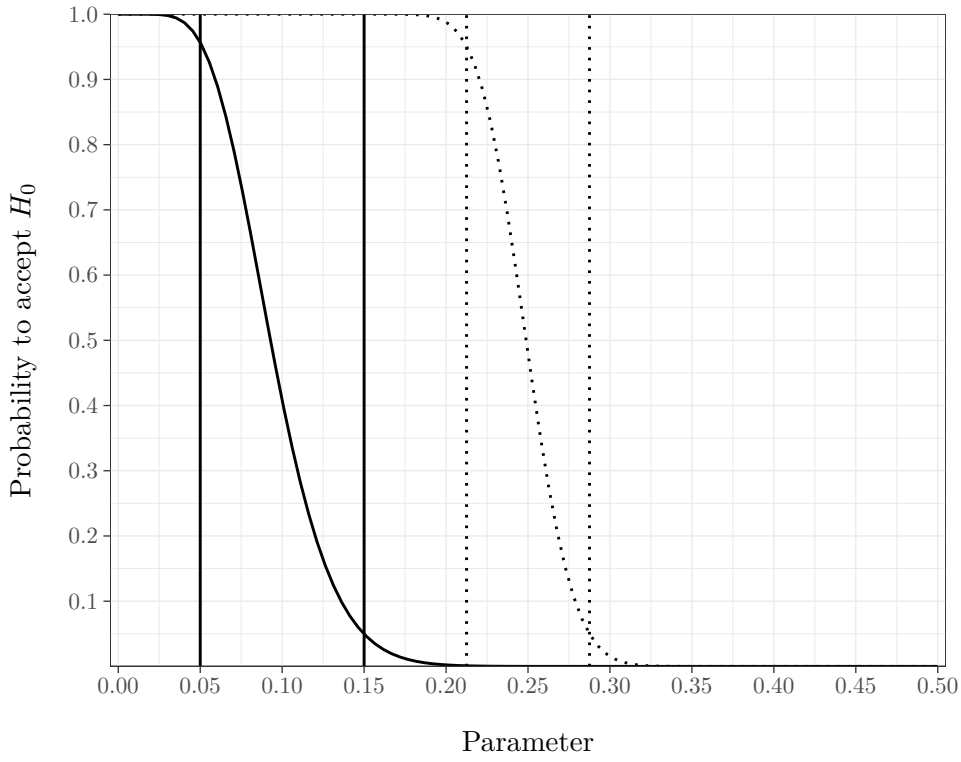


Figure 3: Operating characteristic curve of a curtailed sampling plan. The curves depict the probability of accepting the null hypothesis as a function of the true parameter value for a test of the hypotheses $H_0 : \pi \leq \pi_0 = .05$ vs. $H_1 : \pi \geq \pi_1 = .15$ with $\alpha = \beta = .05$. The solid curve is based directly on the prevalence π. The dotted curve is based on the probability of a "Yes"-response λ in a UQM design with $p = .75$ and $q = .70$. The vertical lines mark the hypothesized values.

than π_0 and π_1 but α and β are held constant at .05, the curve is steeper. In other words the testable hypotheses in RRT applications are stricter, which, again, demonstrates why RRTs are less efficient than direct questions. Therefore, applying a sequential sampling plan becomes even more beneficial in RRT surveys.

The extent of the efficiency gain can be seen in Figure 4. The solid and dotted curves depict the expected sample size of a curtailed sampling plan for the above hypotheses as a function of the true prevalence π in a direct question and an RRT survey, respectively. The horizontal lines depict the maximum sample size N for either. As is to be expected the maximum sample size for the RRT survey is much higher than that of the direct questioning survey. The expected sample size is always lower than the maximum sample size, which equals the pre-specified sample size of a classical Neyman-Pearson test. The possible sample size savings are substantial, especially, when the true prevalence is far from the hypothesized values. Due to the larger maximum sample size in RRTs, the possible savings are even higher in this questioning design.

3.2 Applications

So far, the curtailed sampling plan for RRTs has been applied in two studies and a third is currently in the stage of data collection. The first study was conducted in the context of an unpublished master's thesis (Iberl, 2019). The aim was to replicate the findings of a study on pharmacological neuroenhancement, that is, the use of psychoactive substances with the purpose of improving cognitive or mental performance (Schilling, Hoebel, Müters, & Lange, 2012). The original study (Dietz et al., 2018) investigated pharmacological neuroenhancement among university students and reported a prevalence of 14.9 percent. Thus, the hypotheses in the replication study were:

$$H_0\text{: } \pi \leq \pi_0 = .01$$

The prevalence is lower than in the original study (i.e., nearly absent).

$$H_1\text{: } \pi \geq \pi_1 = .15$$

The prevalence is at least as high as in the original study.

A decision in favor of H_1 was made after reaching the critical number c of "Yes"-responses. At this point the maximum number of responses N was nearly reached. In other words, in this case the application of the curtailed sampling plan did not lead to substantial sample size savings.

The second study was a validation study conducted in the context of an unpublished bachelor's thesis (Hafner, 2019). In a street survey, passers-by were queried about voting in the elections for the European Parliament in 2019 using either the UQM or direct questioning. Because voting is generally perceived as socially desirable (Goerres, 2010), over-reporting in a street survey was expected, but less so using the UQM. The true voter

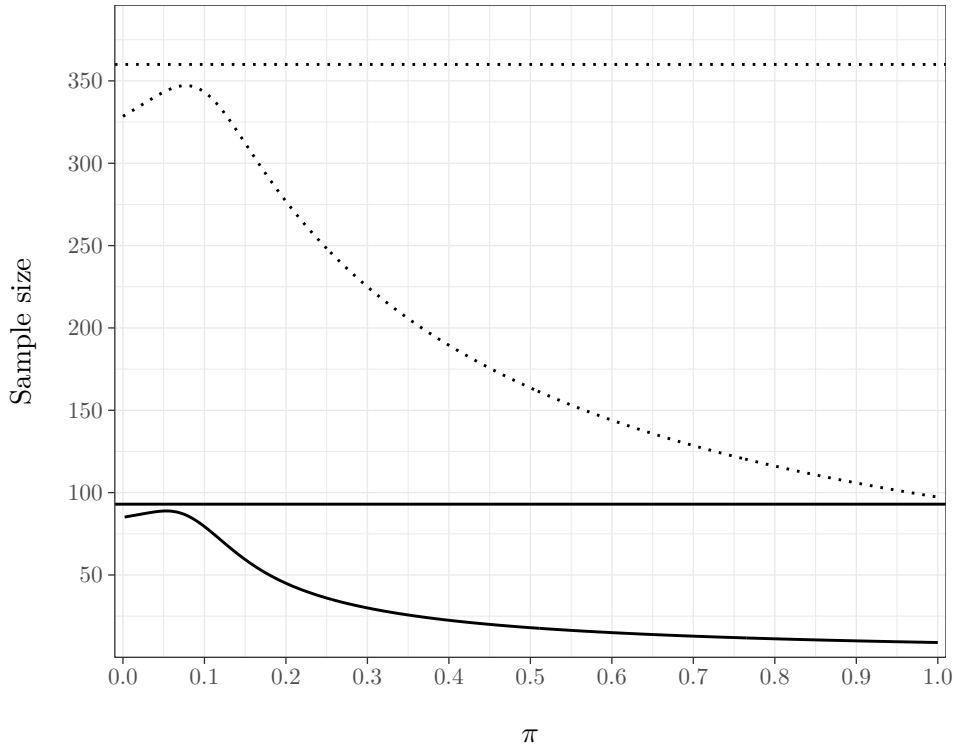


Figure 4: Expected sample size in a curtailed sampling plan. The curves depict the expected sample size in a curtailed sampling plan as a function of the true prevalence π for a test of the hypotheses $H_0 : \pi \leq \pi_0 = .05$ vs. $H_1 : \pi \geq \pi_1 = .15$ with $\alpha = \beta = .05$. The solid curve is based on the curtailed sampling plan for a direct question study. The dotted curve is based on the curtailed sampling plan for a UQM study with $p = .75$ and $q = .70$. The horizontal lines mark the respective maximum sample size N, that is, the pre-specified sample size in a classical Neyman-Pearson test.

turnout was 67.1 percent in the region of survey administration (Stuttgart, Germany). Therefore, the hypotheses were for both questioning techniques,

$$H_0\text{: } \pi \leq \pi_0 = .70$$
Voting is not overreported.
$$H_1\text{: } \pi \geq \pi_1 = .80$$
Voting is overreported.

For both questioning techniques, the decision that voting was overreported was made after reaching the critical number c of "Yes"-responses. Compared to a classical Neyman-Person hypothesis test, the sample size savings were 12.3 percent (of $N = 204$) and 20.9 percent (of $N = 549$) in the direct questioning and UQM conditions, respectively. Thus, in this case, the application of the curtailed smapling plan did lead to substantial sample size savings.

A third study to replicate findings on doping in elite athletics (Striegel, Ulrich, & Simon, 2010) is still in the state of data collection.

3.3 Discussion

The theoretical considerations and first applications demonstrate that curtailed sampling can substantially decrease sample size and thus make RRT applications more feasible. The replication study on pharmacological neuroenhancement shows that there is not always a large improvement but the sample size can never exceed the fixed Neyman-Pearson N. There are other sequential sampling approaches that are on average more efficient. For instance, the *sequential probability ratio test* (SPRT, Wald, 1945) has been proven to be the most efficient test when the true parameter equals or is very close to one of the hypothesized values (i.e., π_0 or π_1; Wald & Wolfowitz, 1948). However, the SPRT incorporates no upper limit to the sample size. Therefore, it is theoretically possible that the sample size becomes extremely large.

A maximum sample size makes studies applying curtailed sampling more plannable. In addition, hypothesis evaluation during sampling is very convenient because researchers only need to count "Yes"- and "No"-responses. In the paper (Reiber, Schnuerch, & Ulrich, 2020), we provide R user scripts and an R shiny web application, to further facilitate study planning and data evaluation.

Another advantage of curtailed sampling compared to other sequential sampling approaches is that it is straightforward to conduct subsequent estimation. Although, as stated above, many RRT research questions call for hypothesis tests, subsequent prevalence estimation can provide further insight. Because sampling is stopped based on the data, conventional maximum likelihood estimators are biased (e.g., Whitehead, 1986). Nevertheless, in case of curtailed sampling, unbiased subsequent estimation is possible

using adjusted *inverse binomial sampling* (Haldane, 1945).

Importantly, however, unbiasedness is not the only criterion for reliable estimation. Another important criterion is precision. However, RRT estimates based on small sample sizes, cannot be precise. In other words, estimation in the context of RRTs is always subject to the trade-off between sample size and precision and curtailed sampling is not designed for precise estimation.

Another limitation of curtailed sampling is that it is restricted to tests of simple hypotheses, like the size of a single prevalence. However, there are conceivable research questions calling for tests of composite hypotheses in the RRT context. For example, it might be of interest whether prevalences of a sensitive attribute differ between subpopulations. Also, whenever an RRT accounting for instruction non-adherence, like the UQMC, is applied, composite hypotheses have to be tested due to the nuisance parameter (e.g., the cheating parameter γ in the UQMC).

The SPRT, on the other hand, can be extended to tests of composite hypotheses. Schnuerch, Erdfelder, and Heck (2020) demonstrate how this is possible using the *sequential maximum likelihood ratio test* (Cox, 1963) in the context of *multinomial processing tree models* (Riefer & Batchelder, 1988) of which the RRT is a special case. An article applying this procedure to RRTs is currently in preparation. This will allow for sequential testing using RRTs measuring instruction non-adherence, such as the UQMC. Consequently, this procedure will enable combining the two approaches for enhancing the applicability of RRTs presented within this dissertation.

4 General Discussion

Investigating sensitive research questions is made difficult by self-protecting response biases of survey respondents. Randomized response techniques (RRTs) provide one way to approach this problem by guaranteeing privacy protection. However, applications of RRTs are impaired by certain restrictions. In the preceding chapters, I have presented two ways to address these restrictions and demonstrated how this can improve RRT applications.

First, I presented the UQMC, a new model to measure a specific type of instruction non-adherence within a standard RRT. The UQMC validation study showed that accounting for instruction non-adherence by means of cheating detection provides a better description of response behavior.

However, the the empirical assessment of the UQMC also indicated that there are more factors influencing responses. There are other types of response styles which are much harder to incorporate in a statistical model. For instance, random responding has been proposed as a factor strongly influencing responses in another popular RRT variant, the *crosswise model* (Yu, Tian, & Tang, 2008). Because of the complicated instructions of RRTs, random responses by respondents who do not understand the instructions might be a severe problem. Such random responding, however, is difficult to model, because estimating randomness is extremely inefficient.

As an alternative to modeling, the low comprehensibility leading to such response styles can be addressed (Meisters, Hoffmann, & Musch, 2020). To this end, simplified instructions and the application of training questions have been suggested (Meisters et al., 2020). Another promising idea, which has to my knowledge not been tested yet, are comprehensive instruction videos, especially in the context of online surveys. Moreover, the RRT should be designed such that the mechanism of the RRT is as intuitive as possible (Höglinger et al., 2016). For example, using dice or a deck of cards as randomization device might be more intuitive than an unrelated question, concerning, for example, the birthday of a close relative. More specifically, to a person not acquainted with probability theory it might not be intuitive that birthdays are randomly distributed. Thus, some respondents might not understand how their privacy is protected by such a randomization procedure and be reluctant to adhere to the instructions. This lack of understandig could either be countered by demonstrating the randomnesss property of birthdays using one of the above mentioned strategies or by using a more intuitive randomization procedure in the first place. As a consequence, not only random responding but also deliberate response

styles based on a lack of trust could be diminished.

These strategies can be devised to decrease deliberate response strategies with the aim of impression management. However, sensitive topics do not only foster impression management strategies but are also subject to unintentional mechanisms such as self-deception, rationalization, and difficulties recalling and reporting unpleasant events (Näher & Krumpal, 2012; Tourangeau & Yan, 2007). Such mechanisms are not under the volitional control of survey respondents and can therefore not be countered by an affirmation of privacy protection. Thus, with respect to reducing these mechanisms, RRTs are naturally restricted.

Second, I suggested sequential testing to ameliorate the problem of high sample size requirements. I demonstrated that the sample size of an RRT study can be substantially decreased in a sequential hypothesis test using curtailed sampling. However, the high sampling variance of RRTs is design inherent and especially when prevalence estimation is the goal of a study this cannot be circumvented.

Therefore, it is reasonable to consider in which cases an RRT application is sensible and to keep alternatives in mind. Like mentioned in the introduction, there are ways to design a study such that honest responding to sensitive questions is facilitated, for example, by self-administration or forgiving wording of the sensitive question (Tourangeau & Yan, 2007). Especially in the context of online surveys, respondents might generally perceive the privacy protection as high enough without extra protection implemented. For example, in an unpublished online study on the perceived privacy protection in direct questioning and different RRT designs with varying randomization probabilities, we observed a ceiling effect.[4] Specifically, the perceived privacy protection was at the top of the scale in all conditions including direct questioning, despite the high sensitivity of the topic (intimate partner violence). It has even been argued that in certain situations the RRT may rather induce a feeling of insecurity by making privacy concerns more salient (see John et al., 2018). Thus, in such studies it can be reasonable to create an evironment that fosters honest responding without applying an RRT.

However, applying RRTs can be very beneficial in certain situations. For example, in face-to-face settings it is much more difficult to create an anonymous environment and due to the presence of an interviewer social desirability becomes even more influencial (Tourangeau & Yan, 2007). Moreover, in face-to-face settings the implementation of intuitive random generators, such as dice or a deck of cards is easily feasible.[5] Additionally, difficulties in understanding the procedure can be more easily ruled out and specific explanations be provided.

[4] A description of the study and the main results is in Appendix C. The ceiling effect with respect to the perceived privacy protection is depicted in Figure 5 of this appendix.

[5] This is more difficult in online studies because one cannot rely on respondents to actually conduct a physical randomization in front of their screens but online tools might not be perceived as trustworthy (Coutts & Jann, 2011).

Another feature making the application of RRTs appropriate is a high sensitivity of the topic, in the sense that it strongly elicits impression management strategies, such as a concrete sensitive behavior with possible legal consequences (e.g., theft or doping in elite athletics). Here, the privacy protection provided by RRTs can elicit more honest responses and lead to more valid prevalence estimates (G. J. Lensvelt-Mulders et al., 2005).

In summary, although RRTs are no panacea for self-protecting response biases in surveys on sensitive attributes, they are a useful tool for specific types of studies. The results presented within this dissertation demonstrate that if an RRT is applied, it is recommendable to use a testable model accounting for instruction non-adherence. Moreover, if the research question implies a hypothesis test, a sequential sampling design can further lower the barrier to apply RRTs.

4.1 Conclusion

Research on sensitive topics often adresses issues of high societal relevance but it is difficult to conduct due to sef-protecting response tendencies in self-reports. Randomized response techniques provide an approach to address this problem by creating privacy protection. However, their empirical applicability is impaired by instruction non-adherence and high sample size requirements. In this dissertation I outlined two routes to increase the applicability of RRTs, namely measuring non-adherence to instructions and sequential hypothesis testing. There is certainly additional work needed to increase instruction adherence in RRTs and alternative ways to facilitate honest responding to sensitive questions have to be considered. However, the presented empirical results show that following these routes is beneficial for RRT applications. Thus, this disseration contributes to increasing the applicability of RRTs for a better understanding of sensitive research topics.

Bibliography

Abernathy, J. R., Greenberg, B. G., & Horvitz, D. G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, *7*, 19 – 29. doi: 10.2307/2060019

Birkel, C., & Guzy, N. (2015). *Viktimisierungsbefragungen in Deutschland* (47.1 ed.; R. Mischkowitz, Ed.). Wiesbaden: Bundeskriminalamt.

Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, *110*, 1304–1319. doi: 10.1080/01621459.2015.1050028

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist*, *6*, 308–311. doi: 10.2307/27701807

Bradbury-Jones, C., & Isham, L. (2020). The pandemic paradox: The consequences of COVID-19 on domestic violence. *Journal of Clinical Nursing*, *29*, 2047–2049. doi: 10.1111/jocn.15296

Chaudhuri, A., & Christofides, T. C. (2013). *Indirect questioning in sample surveys*. Springer Science & Business Media. doi: 10.1007/978-3-642-36276-7

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

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research*, *40*, 169–193. doi: 10.1177/0049124110390768

Cox, D. (1963). Large sample sequential tests for composite hypotheses. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, *25*, 5–12.

Dietz, P., Iberl, B., Schuett, E., Poppel, M. V., Ulrich, R., & Sattler, M. C. (2018). Prevalence estimates for pharmacological neuroenhancement in Austrian university students: Its relation to health-related risk attitude and the framing effect of caffeine tablets. *Frontiers in Pharmacology*, *9*, 1–9. doi: 10.3389/fphar.2018.00494

Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., & Ulrich, R. (2013, jan). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy*, *33*, 44–50. doi: 10.1002/phar.1166

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the preva-

lence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, *16*, 81–106. doi: 10.1207/S15327043HUP1601

Elbe, A.-M., & Pitsch, W. (2018). Doping prevalence among Danish elite athletes. *Performance Enhancement & Health*, *6*, 28–32. doi: 10.1016/j.peh.2018.01.001

Ellsberg, M., Heise, L., Peña, R., Agurto, S., & Winkvist, A. (2001). Researching domestic violence against women: Methodological and ethical considerations. *Studies in Family Planning*, *32*, 1 – 16. doi: 10.1111/j.1728-4465.2001.00001.x

Follingstad, D. R., & Rogers, M. J. (2013). Validity concerns in the measurement of women's and men's report of intimate partner violence. *Sex Roles*, *69*, 149–167. doi: 10.1007/s11199-013-0264-5

Fox, J. A. (2016). *Randomized response and related methods: Surveying sensitive data* (2nd ed.). Thousand Oaks, CA: Sage. doi: 10.4135/9781506300122

Franke, B., Seifert, D., Anders, S., Schröer, J., & Heinemann, A. (2004). Gewaltforschung zum Thema "häusliche Gewalt" aus kriminologischer Sicht. *Rechtsmedizin*, *14*, 193–198. doi: 10.1007/s00194-004-0263-5

Gingerich, D. W. (2010). Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys. *Political Analysis*, *18*, 349–380. doi: 10.1093/pan/mpq010

Goerres, A. (2010). Die soziale Norm der Wahlbeteiligung: Eine international vergleichende Analyse für Europa. *Politische Vierteljahresschrift*, *51*, 275–296. doi: 10.1007/s11615-010-0018-8

Gracia, E. (2004). Unreported cases of domestic violence against women: Towards an epidemiology of social silence, tolerance, and inhibition. *Journal of Epidemiology and Community Health*, *58*, 536–537. doi: 10.1136/jech.2003.019604

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, *64*, 520–539. doi: 10.1080/01621459.1969.10500991

Hafner, S. (2019). *Das Problem sozialer Erwünschtheit bei Befragungen zur Wahlbeteiligung - Eine Anwendung der Randomized-Response-Technik im Vergleich zu direkter Befragung* (Bachelor's Thesis). University of Tübingen.

Haldane, J. B. S. (1945). A labour-saving method of sampling. *Nature*, *155*, 49–50. doi: 10.1038/155049b0

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

Hejri, M. S., Zendehdel, K., Asghari, F., Fotouhi, A., & Rashidian, A. (2013). Academic

disintegrity among medical students: A randomised response technique study. *Medical Education*, *47*, 144–153. doi: 10.1111/medu.12085

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A stochastic lie detector versus the crosswise model. *Behavior Research Methods*, *48*, 1032–1046. doi: 10.3758/s13428-015-0628-6

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

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

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods*, *10*, 171–187. doi: 10.18148/srm/2016.v10i3.6703?c

Iberl, B. (2019). *Anwendung der Curtailed Sampling-Methode bei der Randomized Response Technique zur Ermittlung der Prävalenz von pharmakologischem kognitivem Enhancement bei Studierenden* (Master's Thesis). University of Tübingen.

Jarnecke, A. M., & Flanagan, J. C. (2020). Staying safe curing COVID-19: How a pandemic can escalate risk for intimate partner violence and what can be done to provide individuals with resources and support. *Psychological Trauma: Theory, Research, Practice, and Policy*, *12*, 202–204. doi: 10.1037/tra0000688

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

Krumpal, I., & Voss, T. (2020). Sensitive questions and trust: Explaining respondents' behavior in randomized response surveys. *SAGE Open*, *10*, 1–17. doi: 10.1177/2158244020936223

Landsheer, J. A., Van Der Heijden, P., & Van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality & Quantity*, *33*, 1–12. doi: https://doi.org/10.1023/a:1004361819974

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

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. *Computers in Human Behavior*, *23*, 591–608. doi: 10.1016/j.chb.2004.11.001

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? *PLoS ONE*, *15*, 1–19. doi: 10.1371/journal.pone.0235403

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

Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2010). Reducing socially desirable responses in epidemiologic surveys: An extension of the randomized-response technique. *Epidemiology*, *21*, 379–382. doi: 10.1097/EDE.0b013e3181d61dbc

Näher, A.-F., & Krumpal, I. (2012). Asking sensitive questions : the impact of forgiving wording and question context on social desirability bias. *Quality & Quantity*, *46*, 1601–1616. doi: 10.1007/s11135-011-9469-2

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, *29*, 492–510. doi: 10.1017/S030500410001152X

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

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

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

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

Pitsch, W., Emrich, E., & Klein, M. (2007). Doping in elite sports in Germany: results of a www survey. *European Journal for Sport and Society*, *4*, 89–102. doi: 10.1080/16138171.2007.11687797

Razafimanahaka, J. H., Jenkins, R. K., Andriafidison, D., Randrianandrianina, F., Rakotomboavonjy, V., Keane, A., & Jones, J. P. (2012). Novel approach for quantifying illegal bushmeat consumption reveals high consumption of protected species in Madagascar. *Oryx*, *46*, 584–592. doi: 10.1017/S0030605312000579

Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater Detection Using the Unrelated Question Model. *Sociological Methods and Research*, Advance online publication. doi: 10.1177/0049124120914919

Reiber, F., Schnuerch, M., & Ulrich, R. (2020). Improving the Efficiency of Surveys With Randomized Response Models: A Sequential Approach Based on Curtailed Sampling. *Psychological Methods*, Advance online publication. doi: 10.1037/met0000353

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial Modeling and the Measurement of Cognitive Processes. *Psychological Review*, *95*, 318–339. doi: 10.1037/0033-295X .95.3.318

Schilling, R., Hoebel, J., Müters, S., & Lange, C. (2012). Pharmakologisches Neuroenhancement [Pharmacological neuro-enhancement]. *GBE kompakt. Zahlen und Trends aus der Gesundheitsberichterstattung des Bundes [Facts and Trends from Federal Health Reporting]*, *3*, 1–7.

Schnuerch, M., Erdfelder, E., & Heck, D. W. (2020). Sequential hypothesis tests for multinomial processing tree models. *Journal of Mathematical Psychology*, *95*, 102326. doi: 10.1016/j.jmp.2020.102326

Schröter, H., Studzinski, B., Dietz, P., Ulrich, R., Striegel, H., & Simon, P. (2016). A Comparison of the Cheater Detection and the Unrelated Question Models: A Randomized Response Survey on Physical and Cognitive Doping in Recreational Triathletes. *PLoS ONE*, *11*, e0155765. doi: 10.1371/journal.pone.0155765

Soeken, K. L., & Damrosch, S. P. (1986). Randomized response technique: Applications to research on rape. *Psychology of Women Quarterly*, *10*, 119–126. doi: 10.1111/j.1471-6402.1986.tb00740.x

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

Sugarman, D. B., & Hotaling, G. T. (1997). Intimate violence and social desirability: A meta-analytic review. *Journal of Interpersonal Violence*, *12*, 275 – 290. doi: 10.1177/088626097012002008

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

Ulrich, R., Pope, H. G., Cléret, L., Petróczi, A., Nepusz, T., Schaffer, J., ... Simon, P. (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine*, *48*, 211–219. doi: 10.1007/s40279-017-0765-4

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods*, *17*, 623–641. doi: 10.1037/a0029314

Usher, K., Bhullar, N., Durkin, J., Gyamfi, N., & Jackson, D. (2020). Family violence and COVID-19: Increased vulnerability and reduced options for support. *International Journal of Mental Health Nursing*, *29*, 549–552. doi: 10.1111/inm.12735

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical*

Statistics, *16*, 117–186. doi: 10.1214/aoms/1177731118

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, *19*, 326–339. doi: 10.1214/aoms/1177730197

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69. doi: 10.1080/01621459.1965.10480775

Wetherill, G. B. (1975). *Sequential methods in statistics* (2nd ed.; M. S. Bartlett & D. R. Cox, Eds.). London: Chapman and Hall Ltd.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, *73*, 573–581. doi: 10.1093/biomet/73.3.573

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

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research*, *42*, 321–353. doi: 10.1177/0049124113500474

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

A Acknowledgements

This dissertation would not have been possible without the support of the people who accompanied me in the last three years.

First, I would like to thank Rolf Ulrich for being the best advisor I could have hoped for. Thank you for your optimism, your advice and explanations, and for always taking the time to read and discuss my work. Thank you for being a role-model in your approach to scientific work and priorities and for supporting me with my best personal interest in mind.

I also thank Edgar Erdfelder for being open to discuss my work, for sharing your ideas and experience, and for your feedback and advice.

I would like to thank Martin Schnürch, for sharing your perspectives and expertise with me and for motivating and enriching conversations.

I am grateful to Christian Treffenstädt for introducing me to scientific work and statistics and especially for teaching me the worth and beauty of programming.

I am greatly indebted to the SMiP research training group for offering many opportunities for instructive and enriching experiences. I would especially like to thank Anke Söllner and Annette Förster, for your support and for keeping SMiP running. I also thank Benjamin Hilbig for your feedback and advice as my third supervisor. I am very grateful to my fellow SMiPsters for being a mutually benevolent and supportive group, and, especially Anne, Eileen, Lea, Luisa, Susanne, and Thomas, for great pub conversations, hotel breakfasts and conference adventures.

My time as a PhD candidate in Tübingen would not have been half as stimulating and enjoyable without my dear colleagues. Thank you Cosima, Daniel, Donna, Hera, Karin, Markus, Moritz, Parker, Robert, Ruben, and Victor for fun and inspiring conversations, for hiking tours, and evenings in the Besen and Bahnhofskneipe (while that was still possible). I am especially grateful to Linda for being the best office-mate and PhD-companion and my irreplaceable friend. Thank you for always being there for me, for helping me sort my thoughts and feelings, and for filling our sunshine-office with laughter and beautiful thoughts. I cannot imagine the last three years without you — in- and outside of the office.

Outside of the office, I also found valuable support in my flatmates, Célini, Cilli, Franziski, Jonas, and Tobias. Thank you for being a refreshing and reassuring antipole to work and for being my family in Tübingen. Many friends from earlier stages of my life have

remained a strong source of emotional support throughout the last three years. Thank you, Andreas, Fred, Hanna, Jan, Lena J., Lena R., Robin, and Vanessa, for delightful weekends, holidays and, more recently, zoom-dates, and for being there for me in dark times. I am especially thankful to Linda K. for your fierce friendship and understanding, for your radical honesty and for being an inspiration.

Sometimes the transitions between friends and family get blurred. I would like to thank Ines for being there for me when I most needed you, for your advice and warmth, and Elena for being my close friend throughout all stages of my life and for always believing in me. Thank you, Astrid and Jürgen, for letting me into your family without reservations and for always making me feel welcome, even and especially during the hardest times. Thank you, Constantin, for being my big brother in every regard, for challenging me to question my beliefs, and for dreaming my future. I am forever grateful to my parents, Evelyn and Tobias, for offering me the best starting conditions for this life. Thank you for your unconditional love and support, for teaching me to trust in people, the world, and myself, and for being my safe haven.

Finally, I would like to thank Henrik for baring with me through all the ups and downs of this PhD and life, for helping me to put everything into perspective, and for always seeing in me, what I sometimes could not. The strength and stability of your support help me dare and your humor to do so with joy.

B Copies of Articles

B.1 Article I

Official citation:
Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods and Research*. Advance online publication. doi: 10.1177/0049124120914919

Article

Sociological Methods & Research
1-23

DOI: 10.1177/0049124120914919
journals.sagepub.com/home/smr


Cheater Detection Using the Unrelated Question Model

Fabiola Reiber[1], Harrison Pope[2] and Rolf Ulrich[1]

Abstract

Randomized response techniques (RRTs) are useful survey tools for estimating the prevalence of sensitive issues, such as the prevalence of doping in elite sports. One type of RRT, the unrelated question model (UQM), has become widely used because of its psychological acceptability for study participants and its favorable statistical properties. One drawback of this model, however, is that it does not allow for detecting cheaters—individuals who disobey the survey instructions and instead give self-protecting responses. In this article, we present refined versions of the UQM designed to detect the prevalence of cheating responses. We provide explicit formulas to calculate the parameters of these refined UQM versions and show how the empirical adequacy of these versions can be tested. The Appendices contain R-code for all necessary calculations.

Keywords

sensitive questions, randomized response technique, unrelated question model, cheater detection, instruction nonadherence

[1] University of Tübingen, Tübingen, Germany
[2] Harvard Medical School, Boston, MA, USA

Corresponding Author:
Fabiola Reiber, Department of Psychology, University of Tübingen, Schleichstr. 4, 72076 Tübingen, Germany.
Email: fabiola.reiber@uni-tuebingen.de

Throughout the social sciences, many findings are based on surveys of various groups of individuals. Most such surveys rely on the assumption that respondents will provide honest answers to survey questions. However, this assumption falters when asking respondents sensitive questions (see Tourangeau and Yan 2007)—questions that are perceived as intrusive, stigmatizing, socially undesirable, or even legally incriminating (Tourangeau, Rips, and Rasinski 2000). Faced with sensitive questions, respondents may refuse to participate in the survey or may simply answer dishonestly (Tourangeau et al. 2000), especially if they are carriers of the sensitive attribute being assessed. Thus, direct questioning has frequently been found to underestimate the true prevalence of sensitive attributes, such as having received an abortion (Fu et al. 1998), having been convicted of driving while intoxicated, having engaged in doping in athletics, and many other issues.

To address this problem, several indirect questioning techniques have been developed throughout the last half-century (see Chaudhuri and Christofides 2013). One of these methods, the Randomized Response Technique (RRT), developed by Warner (1965), introduced the idea of creating anonymity by employing random encryption of the respondents' answers. In Warner's model, the respondent receives one of two questions about a sensitive issue. For example, the survey instrument might be designed so that respondents will receive the question S: *Have you ever used illicit drugs?* with probability p (where $p \neq .5$), or they will receive the negative of this question $\neg$S: *Have you never used illicit drugs?* with the complementary probability $1 - p$. A random element (e.g., the throw of a die) determines which of the two questions the respondent receives. The survey is designed so that only the respondent knows the outcome of the randomization (e.g., the respondent is asked to throw the die out of the sight of the investigator). Since only the respondent knows which question he or she has answered the investigator cannot infer the respondent's status when the respondent answers "yes" or "no" to the survey instrument. However, even though the investigators cannot infer the status of any individual respondent, they can nevertheless estimate the prevalence of the sensitive attribute in a large survey population because the probability p underlying the randomization is known, and hence the estimated prevalence of the sensitive attribute can be derived from the proportion of "yes" answers.

Several revisions and modifications of Warner's (1965) model have been proposed over the years (e.g., Kuk 1990; Mangat 1994). One of these is the well-established unrelated question model (UQM; Greenberg et al. 1969; see Figure 1). In the UQM, as in the original Warner model, a randomization procedure determines whether the respondent is instructed to answer the

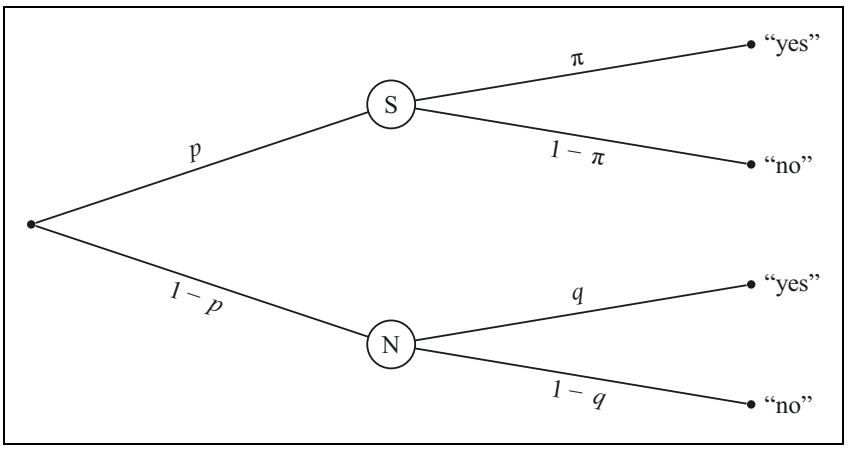


Figure 1. Probability tree of the unrelated question model. The sensitive question S and the neutral question N are randomly received by respondents with probability p and $1 - p$, respectively. The probabilities of responding "yes" and "no" to the neutral question N are q and $1 - q$, and the probabilities of responding "yes" and "no" to the sensitive question S are π and $1 - \pi$.

sensitive question S. The alternative question, however, is not the reversed sensitive question $\neg$S, but instead an unrelated innocuous question, the neutral question N (e.g., "Think of someone close to you whose birthdate you know, and answer "yes" if that individual was born on an odd-numbered day"). Thus, the UQM is potentially more psychologically acceptable to survey respondents than the original Warner method because question N is obviously not related to the sensitive attribute and is therefore clearly not incriminating.

With the UQM, as with Warner's original method, the investigator cannot determine any individual respondent's status on the sensitive attribute. However, given a large sample of respondents, the investigator can still estimate the prevalence π of the sensitive attribute, provided that the randomization probability p and the prevalence of the neutral attribute q are known. Specifically, the prevalence π can be estimated from the observed proportion $\hat{\lambda}$ of "yes" responses by the formula:

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

In several studies, the UQM has elicited prevalence estimates substantially exceeding estimates derived from direct questioning (see Lensvelt-Mulders et al. 2005), such as the prevalence of induced abortion (Abernathy, Greenberg, and Horvitz 1970) and doping in elite athletics (e.g., Ulrich et al. 2018).

However, by introducing an unrelated question N, the UQM opens the possibility that some respondents ("cheaters") will be tempted to answer a self-protective "no" to either of the two alternative questions on the survey regardless of the true answer to the question. Even though a "yes" response does not necessarily imply having the sensitive attribute, a "no" response greatly reduces the possibility of that conclusion. Specifically, under the standard version of the UQM, the conditional probability $P(A|\text{"yes"})$ of being a carrier given a "yes" response is generally larger than the conditional probability $P(A|\text{"no"})$ of being a carrier given a "no" response, when π is less than one. For example, for $p = 0.75$, $q = 0.5$, and $\pi = 0.2$ one computes $P(A|\text{"yes"}) = 0.636$ and $P(A|\text{"no"}) = 0.034$ using Bayes's theorem. Correspondingly, the odds that one is a carrier of the attribute would be 49 times greater given a "yes" response than given a "no" response. Interestingly, this conclusion does not depend on π. As a consequence, this difference in conditional probabilities may encourage cheating behavior in the form of answering "no" under all circumstances.

Another modification of the RRT, the cheater detection model (CDM; Clark and Desharnais 1998), addresses this drawback by dividing respondents into three mutually exclusive categories: (a) honest respondents who are carriers of the sensitive attribute, who will respond "yes" if they receive the sensitive question, (b) honest respondents who are noncarriers of the sensitive attribute, who will respond "no" if they receive the sensitive question, and (c) cheaters who choose the safe option by always responding "no" to any question regardless of whether they are carriers or noncarriers. For illustration, let A be a carrier and $\neg A$ be a noncarrier. Furthermore, let H be an honest respondent and $\neg H$ be a cheater. Then, the probabilities of the three subgroups can be expressed as compound probabilities. These probabilities are for subgroup (a)

$$P(A \cap H) = P(A|H) \cdot P(H) = \varepsilon \cdot (1 - \gamma), \tag{2}$$

for subgroup (b)

$$P(\neg A \cap H) = P(\neg A|H) \cdot P(H) = (1 - \varepsilon) \cdot (1 - \gamma), \tag{3}$$

and for subgroup (c)

$$P(\neg H) = P(C) = \gamma. \tag{4}$$

Note that these three probabilities add to one.

The CDM is based on another RRT variant, the forced response model (Boruch 1971). This model modifies Warner's model by replacing the

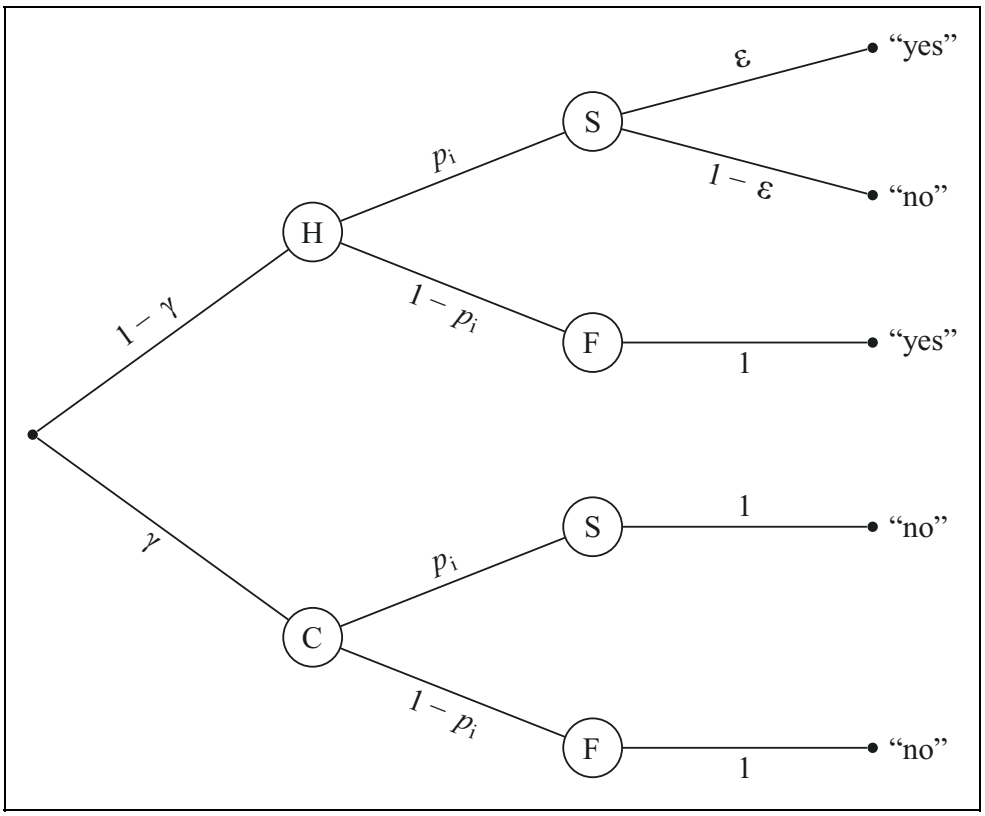


Figure 2. Probability tree of the cheater detection model. Respondents are either cheaters C with probability γ or honest respondents H with probability $1 - \gamma$. All respondents randomly receive either the sensitive question S or the instruction F to respond "yes" with probability p_i and $1 - p_i$, respectively. Cheaters C always answer "no" regardless of their carrier status and regardless of whether they receive question S or instruction F. Honest respondents H respond honestly under all conditions. Specifically, if instructed to say "yes," honest participants always answer "yes." If instructed to answer the sensitive question S, honest participants answer "yes" with probability ε and "no" with probability $1 - \varepsilon$. Thus, participants can be divided into three groups: (a) carriers of the sensitive attribute who will honestly respond "yes" with probability $(1 - \gamma) \cdot \varepsilon = \pi$ when receiving S; (b) noncarriers of this attribute who will honestly respond "no" with probability $(1 - \gamma) \cdot (1 - \varepsilon)$ when receiving S; and (c) cheaters who will respond "no" with probability γ regardless of receiving S or the instruction F to respond "yes."

inverted question ¬ S by the instruction to simply say "yes." In other words, the forced instruction to say "yes" simply replaces the neutral question N in the UQM. Hence, if no cheating is assumed and one is therefore not attempting to assess for cheating, the forced response model is mathematically equivalent to a special case of the UQM, namely when the prevalence q of the neutral attribute equals 1. This situation is depicted in the upper part of Figure 2 starting at node H, representing honest respondents only.

However, note that the temptation to cheat may be especially pronounced in the forced response model because the respondent can completely eliminate any suggestion of being a carrier of the sensitive attribute by simply answering "no." Expressed more formally, in the forced response model, the conditional probability $P(A|\text{"yes"})$ must be always larger than the conditional probability $P(A|\text{"no"})$ because $P(A|\text{"no"}) = 0$ (except in the implausible case where $P(A|\text{"yes"})$ is also 0). For example, for $p = 0.75$ and $\pi = 0.2$, one computes $P(A|\text{"yes"}) = 0.5$ and $P(A|\text{"no"}) = 0$. Correspondingly, the odds that the respondent is a carrier of the attribute would be infinitely greater given a "yes" response than given a "no" response. In other words, answering with "no" is a completely safe option.

Therefore, the CDM includes a parameter to assess the extent of cheating. This is depicted in the lower part of Figure 2 starting at node C and representing cheaters. In this diagram of the CDM, the proportion of cheaters is γ, whereas the proportion of honest respondents is $1 - \gamma$. The proportion of respondents carrying the sensitive attribute cannot be estimated because only the proportion π of honest carriers in the overall respondent population, but not the proportion of carriers who are cheaters in the overall population, can be identified by the model. Importantly, π in the CDM is, therefore, not equivalent to π in the UQM because in the former it is defined as the proportion of honest carriers and in the latter as the total proportion of carriers. Nevertheless, in the CDM, the total proportion of carriers in the population must lie within the range that is defined by the lower bound $\pi = (1 - \gamma) \cdot \varepsilon$ and the upper bound $\pi + \gamma$. The proportions π and γ thus represent two of the above introduced categories, namely (a) honest carriers and (c) cheaters, respectively. Therefore, the proportion of respondents in the remaining category (b)—the honest noncarriers—is simply given by $1 - (\pi + \gamma)$. In order to estimate the parameters π and γ for computing the two bounds, two probabilities λ_1 and λ_2 of responding "yes" are required. They can be estimated by the observed proportion of "yes" responses in two independent samples with $p_1 \neq p_2$. The resulting equation system can then be solved for π and γ.

Several empirical implementations of the CDM (e.g., Elbe and Pitsch 2018; Moshagen et al. 2010; Ostapczuk 2011; Ostapczuk et al. 2009; Pitsch, Emrich, and Klein 2007; Schröter et al. 2016) have provided evidence of cheating behavior—showing the importance of including a cheating parameter in RRTs. However, studies utilizing the forced response model (e.g., Höglinger, Jann, and Diekmann 2016; Kirchner 2015; Wolter and Preisendörfer 2013) have raised doubts about the validity of this particular method. Specifically, it has been shown (Coutts and Jann 2011; Höglinger et al. 2016)

to elicit lower estimates than other indirect questioning techniques, and respondents have reported greater difficulties in understanding this technique. Respondents also seem to express less trust that the technique guarantees anonymity. For example, Lensvelt-Mulders and Boeije (2007) reported that respondents perceived being forced to give a "yes" response as being "forced to be dishonest" (p. 600), which seemingly triggered reluctance.

Ostapczuk et al. (2009) proposed a method to reduce this problem by adding a forced "no" response to the forced "yes" response. In this symmetric design, none of the response options is conclusive of the respondents' status. Specifically, it is not only possible to be forced to respond "yes" even though one is a noncarrier but also to be forced to respond "no" even though one is in fact a carrier. This should increase compliance with the instructions, and indeed, the authors found cheating to be reduced in an empirical comparison to the original design. Still, it is plausible that a forced response can feel like an implicit response to the sensitive question, something that even this approach does not address.

In summary, although it appears important to account for possible cheating when using RRTs, a technique based on the forced response model may not be ideal. By contrast, the UQM is conceptually and mathematically similar without potentially triggering reluctance by forcing responses. Here, responses to the neutral question are clearly not responses to the sensitive question because the neutral question has content of its own. Thus, in the next section, we propose a model combining the greater psychological acceptability of the UQM's design with the CDM's concept of cheating.

Unrelated Question Model—Cheating Extension (UQMC)

Below, we introduce the UQMC, a model combining the basic idea of the CDM (Clark and Desharnais 1998) with the standard version of the UQM (Greenberg et al. 1969). The setup of the UQMC resembles that of the UQM, in that respondents receive the sensitive question S with probability p and the neutral question N with probability $1 - p$. As in the CDM, participants are categorized as being either honest respondents or cheaters. Figure 3 depicts the resulting probabilities. The same parameters generated in the CDM can be estimated using this model. Specifically, γ corresponds to the probability of being a cheater, and $\pi = (1 - \gamma) \cdot \varepsilon$ depicts the probability of being an honest carrier of the sensitive attribute. As in the CDM, the prevalence of the sensitive attribute cannot be inferred because the proportion of carriers can only be estimated among honest respondents and not among cheaters.

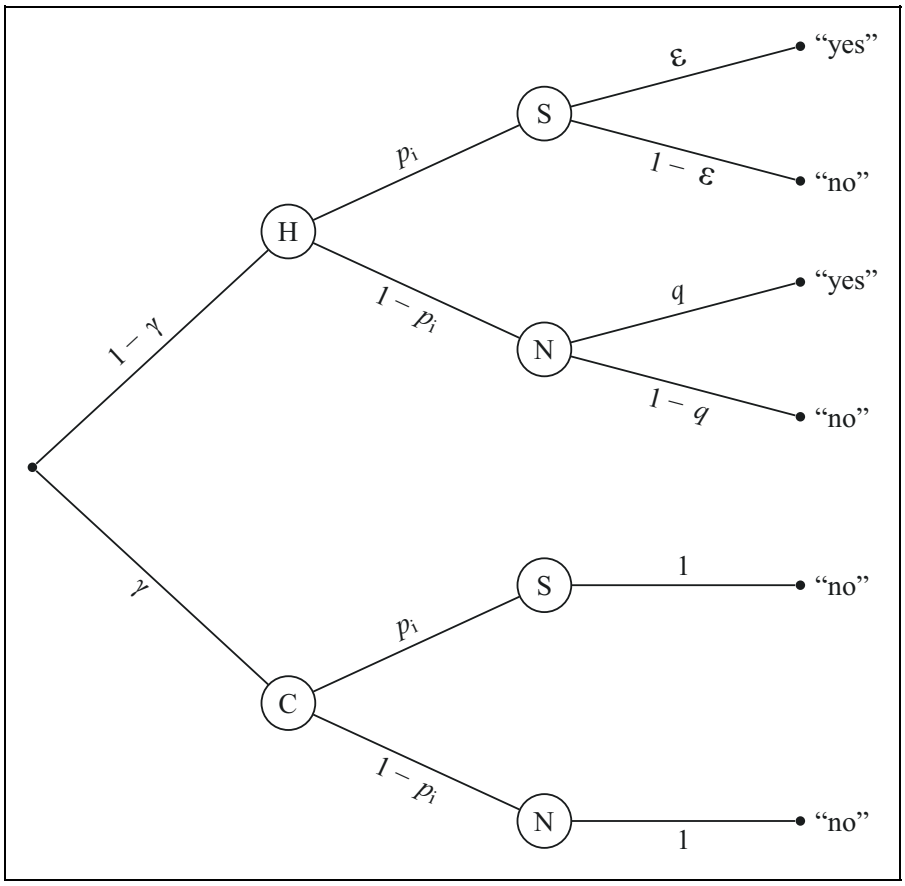


Figure 3. Probability tree of the unrelated question model—cheating extension. The prevalence of cheaters C is γ and the prevalence of honest participants H is $1-\gamma$. In both cases, the sensitive question S and the neutral question N are received by participants with probability p_i and $1-p_i$, respectively. Cheaters always say "no" regardless of the question received. Honest participants respond "yes" with probability q and "no" with probability $1-q$ if instructed to answer the neutral question N. They answer "yes" with probability ε and "no" with probability $1-\varepsilon$, if instructed to answer the sensitive question S. Thus, there are three groups of participants: (a) honest participants who are carriers of the sensitive attribute, who will respond "yes" with probability $(1-\gamma)\cdot\varepsilon=\pi$ if they receive S; (b) honest noncarriers of this attribute who will respond "no" with probability $(1-\gamma)\cdot(1-\varepsilon)$ if they receive S; and (c) cheaters who will respond "no" with probability γ regardless of whether they receive S or N.

However, it is still possible to compute an estimated range for the prevalence, which is defined by the bounds π and $\pi+\gamma$.

As in the CDM, two independent samples of respondents are required to estimate π and γ. Again, different values of p_i must be used with the two samples, $i=1,2$. Thus, the probability of responding "yes" in sample i is given by

$$\lambda_i = p_i \cdot \pi + (1 - p_i) \cdot (1 - \gamma) \cdot q. \tag{5}$$

As λ_1 and λ_2 can be estimated from the corresponding observed proportion $\hat{\lambda}_1$ and $\hat{\lambda}_2$ of "yes" responses in each sample, the resulting equation system can be solved for π and γ,

$$\hat{\pi} = \frac{\hat{\lambda}_2 \cdot (1 - p_1) - \hat{\lambda}_1 \cdot (1 - p_2)}{p_2 - p_1}, \tag{6}$$

and

$$\hat{\gamma} = 1 - \frac{\hat{\lambda}_2 \cdot p_1 - \hat{\lambda}_1 \cdot p_2}{q \cdot (p_1 - p_2)}. \tag{7}$$

The corresponding sampling variances of the two estimates are

$$Var(\hat{\pi}) = \frac{1}{(p_2 - p_1)^2} \left[(1 - p_1)^2 \cdot \frac{\lambda_2(1 - \lambda_2)}{n_2} + (1 - p_2)^2 \cdot \frac{\lambda_1(1 - \lambda_1)}{n_1} \right], \tag{8}$$

and

$$Var(\hat{\gamma}) = \frac{1}{q^2 \cdot (p_1 - p_2)^2} \left[p_2^2 \cdot \frac{\lambda_1(1 - \lambda_1)}{n_1} + p_1^2 \cdot \frac{\lambda_2(1 - \lambda_2)}{n_2} \right]. \tag{9}$$

The covariance of these estimators is

$$Cov(\hat{\pi}, \hat{\gamma}) = \frac{1}{q \cdot (2p_1p_2 - p_1^2 - p_2^2)} \left[(p_1^2 - p_1) \cdot \frac{\lambda_2(1 - \lambda_2)}{n_2} + (p_2^2 - p_2) \cdot \frac{\lambda_1(1 - \lambda_1)}{n_1} \right]. \tag{10}$$

Table 1 provides a numerical example to illustrate the UQMC. This example assumes that the estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ of "yes" responses were obtained from two independent samples. The observed proportions of "yes" responses in this table were simulated with $\pi = 0.2$, and $\gamma = 0.3$. Inserting the values of Table 1 into equations (6–9) yields parameter estimates $\hat{\pi}$ and $\hat{\gamma}$ with their standard errors, which are depicted in Table 2. These estimates can be used to generate the possible range of the prevalence of the sensitive attribute. The lower bound of this range (i.e., the lowest possible estimate of the prevalence) is $\hat{\pi} = 0.190$, with a 95 percent confidence interval of 0.149 to 0.231. The upper bound is $\hat{\pi} + \hat{\gamma} = 0.190 + 0.305 = 0.495$. The sampling variance of this upper bound is given by

Table 1. Numerical Example Illustrating the Unrelated Question Model—Cheating Extension.

Sample	n_i	p_i	q	o_{yi}	o_{ni}	$\hat{\lambda}_i$
1	1,000	.75	.5	229	771	.229
2	1,000	.25	.5	308	692	.308

Note. n_i = size of sample i; p_i = probability of being assigned to the sensitive question in sample i; q = prevalence of the neutral attribute; o_{yi} = observed frequency of "yes" responses in sample i; o_{ni} = observed frequency of "no" responses in sample i; $\hat{\lambda}_i$ = proportion of "yes" responses in sample i.

Table 2. Numerical Example Illustrating the Unrelated Question Model—Cheating Extension (Continued).

Parameter	Prevalence	Estimate	*SE*	CI
π	.200	.190	.021	[.149, .231]
γ	.300	.305	.046	[.215, .395]
$\pi + \gamma$	.500	.495	.079	[.341, .648]

Note. *SE* = standard error of parameter estimate; CI = 95 percent confidence interval of parameter estimate.

$$Var(\hat{\pi} + \hat{\gamma}) = Var(\hat{\pi}) + Var(\hat{\gamma}) + 2 \cdot Cov(\hat{\pi}, \hat{\gamma}), \tag{11}$$

using equations (8–10). Therefore, the 95 percent confidence interval of the upper bound ranges from 0.341 to 0.648. Hence, even though the prevalence of carriers among cheaters remains unknown, one can conclude from this model that the estimated total proportion of carriers is at least 0.190 and at most 0.495 with 95 percent confidence intervals ranging from 0.149 to 0.648.

It is important to note that the size of this range is in large part due to the true cheating proportion, which is 0.3 in this example, and not merely due to random sampling error. A model that does not take cheating into account, such as the original UQM, would therefore yield an estimate with a smaller confidence interval. On first sight, this may look preferable. However, this estimate would be biased, as it disregards the true prevalence of cheating. As such, there is uncertainty in both cases, but only the UQMC makes the degree of this uncertainty explicit by taking cheating into account. If on the other hand, there is in fact no cheating, the UQMC can capture this as well (with $\hat{\gamma}$ approximating 0), and the confidence interval of the prevalence estimate range will decrease correspondingly. By way of illustration, if one changes

the true cheating prevalence in the above example to $\gamma = 0.1$, the estimates resulting from simulation are $\hat{\pi} = 0.239$ with 95 percent confidence interval ranging from 0.193 to 0.284 and $\hat{\pi} + \hat{\gamma} = 0.239 + 0.082 = 0.321$ with 95 percent confidence interval ranging from 0.193 to 0.449. As can be seen from this example positing a lower rate of cheating, the 95 percent confidence interval for the estimated range of the carrier proportion is much smaller, namely 0.193 to 0.449.

In addition to estimating the above parameters, the UQMC can test whether a substantial amount of cheating is present. Indeed, Clark and Desharnais (1998) introduced a likelihood ratio test for this purpose in their initial presentation of the CDM. This test utilizes the ratio of the maximum likelihood of a model setting cheating to $\gamma = 0$ and the maximum likelihood of a model allowing for cheating. It can be applied to the UQMC in a similar manner, where it is formalized as

$$\chi^2(1) = 2 \cdot [\log L(\hat{\pi}, \hat{\gamma}) - \log L(\hat{\pi}^*, \gamma = 0)]. \quad (12)$$

In the above example, this likelihood ratio test supports the hypothesis that cheating is present, with $\chi^2(1) = 41.119, p < .001$. Appendix A (which can be found at http://smr.sagepub.com/supplemental/) contains R-code that can be used for applying the calculations to one's own data.

As is true for all indirect questioning techniques, the sampling variance of the estimates is quite high. Due to the additional estimation of the cheating parameter, this variance becomes even higher than in one-parameter RRMs, such as the original UQM. An optimized choice of p_i and q, and an optimized division of the sample into the two subsamples can minimize this drawback. Appendix B (which can be found at http://smr.sagepub.com/supplemental/) illustrates the influence each of these parameters has on the sum of standard errors and power of the model estimates. In short, more extreme values of p_i and larger values of q make the sum of standard errors smaller and the relative size of the two subsamples within the overall sample has only a small impact, as long as the difference is not too extreme. Thus, a division of the sample into two equal subsamples is desirable. However, minimizing the standard error cannot be the only consideration when choosing the values for p_i and q because in case of values for p_i and q close to 0 or 1, the responses become more indicative of the respondents' status and thus anonymity protection decreases. Therefore, the applied values must be chosen to represent a compromise between efficiency and anonymity protection. Recommended values would therefore be 0.75 and 0.70 for p_1 and q, respectively.

Different parameter combinations might be advantageous if the focus of the study is mainly on prevalence estimation or mainly on cheater estimation. In the former case p_i should be more extreme, q should be smaller, and the larger part of the sample should be allocated to the subsample with higher p_i. In the latter case, p_i should be closer to 0.5, q should be higher, and the larger part of the sample should be allocated to the subsample with lower p_i.

The above recommendations are based on the influence that the design parameters have on the standard error and statistical power, together with an intuitive evaluation of the influence that these parameters have on perceived privacy protection. In specific applications, the parameters should be informed by the specific sensitive question at hand and the implementation of the questioning design. In doing so, one can refer to theoretical as well as empirical work on the optimal choice of design parameters in RRMs with respect to efficiency and perceived privacy protection (e.g., Greenberg et al. 1977; Lanke 1975; Leysieffer and Warner 1976; Ljungqvist 1993; Soeken and Macready 1982). An overview on this topic is given by Fox (2016).

Partial Cheating

As explained above, the UQMC utilizes the cheating concept as initially defined in the CDM, where "cheaters" are assumed to always choose the safe option of a "no" response, regardless of the question presented. However, this may be an unduly restrictive assumption, as there might be respondents who would cheat when confronted with the sensitive question but would answer the neutral question truthfully, since they do not feel threatened by this latter question. Allowing for cheating in this broader and probably more realistic sense implies that the original categories (completely honest respondents and complete cheaters) should be extended by the category "partial cheaters" (i.e., cheating only if presented with the sensitive question). In the following, we refer to the original group of cheaters, who always respond "no," as "complete cheaters."

Figure 4 depicts how partial cheating affects the probabilities for "yes" and "no" responses. Honest respondents still answer honestly to whichever question they are assigned. Complete cheaters, as before, respond "no" to whichever question they are assigned. In this figure, we add partial cheaters, who answer honestly if assigned to the neutral question, but always respond "no" to the sensitive question, regardless of whether they are carriers of the sensitive attribute. Thus, there is a new branch of the probability tree leading to a "yes" response, $\gamma_p \cdot (1 - p_i) \cdot q$. The resulting total probability for answering "yes" if there is partial cheating can be reduced to

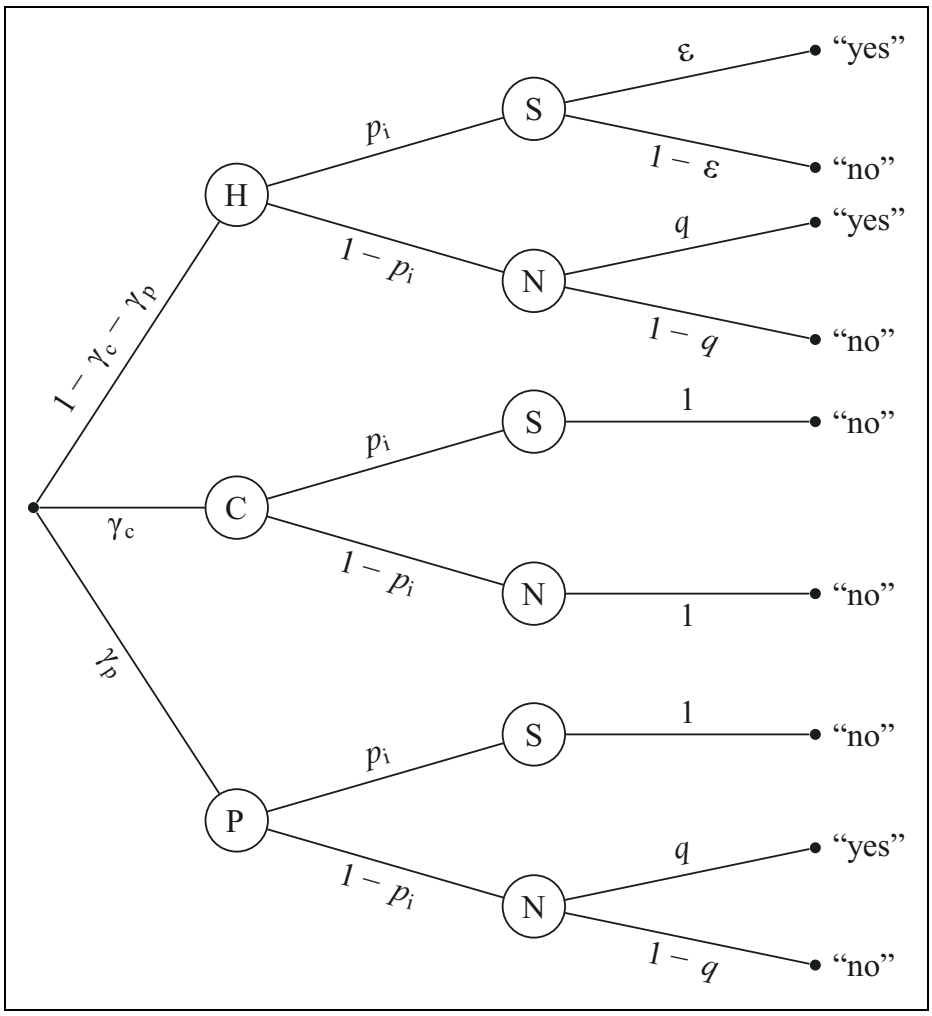


Figure 4. Probability tree of the unrelated question model—cheating extension including partial cheating. Participants are (a) honest H with probability $1-\gamma_c-\gamma_p$, (b) partial cheaters P with probability γ_p, or (c) complete cheaters C with probability γ_c. All types of participants receive the sensitive question S and the neutral question N with probability p_i and $1-p_i$, respectively. (a) Honest participants respond "yes" with probability q and "no" with probability $1-q$, if instructed to answer the neutral question N. They answer "yes" with probability ε and "no" with probability $1-\varepsilon$, if instructed to answer the sensitive question S. (b) Partial cheaters always say "no" if they are instructed to answer the sensitive question S, regardless whether or not they are carriers, but if instructed to answer the neutral question N, they answer honestly by saying "yes" with probability q and "no" with probability $1-q$. (c) Complete cheaters always answer "no" regardless of the question that they receive and regardless of whether or not they carry the attribute.

$$\lambda_i = (1-\gamma_c-\gamma_p)\cdot\varepsilon\cdot p_i + (1-p_i)\cdot(1-\gamma_c)\cdot q. \tag{13}$$

It should be stressed that not all three parameters γ_c, γ_p, and ε can be estimated from empirical data. In other words, the same value of π can be

achieved by an infinite number of combinations of γ_p and ε, which would give rise to the same probability λ_i. Therefore, this extension can be only partially solved for parameters $\pi = (1 - \gamma_c - \gamma_p) \cdot \varepsilon$ and γ_c. As such, π can be inserted into equation (13) resulting in

$$\lambda_i = \pi \cdot p_i + (1 - p_i) \cdot (1 - \gamma_c) \cdot q. \tag{14}$$

It is clear that equation (14) is equivalent to equation (5), except that γ is replaced by γ_c. Thus, the lower bound for the estimated prevalence of the sensitive attribute is still defined by π when allowing for partial cheaters. However, the upper bound of the estimated prevalence, which was formerly given by $\pi + \gamma$, may no longer be given by $\pi + \gamma_c$ after allowing for partial cheaters because the remaining category now comprises not only the proportion of honest noncarriers but additionally γ_p. Since partial cheaters can be carriers of the attribute, γ_p should be added to the possible prevalence range. This results in an increased upper bound of $\pi + \gamma_c + \gamma_p$, which cannot be determined because γ_p is not identifiable.

For the above numerical example, this would mean that the estimate for the lower bound of the prevalence range would remain at $\hat{\pi} = 0.188$. The estimate for the upper bound, however, would potentially exceed $\hat{\pi} + \hat{\gamma}_c = 0.188 + 0.304 = 0.492$ because there could be an additional unknown proportion of partial cheaters. In other words, if one computes the prevalence of the sensitive attribute using the UQMC, which formally assumes only the possibility of complete cheating, the estimate of the lower bound of carrier prevalence is not affected by the presence of partial cheaters, but the upper bound of this range may be underestimated if partial cheaters are present. This consideration should be kept in mind when interpreting the results of a study using the UQMC. In other words, if one wants to address partial cheating within the UQMC framework, the same estimates can be calculated but need to be interpreted differently concerning the upper bound of the prevalence estimate.

It is worth mentioning that the same line of reasoning would apply to the CDM. That is, the possibility of partial cheating would involve a reinterpretation of the parameters estimated by the CDM. Specifically, as before, in the presence of partial cheating, the lower bound of the prevalence would remain at π. However, the upper bound could exceed $\pi + \gamma$ if partial cheaters are present.

A Survey Design for Testing the UQMC

A limitation of RRTs in general is that their empirical adequacy cannot be tested because the number of unknown parameters usually equals the number

Table 3. Numerical Example Illustrating the Unrelated Question Model—Cheating Extension with Four Samples.

Sample	n_{ij}	p_i	q_j	o_{yij}	o_{nij}	$\hat{\lambda}_{ij}$
11	500	.75	.7	129	371	.258
12	500	.75	.3	96	404	.192
21	500	.25	.7	204	296	.408
22	500	.25	.3	98	402	.196

Note. n_{ij} = size of sample *ij*; p_i = probability to be assigned to the sensitive question in samples *i*; q_j = prevalence of the neutral attribute in samples *j*; o_{yij} = observed frequency of "yes" responses in sample *ij*; o_{nij} = observed frequency of "no" responses in sample *ij*; $\hat{\lambda}_{ij}$ = proportion of "yes" responses in sample *ij*.

of independent samples, and therefore, there are no degrees of freedom left for testing empirical adequacy. Thus, empirical adequacy must simply be assumed. Fortunately, this drawback can be resolved in the UQMC by varying the prevalence of the neutral attribute q. In the basic UQMC, p_1 and p_2 are applied to two independent samples in order to generate two independent equations for λ_1 and λ_2, allowing for two parameters to be identified. However, if q_j is varied orthogonally to p_i, four independent samples can be drawn, each with a unique combination of these design parameters, (p_1, q_1), (p_1, q_2), (p_2, q_1), and (p_2, q_2). The resulting model with four independent equations for λ_{ij} (λ_{11}, λ_{12}, λ_{21}, and λ_{22}) provides two degrees of freedom, allowing for an empirical test of adequacy.

Table 3 illustrates what the setup of the UQMC with four samples could look like, including exemplary estimates $\hat{\lambda}_{ij}$. Like in the first example, the observed proportions of "yes" responses in this table were simulated with $\pi = 0.2$ and $\gamma = 0.3$. In this case, there is no explicit solution for the estimation of the model parameters. Parameter estimates $\hat{\pi}$ and $\hat{\gamma}$ can be obtained by numerical maximum likelihood estimation. Furthermore, the standard errors of the estimated parameters can be numerically evaluated using the observed Fisher information. For the example in Table 3, these estimates are depicted in Table 4. The likelihood ratio test can also be conducted in the four-sample extension. In the numerical example here, the results are in favor of the hypothesis that cheating is present, with $\chi^2(1) = 55.029$, $p < .001$. The exemplary results shown so far are equivalent to those obtainable by the UQMC with two samples. However, the four-sample extension additionally enables testing of the model's adequacy using Pearson's χ^2 goodness-of-fit test. In the UQMC, this is formalized as

Table 4. Numerical Example Illustrating the Unrelated Question Model—Cheating Extension with Four Samples (Continued).

Parameter	Prevalence	Estimate	*SE*	CI
π	.200	.186	.020	[.146, .225]
γ	.300	.317	.042	[.234, .400]
$\pi + \gamma$	.500	.502	.056	[.393, .612]

Note. *SE* = standard error of parameter estimate; CI = 95 percent confidence interval of parameter estimate.

$$\chi^2(2) = \sum_{i=1}^{2}\sum_{j=1}^{2}\left[\frac{(o_{yij} - e_{yij})^2}{e_{yij}} + \frac{(o_{nij} - e_{nij})^2}{e_{nij}}\right], \tag{15}$$

where o_{yij} and o_{nij} are the observed frequencies of "yes" responses and "no" responses, respectively, in each sample with p_i and q_j. Likewise, e_{yij} and e_{nij} are the corresponding expected frequencies. The test supports the fit of the UQMC in the numerical example, $\chi^2(2) = 0.080$, $p = .961$. Appendix C (which can be found at http://smr.sagepub.com/supplemental/) contains R-code for parameter estimation and the goodness-of-fit test that can be applied to one's own data.

Discussion

The present article extends the UQM to allow it to assess cheating while still ensuring respondents' anonymity. This extension incorporates the basic idea of the CDM (Clark and Desharnais 1998) while preserving the more psychologically acceptable design of the UQM. Such an extension seems appropriate because there is ample evidence that many respondents cheat by always answering "no" in randomized response surveys (e.g., Elbe and Pitsch 2018; Moshagen et al. 2010; Ostapczuk 2011; Ostapczuk et al. 2009; Pitsch et al. 2007; Schröter et al. 2016), probably because a "no" response reduces the fear of embarrassment or other negative consequences. In particular, when a respondent is administered the UQM, such cheating would greatly diminish the conditional probability of being deemed a carrier of the sensitive attribute. For example, as noted earlier, Bayesian analysis reveals that for the design parameters $p = 0.75$ and $q = 0.50$, the odds of carrying the sensitive attribute would be 49 times higher in the presence of a "yes" response as opposed to a "no" response, if respondents were to obey the UQM's instructions. Therefore, disobeying these instructions by cheating

with uniform "no" responses is potentially attractive as a self-protecting strategy.

In the present article, we have first introduced an extension of the UQM utilizing the standard assumptions of the CDM—namely the assumption that cheaters will always respond "no" regardless of whether they are directed to the sensitive or to the neutral question. For this extension of the UQM, which we have termed the UQMC, we provide explicit formulae to compute the lower and upper bound of the prevalence estimate range, together with a likelihood ratio test to statistically assess the presence of cheating.

Second, we have discussed in this article the possibility of partial cheating in addition to complete cheating—a perhaps more realistic assumption. Partial cheaters answer honestly if directed to the neutral question but always respond "no" if directed to the sensitive question, even if they are in fact carriers of the sensitive attribute. The parameters of a model including partial cheating are only partially identifiable. Currently, we are not aware of a mathematical or experimental solution for this limitation. However, we have shown that even if partial cheating is disregarded, as in the UQMC, the lower prevalence limit is not affected if partial cheaters are present, although the upper limit may be higher than that estimated by the UQMC if partial cheaters are present. Importantly, such a lower bound provides relevant information like, for example, in a study on the prevalence of doping in elite athletics using the UQM (Ulrich et al. 2018). The UQM estimates of more than 30 percent were clearly much higher than the prevalence estimates from physical doping tests, which indicated a prevalence of about 2 percent at the time (World Anti-Doping Agency 2012). Consequently, even if this only represents a lower bound to the prevalence, the implications are considerable. In addition, the UQMC can account for a very likely type of nonadherence, namely complete cheating. Thus, even if one wants to avoid overconfident conclusions and regards partial cheating, UQMC estimates can have important implications.

Third, we have also shown how the adequacy of the UQMC can be empirically tested. Finally, we have performed power analyses to show that reliable parameter estimates can be obtained even with modest total sample sizes.

The described RRT cheating models assume the presence of "no" cheating for self-protective reasons. Nevertheless, it is at least conceivable that some respondents could cheat with a false "yes" response. For example, a clean athlete might be tempted to cheat with "yes" in order to inflate the prevalence estimate of doping in the hope that this would lead to stricter anti-doping policies (Elbe and Pitsch 2018). In light of this possibility, Feth et al.

(2017) extended the CDM to address not only "no" cheating but also "yes" cheating. These authors regard the idea of the CDM in the context of a more general variant of the forced response method, in which there is a forced "no" response in addition to the forced "yes" response. The authors provide an in-depth discussion of the estimation of "yes" and "no" cheating within this framework and also mention the possibility of transferring this idea to the UQM. This CDM extension was recently applied to estimate the prevalence of doping among elite Danish athletes (Elbe and Pitsch 2018). Although the model revealed a high proportion of "no" cheaters, the proportion of "yes" cheaters was virtually nil. A similar conclusion was reached in a recent experimental individual-level validation study (Höglinger and Jann 2018), which examined whether cheating in a dice game could be accurately assessed by several indirect questioning techniques—and, if not, in which direction respondents misreport on their actual behavior. In case of the UQM, these investigators found a substantial prevalence of false-negative responses (i.e., "no" cheating), but not of false-positive responses (i.e., "yes" cheating). These findings are consistent with several lines of evidence indicating that misreporting usually occurs in the socially desirable direction (see Tourangeau and Yan 2007). In the present article, we have extended the standard UQM only for "no" cheating, but future extensions of the UQM could include the possibility of "yes" cheating (including, at least in theory, the possibilities of both complete and partial "yes" cheating). However, assessing for "yes" cheating would likely be useful only in rare situations where social desirability plays a subordinate role, or where there might be a plausible motivation for "yes" cheating.

In the UQMC, the estimation of two parameters requires independent subsamples. A possible limitation of this approach is that it relies on the assumption that these subsamples do not differ with respect to the true parameter values. In case of the cheating parameter, this assumption could be violated because different probabilities of receiving the sensitive question might induce different levels of trust and hence different levels of cheating. There are alternative approaches to estimate nonadherence parameters that do not rely on independent subsamples (e.g., Böckenholt and van der Heijden 2007; Böckenholt, Barlas, and van der Heijden 2009; Cruyff, Böckenholt, and van der Heijden 2016). However, these approaches usually involve the assessment of multiple RRM questions instead of using independent subsamples. Thus, these alternative approaches are not equally suited to the same research questions as approaches using subsamples. When applying the UQMC, this risk of violating the above-mentioned assumption can be minimized by defining the design parameters such that the motivation to

cheat would not be expected to strongly differ between subsamples. Additionally, and most crucially, the model test proposed in this article allows one to assess the adequacy of these assumptions.

In this article, we have focused on the UQM and CDM. The Crosswise Model (Yu, Tian, and Tang 2008) provides an alternative to these two models. An advantage of this model is that it does not necessitate a randomization device, nor does it require a "yes"/"no" response. Thus, a response cannot be interpreted as a direct response to the sensitive question, which seems to increase perceived anonymity (Hoffmann et al. 2017). Despite these advantages, this model also has drawbacks. First, the sampling variance of this model's prevalence estimate is relatively high and thus samples much larger than those typically used in the original UQM are required (Ulrich et al. 2012). Second, the Crosswise Model has been shown to be susceptible to other types of instruction nonadherence, which may distort the prevalence estimate (e.g., Höglinger and Diekmann 2017; Höglinger and Jann 2018).

In summary, the present article attempts to enrich the RRT toolbox by extending one of the most common RRT models, the UQM, to allow for the estimation of cheaters. This extended model is relatively easy to implement in surveys. Therefore, we recommend that cheating and model adequacy should be routinely taken into account in future RRT surveys that will employ the UQM.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Deutsche Forschungsgemeinschaft (DFG), grant 2277, Research Training Group "Statistical Modeling in Psychology" (SMiP).

ORCID iD

Fabiola Reiber https://orcid.org/0000-0002-5654-4985

Supplemental Material

Supplemental material for this article is available online.

References

Abernathy, James R., Bernard G. Greenberg, and Daniel G. Horvitz. 1970. "Estimates of Induced Abortion in Urban North Carolina." *Demography* 7:19-29.

Böckenholt, Ulf and Peter G. M. van der Heijden. 2007. "Item Randomized-Response Models for Measuring Noncompliance: Risk-return Perceptions, Social Influences, and Self-protective Responses." *Psychometrica* 72:245-62. doi:10.1007/s11336-005-1495-y.

Böckenholt, Ulf, Sema Barlas, and P. G. M. van der Heijden. 2009. "Do Randomized-Response Designs Eliminate Response Biases? An Empirical Study of Noncompliance Behavior." *Journal of Applied Economics* 24:377-92. doi:10.1002/jae.1052.

Boruch, Robert F. 1971. "Assuring Confidentiality of Responses in Social Research: A Note on Strategies." *The American Sociologist* 6:308-11.

Chaudhuri, Arijit and Tasos C. Christofides. 2013. *Indirect Questioning in Sample Surveys*. Berlin, Germany: Springer.

Clark, S. J. and R. A. Desharnais. 1998. "Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model." *Psychological Methods* 3:160-68.

Coutts, Elisabeth and Ben Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)." *Sociological Methods and Research* 40: 169-93. doi:10.1177/0049124110390768.

Cruyff, Maarten J. L. F., Ulf Böckenholt, and Peter G. M. van der Heijden. 2016. "The Multidimensional Randomized Response Design: Estimating Different Aspects of the Same Sensitive Behavior." *Behavior Research* 48:390-99. doi: 10.3758/s13428-015-0583-2.

Elbe, Anne-Marie and Werner Pitsch. 2018. "Doping Prevalence Among Danish Elite Athletes." *Performance Enhancement and Health* 6:28-32. doi:10.1016/j.peh.2018.01.001.

Feth, Sascha, Monika Frenger, Werner Pitsch, and Patrick Schmelzeisen. 2017. *Cheater Detection for Randomized Response-Techniques: Derivation, Analyses and Application*. Saarbrücken, Germany: Saarland University Press.

Fox, James Alan. 2016. *Randomized Response and Related Methods: Surveying Sensitive Data. Quantitative Applications in the Social Sciences*, 2nd ed. Thousand Oaks, CA: Sage.

Fu, Haishan, Jacqueline E. Darroch, Stanley K. Henshaw, and Elizabeth Kolb. 1998. "Measuring the Extent of Abortion Underreporting in the 1995 National Survey of Family." *Family Planning Perspectives* 30:128-133.

Greenberg, B.G., R.R. Kuebler, J.R. Abernathy, and D.G. Horvitz. 1977. "Respondent Hazards in the Unrelated Question Randomized Response Model." *Journal of Statistical Planning and Inference* 1:53-60. doi:10.1016/0378-3758(77)90005-2.

Greenberg, Bernard G., Abdel-Latif A. Abul-Ela, Walt R. Simmons, and Daniel G. Horvitz. 1969. "The Unrelated Question Randomized Response Model: Theoretical Framework." *Journal of the American Statistical Association* 64:520-39.

Hoffmann, Adrian, Berenike Waubert De Puiseau, Alexander F. Schmidt, and Jochen Musch. 2017. "On the Comprehensibility and Perceived Privacy Protection of Indirect Questioning Techniques." *Behavior Research Methods* 49:1470-1483. doi:10.3758/s13428-016-0804-3.

Höglinger, Marc and Andreas Diekmann. 2017. "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT." *Political Analysis* 25:131-37. doi:10.1017/pan.2016.5.

Höglinger, Marc and Ben Jann. 2018. "More Is Not Always Better: An Experimental Individual-level Validation of the Randomized Response Technique and the Crosswise Model." *PLoS ONE* 13:e0201770. doi:10.1371/journal.pone.0201770.

Höglinger, Marc, Ben Jann, and Andreas Diekmann. 2016. "Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model." *Survey Research Methods* 10:171-87. doi:10.18148/srm/2016.v10i3.6703.

Kirchner, Antje. 2015. "Validating Sensitive Questions: A Comparison of Survey and Register Data." *Journal of Official Statistics* 31:31-59. doi:10.1515/JOS-2015-0002.

Kuk, Anthony Y. C. 1990. "Asking Sensitive Questions Indirectly." *Biometrica* 77: 436-38.

Lanke, Jan. 1975. "On the Choice of the Unrelated Question in Simmons' Version of Randomized Response." *Journal of the American Statistical Association* 70: 80-83. doi:10.1080/01621459.1975.10480265.

Lensvelt-Mulders, G. J. L. M., J. J. Hox, P. G. M. van der Heijden, and C. J. M. Maas. 2005. "Meta-analysis of Randomized Response Research: Thirty-five Years of Validation." *Sociological Methods and Research* 33:319-48. doi:10.1177/0049124104268664.

Lensvelt-Mulders, Gerty J. L. M. and Hennie R. Boeije. 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.

Leysieffer, Frederick W. and Stanley L. Warner. 1976. "Respondent Jeopardy and Optimal Designs in Randomized Response Models." *Journal of the American Statistical Association* 71:649-56. doi:10.1080/01621459.1976.10481541.

Ljungqvist, Lars. 1993. "A Unified Approach to Measures of Privacy in Randomized Response Models: A Utilitarian Perspective." *Journal of the American Statistical Association* 88:97-103.

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

Moshagen, Morten, Jochen Musch, Martin Ostapczuk, and Zengmei Zhao. 2010. "Reducing Socially Desirable Responses in Epidemiologic Surveys: An Extension of the Randomized-Response Technique." *Epidemiology* 21:379-82. doi:10.1097/EDE.0b013e3181d61dbc.

Ostapczuk, Martin, Morten Moshagen, Zengmei Zhao, and Jochen Musch. 2009. "Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry." *Journal of Educational and Behavioral Statistics* 34:267-87. doi:10.3102/1076998609332747.

Ostapczuk, Martin. 2011. "Improving Self-report Measures of Medication Non-adherence Using a Cheating Detection Extension of the Randomised-Response-Technique." *Statistical Methods in Medical Research* 20:489-503. doi:10.1177/0962280210372843.

Pitsch, Werner, Eike Emrich, and Markus Klein. 2007. "Doping in Elite Sports in Germany: Results of a www Survey." *European Journal for Sport and Society* 4: 89-102. doi:10.1080/16138171.2007.11687797.

Schröter, Hannes, Beatrix Studzinski, Pavel Dietz, Rolf Ulrich, Heiko Striegel, and Perikles Simon. 2016. "A Comparison of the Cheater Detection and the Unrelated Question Models: A Randomized Response Survey on Physical and Cognitive Doping in Recreational Triathletes." *PLoS One* 11:e0155765. doi:10.1371/journal.pone.0155765.

Soeken, Karen L, and George B. Macready. 1982. "Respondents' Perceived Protection When Using Randomized Response." *Psychological Bulletin* 92:487-89.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859-83. doi:10.1037/0033-2909.133.5.859.

Ulrich, Rolf, Hannes Schröter, Heiko Striegel, and Perikles Simon. 2012. "Asking Sensitive Questions: A Statistical Power Analysis of Randomized Response Models." *Psychological Methods* 17:623-41. doi:10.1037/a0029314.

Ulrich, Rolf, Harrison G. Pope, Léa Cléret, Andrea Petróczi, Tamás Nepusz, Jay Schaffer, Gen Kanayama, R. Dawn Comstock, and Perikles Simon. 2018. "Doping in Two Elite Athletics Competitions Assessed by Randomized-Response Surveys." *Sports Medicine* 48:211-19. doi:10.1007/s40279-017-0765-4.

Warner, S. L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60: 63-66.

Wolter, Felix and Peter Preisendörfer. 2013. "Asking Sensitive Questions: An Evaluation of the Randomized Response Technique versus Direct Questioning Using

Individual Validation Data." *Sociological Methods and Research* 42:321-53. doi: 10.1177/0049124113500474.

World Anti-Doping Agency. 2012. *2011 Laboratory Testing Figures*. Montreal.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. "Two New Models for Survey Sampling with Sensitive Characteristics: Design and Analysis." *Metrika* 67:251-63. doi:10.1007/s00184-007-0131-x.

Author Biographies

Fabiola Reiber is currently pursuing her PhD in psychology at the University of Tübingen, Germany, in the research training group Statistical Modeling in Psychology.

Harrison Pope is professor of psychiatry at Harvard Medical School in Boston, USA, and Chief of the Biological Psychiatry Laboratory at McLean Hospital in Belmont, USA. His research interests include the diagnosis and treatment of various psychiatric and substance abuse disorders, together with epidemiology and statistical methods.

Rolf Ulrich is professor of cognitive psychology at the University of Tübingen, Germany. His main research focuses on mathematical psychology and cognition.

Online Supplement

Appendix A

R-code for UQMC parameter estimation

```
library(bbmle)

## Parameter definition
# pi: proportion of honest carriers
# gm: proportion of cheaters
q <- 0.5       # prevalence of neutral attribute
p1 <- 0.75     # probability of the sensitive question in sample 1
p2 <- 0.25     # probability of the sensitive question in sample 2
n1 <- 1000     # sample size in sample 1
x1 <- 229      # frequency of "Yes"-responses in sample 1
lmb1 <- x1/n1  # proportion of a "Yes"-response in sample 1
n2 <- 1000     # sample size in sample 2
x2 <- 308      # frequency of "Yes"-responses in sample 2
lmb2 <- x2/n2  # proportion of a "Yes"-response in sample 2

## Calculation of model estimates

pi = (lmb2*(1-p1)-lmb1*(1-p2))/(p2-p1)
se_pi = sqrt((1/((p2-p1)^2))*((((1-p1)^2)*lmb2*(1-lmb2))/n2
                              + (((1-p2)^2)*lmb1*(1-lmb1))/n1))

gm = 1 - (lmb2 * p1 - lmb1 * p2)/(q*(p1 - p2))
se_gm = sqrt((1/((q^2)*((p1-p2)^2)))*(((p2^2)*lmb1*(1-lmb1))/n1
                                      + ((p1^2)*lmb2*(1-lmb2))/n2))

piPlusGm = pi + gm
se_piPlusGm = sqrt(se_pi^2 + se_gm^2 +
          (2/(q*(2*p1*p2 -p1^2 -p2^2))) *
          ((((p1^2-p1)*s_lmb2*(1-s_lmb2))/n2) +
             (((p2^2-p2)*s_lmb1*(1-s_lmb1))/n1)))

Estimate <- c(pi, gm, piPlusGm)
SE <- c(se_pi, se_gm, se_piPlusGm)
lowerCI <- Estimate - 1.96*SE
upperCI <- Estimate + 1.96*SE
UQMC_Estimates <- cbind(Estimate,SE,lowerCI,upperCI)
rownames(UQMC_Estimates) <- c('Pi','Gamma','(Pi + Gamma)')
UQMC_Estimates

## Likelihood ratio test for cheating
# negative log-Likelihood function:
l <- function (pi,gm){
  -((x1*log(p1*pi + (1-p1)*(1-gm)*q) +
       (n1-x1)*log(1-(p1*pi + (1-p1)*(1-gm)*q)) +
       x2*log(p2*pi + (1-p2)*(1-gm)*q) +
       (n2-x2)*log(1-(p2*pi + (1-p2)*(1-gm)*q))))}
```

```
# Unrestricted negative log-Likelihood:
lim <- 1e-15
ML1 <- mle2(l, start = list(pi = pi, gm = gm),
            method = 'L-BFGS-B',
            lower = c(lim,lim), upper = c((1-lim),(1-lim)))
summary(ML1)

# Restricted negative log-Likelihood with gamma = 0:
ML0 <- mle2(l, start = list(pi = pi),
            fixed = list(gm = 0),
            method = 'L-BFGS-B',
            lower = lim, upper = (1-lim))
summary(ML0)

# Likelihood ratio:
(LR = 2*((-ML1@min) - (-ML0@min)))
pchisq(LR, df = 1, lower.tail = F)
```

Appendix B

Parameter optimization and power analysis

Because the randomization procedure in all RRTs adds extra variance, it seems important to enhance the efficiency of the UQMC by optimizing the design parameters. These are the probabilities p_1 and p_2 of receiving the sensitive question in the first and second sample, the probability q of a "yes"-answer to the neutral question by an honest respondent, and the parameter f_1, which is defined as the proportion of all respondents assigned to the first sample, $f_1 = \frac{n_1}{n_1+n_2}$. In the following discussion, $p_2 = 1 - p_1$ is assumed. To suggest the optimal choice of these parameters, we examine their influence on the sum of the standard errors of the model parameters π and γ.

The sum of the standard errors is minimized if p_1 is as far from 0.5 as possible, as can be seen in Figure B1. Which value of p_1 minimizes the sum of standard errors does not depend on q and depends only slightly on f_1. Importantly, this also holds for other values of π and γ. When choosing an extreme value for p_1, however, one has to keep in mind psychological effects. A very high or very low probability of receiving one of the questions reduces the degree that one's anonymity is protected. Therefore a value of about 0.75 (or 0.25) seems advisable.

Figure B2 depicts the influence of the choice of q on the sum of the standard errors. This value becomes smaller with higher values chosen for q. The remaining parameters do not influence at which value of q the sum the standard errors becomes minimal. Again, this also holds for other values of π and γ. Choosing the optimal value for q by means of standard error minimization is not ideal because a very high prevalence of the neutral characteristic reduces the anonymity protection. Thus, we suggest a value around 0.7.

As can be seen in Figure B3 the value of f_1 should be chosen to be about 0.5, meaning that the respondents should be equally divided between the two subsamples. The minimum sum of standard errors depends on the choice of p_1. Specifically, when p_1 is higher than 0.5, the sum of the standard errors is smaller if the larger proportion of the sample is assigned to the second subsample, i.e. $f_1 < 0.5$. When p_1 is less than 0.5, and therefore p_2 is higher than p_1, the sum of standard errors can be minimized by $f_1 > 0.5$. Thus, the subsample with the lower probability of receiving the sensitive question should contain more respondents. The sum of standard errors depends on the choice of q in the sense that f_1 should be slightly further from 0.5 if the prevalence of the neutral question is lower (towards 0 if $p_1 > 0.5$ and towards 1 if $p_1 < 0.5$). Again, the same pattern holds for other values of π and γ. As the differences in the area around 0.5 are quite small, equal allocation to the two subsamples seems recommendable.

The influence of the design parameters on efficiency can also be observed by using a power analysis. Ulrich et al. (2012) provide a general framework for the power analysis of RRTs. Implementing the UQMC sampling

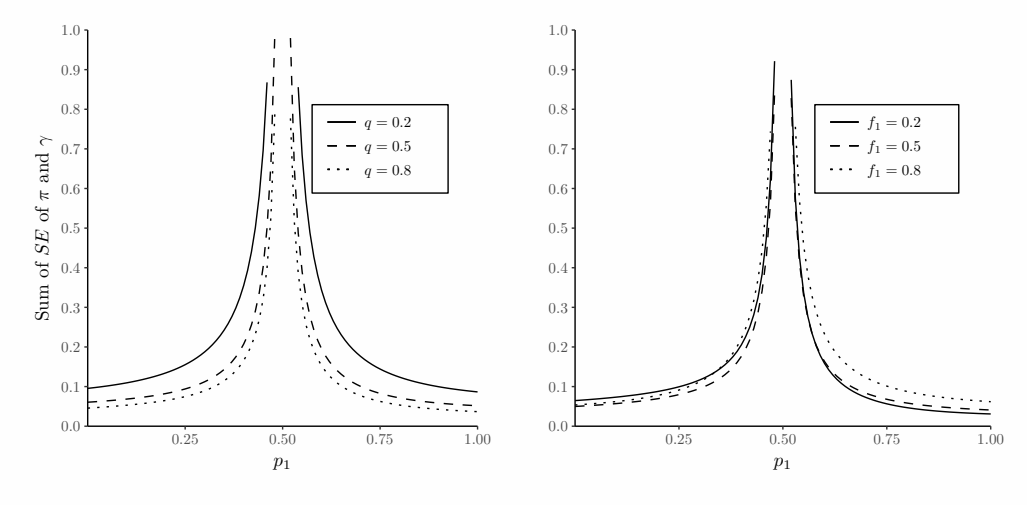


Figure B1: Curves showing the sum of the standard errors of π and γ as a function of the probability p_1 of receiving the sensitive question in the first sample, with $p_2 = 1 - p_1$. Parameters π and γ are kept constant at 0.2 and 0.3, respectively. In the left panel, f_1 is kept constant at 0.5 and the curves differ with respect to q. In the right panel, q is kept constant at 0.7 and the curves differ with respect to f_1.

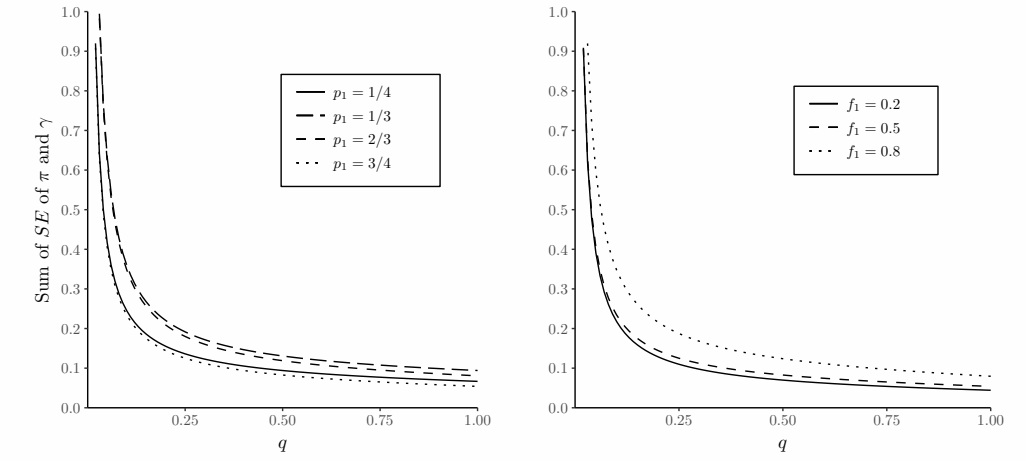


Figure B2: Curves showing the sum of the standard errors of π and γ as a function of the prevalence q of the neutral characteristic. Parameters π and γ are kept constant at 0.2 and 0.3, respectively. In the left panel, f_1 is kept constant at 0.5 and the curves differ with respect to p_1, with $p_2 = 1 - p_1$. In the right panel, p_1 is kept constant at 3/4, with $p_2 = 1 - p_1$, and the curves differ with respect to f_1.

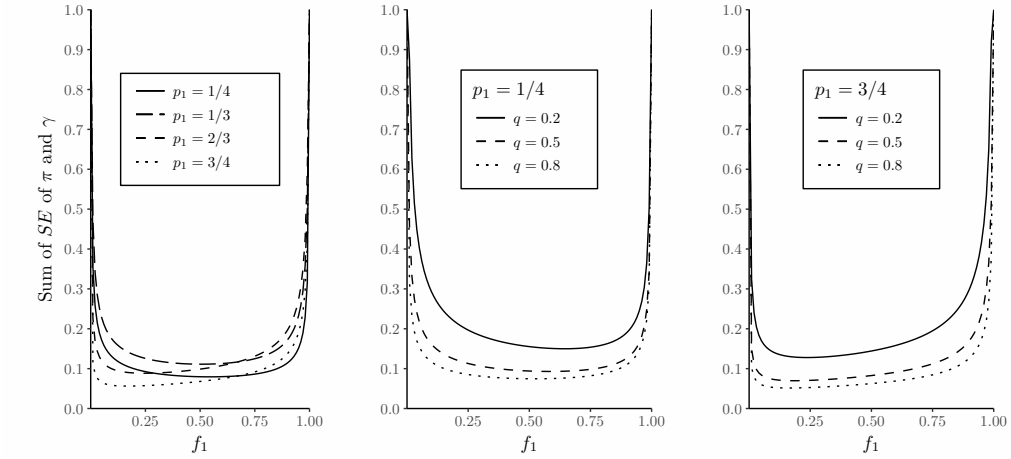


Figure B3: Curves showing the sum of the standard errors of π and γ as a function of the distribution parameter f_1, defined as the proportion of participants assigned to the first sample. Parameters π and γ are kept constant at 0.2 and 0.3, respectively. In the left panel, q is kept constant at 0.7 and the curves differ with respect to p_1, with $p_2 = 1 - p_1$. In the middle and right panels, p_1 is kept constant at 1/4 and 3/4, respectively, with $p_2 = 1 - p_1$, and the curves differ with respect to q.

variance of π yields the following statistical power for accepting the hypothesis H_1 that π takes on a certain value π_1,

$$\text{power of detecting } \pi_1 = \Phi\left(\frac{\pi_1 + z_\alpha \cdot \sqrt{Var(\hat{\pi}|\pi = 0)}}{\sqrt{Var(\hat{\pi}|\pi = \pi_1)}}\right). \tag{1}$$

The cumulative distribution function of the standard normal distribution is thereby denoted by Φ and α is the probability for erroneously rejecting the hypothesis H_0, that $\pi = 0$, with z_α being the $(100 \cdot \alpha)$th percentile of the standard normal distribution, $z_\alpha = \Phi^{-1}(\alpha)$. $Var(\hat{\pi}|\pi = 0)$ is given by Equation 8 with $\pi = 0$ and $Var(\hat{\pi}|\pi = \pi_1)$ is given by Equation 8 with $\pi = \pi_1$. The equation for hypotheses concerning parameter values of γ is composed correspondingly, by

$$\text{power of detecting } \gamma_1 = \Phi\left(\frac{\gamma_1 + z_\alpha \cdot \sqrt{Var(\hat{\gamma}|\gamma = 0)}}{\sqrt{Var(\hat{\gamma}|\gamma = \gamma_1)}}\right). \tag{2}$$

$Var(\hat{\gamma}|\gamma = 0)$ is derived from Equation 9 with γ set to 0. $Var(\hat{\gamma}|\gamma = \gamma_1)$ is derived from the same equation with the parameter set to $\gamma = \gamma_1$.

Power curves of the two model parameters as a function of sample size, given different choices of p_1, q and f_1, are depicted in Figures B4, B5 and B6, respectively. Naturally, the power for all parameters increases, as the parameter value itself increases. Across all variations of p_1, q and f_1, π has a higher power than γ. Choosing a more extreme value for p_1 (3/4 compared to 2/3) increases the power for π and decreases the power for γ. Increasing the value for q slightly decreases the power for π but strongly increases the power for γ. Higher values of f_1 lead to slightly higher power for π and markedly lower power for γ, if $p_1 > 0.5$. The opposing influence of p_1, q and f_1 on the power for π and the power for γ must be considered when setting the values of the design parameters for a given study, depending of the study's focus. Specifically, if the study's focus is on estimating the minimal prevalence of the sensitive attribute, p_1 should be chosen to be farther from 0.5, q should be smaller and f_1 higher (if $p_1 > 0.5$). If the study's focus is on the prevalence of cheating, however, p_1 should be chosen closer to 0.5, q higher and f_1 smaller (if $p_1 > 0.5$). If both outcome measures are of equal interest (e.g. because the maximal prevalence estimate is the focus), the above suggestions based on the sum of the standard errors can be applied.

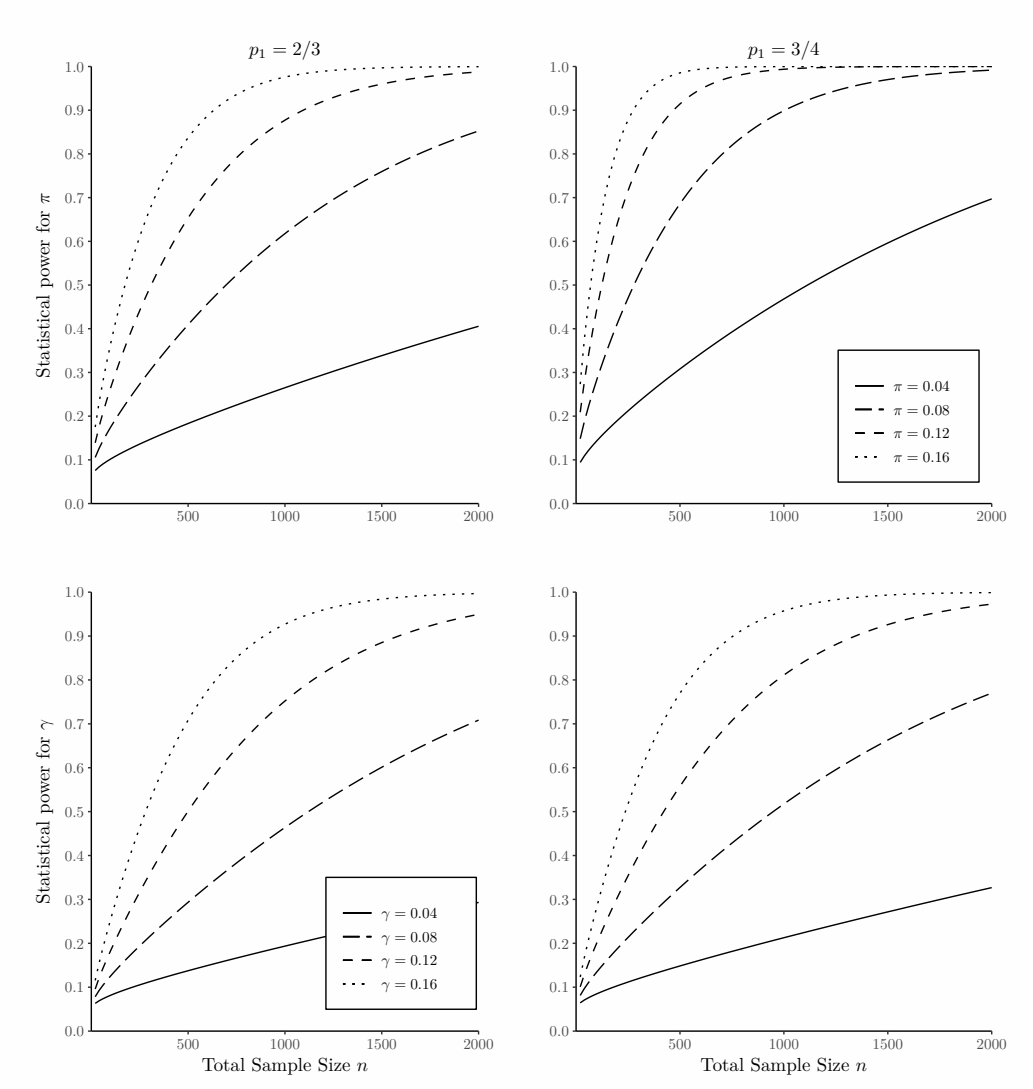


Figure B4: Power curves showing the effect of different choices of p_1. Each curve shows statistical power as a function of sample size n with $f_1 = 0.5$, i.e. $n_1 = n_2 = n/2$. The prevalence q of the neutral characteristic is set at 0.7. The probability p_1 of answering the sensitive question is set at $p_1 = 2/3$, with $p_2 = 1 - p_1 = 1/3$, for the left panels and at $p_1 = 3/4$, with $p_2 = 1 - p_1 = 1/4$, for the right panels. The top and bottom panels show power curves for π and γ, respectively. The curves within each panel differ with respect to the size of the respective model parameter.

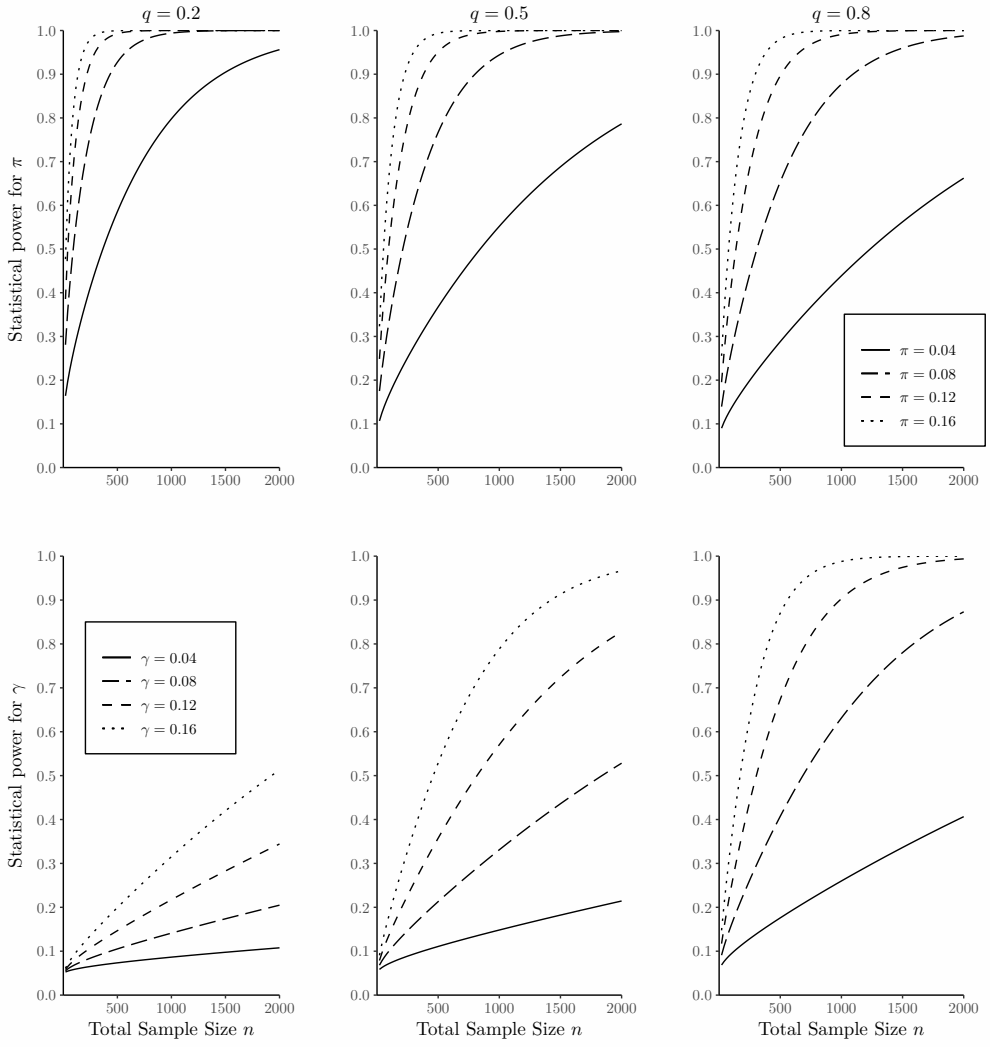


Figure B5: Power curves showing the effect of different choices of q. Each curve shows statistical power as a function of sample size n with $f_1 = 0.5$, i.e. $n_1 = n_2 = n/2$. The probability p_1 of answering the sensitive question is set at $p_1 = 3/4$, with $p_2 = 1 - p_1 = 1/4$. The prevalence q of the neutral characteristic is set at $q = 0.2$ in the left panels, $q = 0.5$ in the middle panels and $q = 0.8$ in the right panels. The top and bottom panels show power curves for π and γ, respectively. The curves within each panel differ with respect to the size of the respective model parameter.

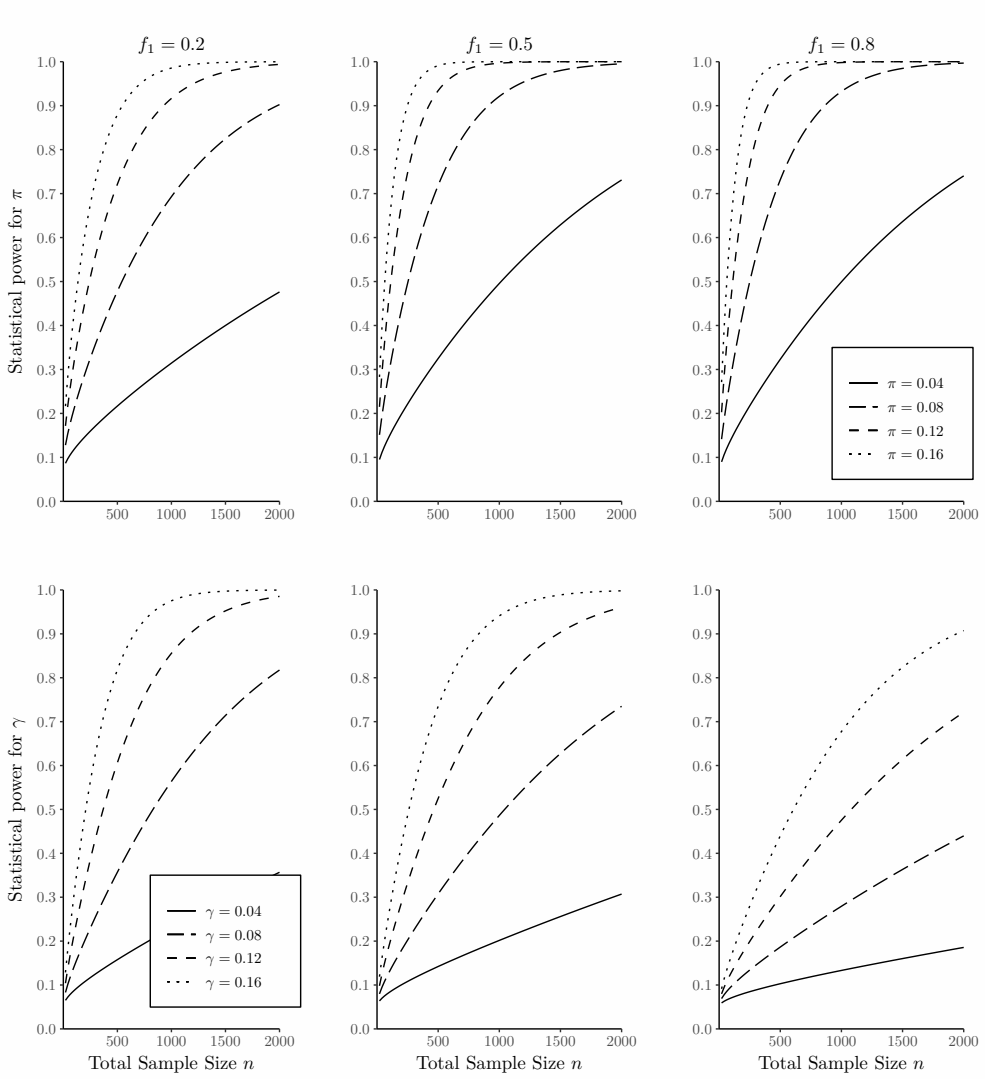


Figure B6: Power curves showing the effect of different choices of f_1. Each curve shows statistical power as a function of sample size n. The probability p_1 of answering the sensitive question is set at $p_1 = 3/4$, with $p_2 = 1 - p_1 = 1/4$, and the prevalence q of the neutral characteristic is set at $q = 0.7$. The distribution parameter f_1 determining the proportion of participants assigned to the first sample is set at $f_1 = 0.2$ in the left panels, $f_1 = 0.5$ in the middle panels and $f_1 = 0.8$ in the right panels. The top and bottom panels show power curves for π and γ, respectively. The curves within each panel differ with respect to the size of the respective model parameter.

For the numerical example in the main text with two samples (total size $n = 1000$ and $\alpha = .05$), the statistical power for obtaining a significant result with $\pi = 0.2$ is equal to 0.999, and that for $\gamma = 0.3$ would be even higher than 0.999.

Appendix C

R-code for parameter estimation in the testable UQMC

```
library(bbmle)

## Parameter definition
# pi: proportion of honest carriers
# gm: proportion of cheaters
q1 <- 0.7     # prevalence of neutral attribute in sample 1 and 3
q2 <- 0.3     # prevalence of neutral attribute in sample 2 and 4
p1 <- 0.75    # probability of the sensitive question in sample 1 and 2
p2 <- 0.25    # probability of the sensitive question in sample 3 and 4
n11 <- 500    # sample size in sample 1 (with p1 and q1)
x11 <- 129    # frequency of "Yes"-responses in sample 1
n12 <- 500    # sample size in sample 2 (with p1 and q2)
x12 <- 96     # frequency of "Yes"-responses in sample 2
n21 <- 500    # sample size in sample 3 (with p2 and q1)
x21 <- 204    # frequency of "Yes"-responses in sample 3
n22 <- 500    # sample size in sample 4 (with p2 and q2)
x22 <- 98     # frequency of "Yes"-responses in sample 4

## Numerical maximum likelihood estimation
# negative log-Likelihood function:
l <- function (pi,gm){
  lambda11 <- pi*p1 + (1-gm)*(1-p1)*q1
  lambda12 <- pi*p1 + (1-gm)*(1-p1)*q2
  lambda21 <- pi*p2 + (1-gm)*(1-p2)*q1
  lambda22 <- pi*p2 + (1-gm)*(1-p2)*q2
  l11 <- x11*log(lambda11) + (n11-x11)*log(1-lambda11)
  l12 <- x12*log(lambda12) + (n12-x12)*log(1-lambda12)
  l21 <- x21*log(lambda21) + (n21-x21)*log(1-lambda21)
  l22 <- x22*log(lambda22) + (n22-x22)*log(1-lambda22)
  l = -(l11 + l12 + l21 + l22)}

# Unrestricted likelihood:
lim <- 1e-15
ML1 <- mle2(l, start = list(pi = 0.5, gm = 0.2),
            method = 'L-BFGS-B',
            lower = c(lim,lim), upper = c((1-lim),(1-lim)))
summary(ML1)
Estimate <- c(coef(ML1),sum(coef(ML1)))
SE <- c(sqrt(vcov(ML1)[1,1]),sqrt(vcov(ML1)[2,2]),sqrt(sum(vcov(ML1))))
lowerCI <- Estimate - 1.96*SE
upperCI <- Estimate + 1.96*SE
UQMC_Estimates <- cbind(Estimate,SE,lowerCI,upperCI)
rownames(UQMC_Estimates) <- c('Pi','Gamma','(Pi + Gamma)')
UQMC_Estimates

## Likelihood-Ratio test for cheating
```

```
# Restricted likelihood with gamma = 0:
ML0 = mle2(l, start = list(pi = 0.5),
           fixed = list(gm = 0),
           method = 'L-BFGS-B',
           lower = lim, upper = (1-lim))
summary(ML0)
# Likelihood ratio:
(LR = 2*((-ML1@min) - (-ML0@min)))
pchisq(LR, df = 1, lower.tail = F)

## Chi^2 model test
f <- function (par){
  pi <- par[1]
  gm <- par[2]
  lambda11 <- pi*p1 + (1-gm)*(1-p1)*q1
  lambda12 <- pi*p1 + (1-gm)*(1-p1)*q2
  lambda21 <- pi*p2 + (1-gm)*(1-p2)*q1
  lambda22 <- pi*p2 + (1-gm)*(1-p2)*q2
  y11 <- (x11 - n11*(lambda11))^2 / (n11*(lambda11))
  n11 <- ((n11-x11) - n11*(1-lambda11))^2 / (n11*(1-lambda11))
  y12 <- (x12 - n12*(lambda12))^2 / (n12*(lambda12))
  n12 <- ((n12-x12) - n12*(1-lambda12))^2 / (n12*(1-lambda12))
  y21 <- (x21 - n21*(lambda21))^2 / (n21*(lambda21))
  n21 <- ((n21-x21) - n21*(1-lambda21))^2 / (n21*(1-lambda21))
  y22 <- (x22 - n22*(lambda22))^2 / (n22*(lambda22))
  n22 <- ((n22-x22) - n22*(1-lambda22))^2 / (n22*(1-lambda22))
  f = sum(y11,n11,y12,n12,y21,n21,y22,n22)}
xsq <- optim(par = c(0.5,0.5), fn = f,
             method = 'L-BFGS-B',
             lower = c(lim,lim), upper = c((1-lim),(1-lim)))
xsq$value
pchisq(xsq$value, df = 2, lower.tail = F)
```

B.2 Article II

Article in press.

Official citation:

Reiber, F., Bryce, D., & Ulrich, R. (in press). Self-protecting responses in randomized response designs: A survey on intimate partner violence during the COVID-19 pandemic. *Sociological Methods and Research*.

Self-protecting responses in randomized response designs: A survey on intimate partner violence during the COVID-19 pandemic

Fabiola Reiber, Donna Bryce, and Rolf Ulrich

University of Tübingen

Author Note

This research was funded by the Deutsche Forschungsgemeinschaft (DFG), grant 2277, Research Training Group "Statistical Modeling in Psychology" (SMiP). Data and all analysis scripts are available on the Open Science Framework (https://osf.io/9bna3/).

Correspondence concerning this article should be addressed to Fabiola Reiber, Schleichstraße 4, 72076 Tübingen. E-mail: fabiola.reiber@uni-tuebingen.de

Abstract

Randomized response techniques (RRTs) are applied to reduce response biases in self-report surveys on sensitive research questions (e.g., on socially undesirable characteristics). However, there is evidence that they cannot completely eliminate self-protecting response strategies. To address this problem, there are RRTs specifically designed to measure the extent of such strategies. Here, we assessed the recently devised cheating extension of the unrelated question model (UQMC, Reiber, Pope, & Ulrich, 2020, Soc. Methods and Research) in a preregistered online survey on intimate partner violence (IPV) victimization and perpetration during the first contact restrictions as containment measures for the outbreak of the COVID-19 pandemic in Germany in early 2020. The UQMC accounting for self-protecting responses described the data better than its predecessor model which assumes instruction adherence. The resulting three-month prevalence estimates were alarmingly high ($\sim$ 10 percent) and we found a high proportion of self-protecting responses in the group of female participants queried about IPV victimization. However, unexpected results concerning the differences in prevalence estimates across the groups queried about victimization and perpetration highlight the difficulty of investigating sensitive research questions even using methods that guarantee anonymity and the importance of interpreting the respective estimates with caution.

Keywords: sensitive research questions, randomized response techniques, cheater detection, intimate partner violence, self-protecting responses

Word count: 7998

Self-protecting responses in randomized response designs: A survey on intimate partner violence during the COVID-19 pandemic

Many social and psychological phenomena of high societal relevance are difficult to investigate empirically because of their sensitive nature. For instance, the German news broadcaster Tagesschau recently reported an alarming increase in the incidence of intimate partner violence (IPV) in criminal statistics during the ongoing COVID-19 pandemic (Emundts 2020). However, criminal statistics are assumed to underestimate the actual numbers, because they only capture legally reported cases and the dark figure, that is, the number of non-registered cases, might substantially exceed these numbers. Problematically, the dark figure of cases of IPV is difficult to investigate because it is a highly stigmatized topic (e.g., Ellsberg et al. 2001; Gracia 2004). Both victimization and perpetration of IPV are perceived as socially undesirable and reporting is associated with negative consequences (e.g., Birkel and Guzy 2015; Franke et al. 2004). Social desirability and fear of stigmatization or other negative consequences can influence response behavior in surveys and interviews (Tourangeau and Yan 2007). Specifically, survey respondents can be inclined not to respond at all, especially if they carry the investigated undesirable or stigmatized attribute, or to give an untruthful self-protecting response. Although both these behaviors are employed to disguise ones own individual status, they bias group level estimates of dark figures as well. The consequence is that the extent of societal problems such as IPV can be underestimated by surveys (Tourangeau and Yan 2007). This impairment concerns a variety of research fields in the social sciences that address sensitive characteristics.

Randomized Response Techniques

To overcome self-protecting response strategies in surveys on sensitive attributes, randomized response techniques (RRT, Warner 1965) were developed to assure the protection of the respondents' anonymity. Specifically, a randomization device (such as a die) is employed to ambiguate single responses and thus make them inconclusive towards the carrier status of a single respondent. For instance, in the unrelated question model

(UQM, Greenberg et al. 1969) version of the RRT, a randomization device decides whether a respondent shall answer the sensitive question S of interest, such as "Have you ever been physically assaulted by a partner?" or an unrelated neutral question N, such as "Is your mother's birthday in the first half of the year?" In case of employing a die as randomization device, the instruction could be to answer the sensitive question S, if the die comes up 1 through 4, and the neutral question N, if it comes up 5 or 6. Importantly, only the response to either question but not the outcome of the randomization is reported. Therefore, a "Yes"-response could either mean that the respondent has been physically assaulted by a partner or that their mother's birthday is in the first half of the year. Consequently, it remains concealed whether a specific respondent was physically assaulted by a partner and, theoretically, respondents have no reason to employ self-protecting response strategies that could bias prevalence estimates. In the current study, we applied such a technique to estimate the prevalence of IPV during the first COVID-19 related contact restrictions in Germany in spring 2020.

Importantly, it is possible to compute these prevalence estimates using the known probabilities underlying the questioning design. Figure 1 depicts the probabilities underlying "Yes"- and "No"-responses in the UQM. A "Yes"-response can come (a) from a respondent who was instructed to respond to the sensitive question S with probability p and carries the sensitive attribute with probability π and (b) a respondent who was instructed to respond to the neutral question N with probability $(1 - p)$ and carries the neutral attribute with probability q. Therefore, the overall probability to respond "Yes" is

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

The randomization probability p is known - in the example using a die, above, it is $p = 4/6 = .67$. The neutral question can be chosen such that the neutral prevalence q is also known. In the example above it is $q \approx .5$, assuming a uniform distribution of birthdays across the year, which is a reasonable assumption based on the birthdate records over the last 50 years in Germany (Statistisches Bundesamt 2020). The probability λ to respond "Yes" can be estimated from the proportion of "Yes"-responses

in a sufficiently large sample such that Equation 1 can be rearranged to estimate the prevalence of the sensitive attribute:

$$\hat{\pi}_{UQM} = \frac{\hat{\lambda} - (1-p) \cdot q}{p}. \tag{2}$$

Because the respondents' anonymity is protected, this estimate is expected to be less biased due to self-protecting response strategies. In fact, there is evidence that RRT applications elicit prevalence estimates that are less biased towards the socially desirable response option (e.g., Moshagen et al. 2010; Ulrich et al. 2018; Wimbush and Dalton 1997) and closer to a known true prevalence (e.g., Horvitz, Greenberg, and Abernathy 1976; Van Der Heijden et al. 2000).

Non-Adherence to Instructions in Randomized Response Techniques

However, there are reasons to doubt that even with the RRT there is full honesty in responding. A number of studies did not find RRT estimates to be more valid than those from studies using direct questions (Holbrook and Krosnick 2010; e.g. Höglinger and Diekmann 2017; Höglinger and Jann 2018). A possible explanation for this finding is that the instructions of the RRT are difficult to understand (Hoffmann et al. 2017) and there is still a lack of trust in the anonymity protection (Höglinger, Jann, and Diekmann 2016). One way to address this problem is to increase comprehensibility (e.g., Meisters, Hoffmann, and Musch 2020). Another way is to quantify the extent of non-compliance with instructions. In this vein, some RRT extensions, such as the cheater detection model (CDM, Clark and Desharnais 1998) or the stochastic lie detector (Moshagen, Musch, and Erdfelder 2012) include parameters for specific types of instruction non-adherence. Especially the CDM has been applied in a number of studies (e.g., Elbe and Pitsch 2018; Moshagen et al. 2010; Ostapczuk 2011; Pitsch, Emrich, and Klein 2007; Schröter et al. 2016). It is based on another RRT variant, the forced response technique (Boruch 1971), which is similar to the UQM. The only difference is that the alternative to the sensitive question is not a neutral question but the instruction to respond "Yes." In the CDM, respondents are considered to be either honest and follow the instructions or to be cheaters and give a "No"-response irrespective of the outcome of the randomization and

their carrier status. The latter can serve to evade being seen as a carrier of the sensitive attribute and has thus been termed a self-protective response strategy (Böckenholt and Van Der Heijden 2007). Based on this categorization, two parameters can be estimated: The proportion of cheaters γ and the proportion of honest carriers π_{CDM}.[1] To allow for the estimation of both these parameters, two independent estimates of the probability of a "Yes"-response are required. To that end, two independent samples are assessed using varying levels of the randomization probability p. Studies applying this design found substantial proportions of cheating (Elbe and Pitsch 2018; Moshagen et al. 2010; Ostapczuk 2011; Pitsch, Emrich, and Klein 2007; Schröter et al. 2016). Thus, it seems to be reasonable to include a cheating parameter in RRTs.

It is important to note, however, that the CDM still makes strong assumptions about the nature of instruction non-adherence. For instance, the varying levels of the randomization probability p, which are employed to enable the estimation of the cheating parameter, are assumed not to influence responses. In practice, different randomization probabilities could influence the subjective anonymity protection and thereby the probability to cheat.[2] Unfortunately, assumptions, like this assumption of randomization probability independence, are not testable within the CDM. The variation of p across two independent samples allows for the estimation of both parameters. However, the resulting model is saturated, which means that it is not possible to assess model fit or, for instance, test whether cheating differs between the subsamples.[3]

The Unrelated Question Model - Cheating Extension

The recently proposed unrelated question model - cheating extension (UQMC, Reiber, Pope, and Ulrich 2020) transfers the CDM's concept of cheating to the UQM's design. The reason for devising this extension was that the psychological acceptability of the UQM has been found to be superior to that of the forced response method (Höglinger, Jann, and Diekmann 2016). As such, the UQM can be seen as less fallible to self-protecting responses, since there is no response option that clearly rules out being a carrier of the sensitive attribute (one could respond "No" to the neutral question and still

be a carrier of the sensitive attribute). However, also in the UQM "No" can be seen as a self-protecting response since the conditional likelihood of being a carrier is always lower given a "No"- than given a "Yes"-response. Additionally, and probably more intuitively from a respondent's perspective, a "No"-response to the neutral question can naively be interpreted as a response with which being a carrier of the sensitive attribute is negated. Thus, it is worthwhile to investigate whether cheating occurs in the UQM as well.

Another major advantage of embedding the cheating concept within the UQM is that it is possible to test the model's assumptions. In contrast to the CDM, the UQM incorporates a second design parameter that can be varied, namely the prevalence of the neutral attribute q. Therefore, four independent samples can be assessed and the degrees of freedom gained make the model testable.

From Figure 2 the probability of a "Yes"-response in sample i in the UQMC is

$$\lambda_i = (1-\gamma) \cdot [p_i \cdot \epsilon + (1-p_i) \cdot q_i] \,. \tag{3}$$

Only respondents who do not cheat would respond "Yes." If they are assigned to the sensitive question S (with randomization probability p_i in sample $i \in \{1,2,3,4\}$), honest respondents answer "Yes" with probability ϵ, that is the true prevalence of the sensitive attribute. If they are assigned to the neutral question N (with probability $1-p_i$ in sample i), honest respondents answer "Yes" with probability q_i, that is the prevalence of the neutral attribute in sample i. Following the logic of the CDM, $\pi_{UQMC} = (1-\gamma) \cdot \epsilon$ is the prevalence of honest carriers, that is, the joint probability of not cheating and of being a carrier of the sensitive attribute. Therefore,

$$\lambda_i = p_i \cdot \pi_{UQMC} + (1-\gamma) \cdot (1-p_i) \cdot q_i. \tag{4}$$

There are no closed form equations to compute estimates for π_{UQMC} and γ from the four samples' estimated probabilities of a "Yes"-response $\hat{\lambda}_i$.[4] Instead, the parameters must be estimated using numerical likelihood optimization.

The development and properties of the UQMC are described in more detail in Reiber, Pope, and Ulrich (2020). However, the validity of the model has so far not been

investigated empirically. The aim of the present study was, therefore, to test the UQMC's validity in an empirical investigation and to assess whether it provides an advantage over its predecessor model, the original UQM.

Present Study

There are different approaches to assessing a model's validity. One widely accepted approach is to compare prevalence estimates with a known criterion, optimally on an individual level (e.g., Hoffmann et al. 2015). Unfortunately, the prevalences of highly sensitive topics are often not known, especially not on an individual level. Therefore, studies using this approach often use experimentally induced behaviors for sensitive characteristics, such as cheating for extra pay-off in the survey (e.g., Hoffmann et al. 2015). However, these characteristics differ from those addressed in typical RRT applications because RRTs are most useful for investigating highly sensitive topics (see Lensvelt-Mulders et al. 2005). Therefore, in the present study, we chose another approach to assess the UQMC's validity. Specifically, to test whether the cheating extension provides a more realistic model than the original UQM, the occurrence of cheating in a survey sample was tested and the general model fit assessed and compared to that of the original UQM. Since we wanted the study to resemble a typical RRT application, we assessed a highly sensitive characteristic, that is, intimate partner violence (IPV).

Intimate partner violence

The term IPV incorporates physical, sexual, and psychological violence and controlling behavior towards a former or current intimate partner (World Health Organization 2012). The present study focused on physical IPV because this facet is easiest to explain to respondents in an online survey using concrete examples of behavior (here "shoving, slapping, hitting, kicking, or punching"). Other forms of violence, such as "humiliation" as an example for psychological violence, can be much more difficult to identify as violence for survey respondents. The lifetime prevalence of physical and sexual IPV against women in the EU was estimated to be 22 percent in a survey by the European Agency for Fundamental Rights (2014) and the 12 month prevalence to be

4 percent. The Federal Criminal Police Office reported 141,792 cases of attempted or committed IPV in Germany in 2019 (Bundeskriminalamt 2020), that is 17.3 percent of all reported violent crimes (incl. non-partner violence). Of these, 61.2 percent were actual bodily harm ("einfache Körperverletzung"). Of all IPV victims in the criminal statistic, 26,889 were male and 114,903 female. Numbers like these contribute to the assumption that IPV is mainly perpetrated by men against women. However, there is an ongoing debate about gender (a-)symmetry with respect to varying characteristics of both the specific type of violence investigated and the survey method (e.g., Archer 2000; Johnson 2006; Kimmel 2002). For instance, the lifetime prevalence of physical IPV victimization in the US was estimated to be 30.6 percent among women and 31.0 percent among men in the National Intimate Partner and Sexual Violence Survey (Smith et al. 2018), whereas the prevalence of severe physical violence victimization was estimated to be 21.4 percent among women and 14.9 percent among men. Generally speaking, estimates vary strongly between studies due to differences in the applied measures and samples (see, e.g., Devries et al. 2013; Garcia-Moreno et al. 2006; Kimmel 2002; Waltermaurer 2005).

As outlined in the beginning of this paper, IPV is a highly sensitive topic, that is, exactly the kind of topic for which RRTs were developed and thus suitable for the present validation study. Furthermore, several articles in scientific journals and the media reported rising numbers of IPV in the context of the impact of the spread of the coronavirus disease 2019 (COVID-19), which was declared a pandemic by the World Health Organization in March 2020 (e.g., Bradbury-Jones and Isham 2020; Emundts 2020; Jarnecke and Flanagan 2020). The pandemic, and the measures implemented to contain it, are believed to foster factors associated with IPV, such as increased material worries or restricted possibilities to avoid the perpetrator and seek help (Usher et al. 2020). The rising numbers of criminal reports corroborate this argumentation, highlighting the relevance of investigating the dark figure of IPV. Thus, we applied the UQMC to estimate the prevalence of physical IPV during the first COVID-19 contact restrictions in spring and early summer 2020 in Germany to assess the models empirical adequacy in a context which is relevant and representative of RRT applications.

Sensitivity manipulation

To further test the UQMC we employed an experimental manipulation of the sensitivity of the question: Respondents were either queried about their role as victim of IPV or as perpetrator of IPV. As mentioned before, both roles are associated with stigma and perceived as socially undesirable. However, being a perpetrator is even legally incriminating and has been shown to have an even stronger association with social desirability (Sugarman and Hotaling 1997). We therefore expected the question on perpetration of IPV to be more sensitive than the question on victimization of IPV. Consequently, we expected cheating to be more pronounced in the subsample queried about perpetration. We restricted our sample to participants who were, at the time of the investigation, in a romantic relationship with exactly one person. This way, the true proportion of perpetrators and victims should be equal in our sample. Assuming that differences in honesty of responding would be captured by the cheating parameter, any differences between estimates of the prevalence of honest carriers should be reflected in complementary differences in cheating estimates. Specifically, we expected that, if there was significant cheating, (a) it would be estimated to be higher in the subsample queried about perpetration and that (b) the prevalence of honest carriers would be estimated to be lower in the subsample queried about perpetration. If there was no significant cheating, the prevalence of honest carriers was expected not to differ between the subsamples.

Objective

To summarize, the aim of the present study was to assess the empirical validity of the recently devised cheating extension of the unrelated question model (UQMC; Reiber, Pope, and Ulrich 2020) in a survey on the prevalence of IPV. To this end, the fit of the UQMC was compared to that of its predecessor, the original UQM, and the occurrence of cheating was tested. Additionally, the queried IPV role was experimentally manipulated to investigate the differential influence of the question sensitivity on cheating.

Methods

Participants

Participants were recruited from the participant panel of the market research institute respondiAG with a target sample size of 4800. Quotas to approximate population proportions were installed for gender, age and highest educational achievement. The target quotas are depicted in Table 1.

To participate, respondents had to declare that they were at least 18 years old and currently in a relationship with one person. Participants who indicated that they were younger than 18 years or that they were in no romantic relationship or in a romantic relationship of equal importance with more than one person, were screened out before answering the questionnaire. Participants who fell into an age, gender, or education level category for which the quota was already full were also screened out. To ensure data quality, an attention check question was included in the questionnaire and participants who failed to answer this question correctly were screened out before finishing the questionnaire. Section A of the online supplemental materials contains information on participant dropout per page.

The total data set consisted of 4804 participants who reached the last page of the questionnaire. Of these, 1326 participants who failed to answer the second of two training questions correctly, described in more detail later, and 183 participants with a mean response time less than half of the median of each page (relative speed index, Leiner 2019) were excluded from the analysis.

After exclusion, the final sample consisted of 3295 participants with a mean age of 47.35 ($SD = 15.44$); 1732 (52.56 percent) indicated that their gender was female and 10 (0.30 percent) indicated diverse. Age and gender categories approximated the target quotas very well as can be seen in Table 1. The target quotas for education could not be attained because too few participants in the lower education level groups were reached, as can be seen in the table. More specifically, people with a high education level were

over-represented in the sample.

Of the final sample, 1618 (49.10 percent) answered the question on victimization of IPV. They did not differ from those who answered the question on perpetration with respect to age, $t(3293) = 1.79$, $p = .073$, gender, $p = .623$, Fisher's exact test, or highest educational achievement, $\chi^2(5) = 7.22$, $p = .205$.

Design

The prevalence of IPV was assessed using one of two sensitive questions. Participants were either asked if they had experienced IPV (victimization role) or if they had committed IPV (perpetration role). They were randomly assigned to either of these role conditions, which only differed in the phrasing of the sensitive question itself. The sensitive question in the victimization condition read: "Have you, in your current relationship, since March 23rd, been intentionally physically assaulted by your partner?"[5] In the perpetration condition it read: "Have you, in your current relationship, since March 23rd, intentionally physically assaulted your partner?"[6] The date March 23rd was chosen because it marks the date on which contact restrictions as a means of containing the spread of COVID-19 were officially announced in Germany. Participants were reminded of this context before answering the IPV question.

The sensitive question was presented within a UQMC design. Specifically, participants were instructed to think of a person whose birthday they knew and keep that birthday in mind. If the birthday was within a certain range of days in a month, they were asked to respond to a neutral question A and if it was in the remaining days of a month they were asked to respond to the sensitive question B. This range of days in a month determined the randomization probability p to respond to the sensitive question B. It was varied between participants on two levels: 1st to 10th day ($p_1 = 2/3$) or 1st to 20th day ($p_2 = 1/3$). The sensitive question B was the question on IPV. The neutral question A asked whether the memorized birthday was in a certain range of months in the year. This question was also varied between participants on two levels to obtain two

neutral prevalences q: January to September ($q_1 = .75$) or January to March ($q_2 = .25$). The birthday probabilities were reconciled with German birth rate records since 1950 (Statistisches Bundesamt 2020). Participants were instructed to mark their response ("Yes" or "No") to the question they were assigned to and reminded that only they knew which question they were responding to.

The combination of the factors role condition, p, and q resulted in eight groups, which are depicted in Table 2. This design was implemented to allow for testing the UQMC's assumptions and model fit.

Procedure

The questionnaire was created using the software SoSci Survey (Leiner 2020). The survey administration period lasted from June 29th to July 15th 2020.

On being directed to the survey via a link distributed by respondiAG, participants received general information about the study and were asked to confirm their informed consent. Only participants who did so were directed to following pages of the questionnaire. First, they answered demographic questions on age, gender, highest educational achievement, and relationship status for screening and quota checks. Then they received detailed instructions on the UQMC together with an example involving the abuse of illicit drugs as the sensitive question. All participants completed two UQMC training questions. In each one they received a vignette of a fictional person who is asked whether they took illicit drugs within a UQMC design. This design was, for each participant, exactly the same as in the question on IPV, with the difference that the sensitive question was on taking illicit drugs instead of IPV and that participants did not have to answer for themselves but for the fictional person. This way, it was possible to provide feedback on the response, because the correct answer was known from the vignette. In both cases the correct response was "Yes" but only once because the fictional person had taken illicit drugs. In the other case the correct response was "Yes" as an answer to the neutral question although the person had not taken illicit drugs. This was

meant to demonstrate the anonymity protection. Participants who did not respond correctly to the second training question were later excluded from the analysis. After completing the training questions participants were informed about the definition of physical IPV and the relevant time period beginning March 23rd, that is, during the first contact restrictions due to the COVID-19 pandemic in Germany. Participants then completed the IPV question within the UQMC design in one of the above described eight conditions. On the following two pages, participants were asked to provide information on their living conditions during the considered time period. A list of the questions is in Section B of the online supplemental materials. Among the additional questions was an attention check ("Which of the following cities is not in Germany?" - Berlin, Hamburg, Cologne, *London*, Frankfurt, Munich). Participants were expected to be able to answer this question if they were paying attention and, thus, participants who failed to answer correctly were excluded from the survey. On the last survey page participants were provided helpline information for victims and perpetrators of IPV before being redirected to the site of respondiAG.

Data analysis[7]

Data exclusion

Participants who responded incorrectly to the second of two UQMC training questions were excluded from the analysis. Because the training questions were very similar to the IPV question, failing to answer the second training question correctly was taken as an indicator for unreliable statements in the IPV question. We excluded 1326 participants, that is 27.60 percent, because they did not meet this criterion. This is a surprisingly high number, especially because only 866, that is 18.03 percent failed to respond correctly to the first training question. Even though it is unclear why so many participants failed to answer the second training question correctly, this casts doubts on the validity of this criterion. However, the main results of this study are not strongly affected by in- or exclusion of the respective participants. Section C of the online supplemental materials contains the results of the analyses including participants who

answered the second training question incorrectly. Differences between the two analyses are largely explainable by differences in power.

Additionally, 183 fast respondents with a relative speed index (RSI, Leiner 2019) above 2.00 were excluded from the main analysis. The RSI measures the participants' screen processing times relative to the screens' median processing times averaged across all screens. Therefore, an RSI above two indicates that the participant, on average, proceeded to the next screen twice as fast as the median of respondents. This can be used as an indicator for careless responding (Leiner 2019).

The participants' gender was included as a control variable in most analyses because of the inconclusive findings in the literature concerning its association with IPV. Whenever it was included, participants who indicated diverse gender were excluded from the analyses because the group was too small to be included as separate factor level.

Parameter estimation and assessment of model fit

All models were fitted by optimizing the G^2 statistic, which is a measure for the deviance of observed and model predicted response frequencies, using the method by Nelder and Mead (1965) implemented in the function *optim* provided in the *stats* R-package. In a first step, the sample was split into four subsamples following from the combination of the two factors Gender (excluding diverse gender) and Role. For each of these subsamples, the IPV prevalence π_{UQM} in the UQM and the prevalence of honest carriers π_{UQMC} and the cheating prevalence γ in the UQMC were estimated separately. The model fit of both models in the four subsamples was assessed using the asymptotically χ^2-distributed G^2 statistic. Additionally, the overall fit of both models was assessed by summing the G^2 values from the subsamples and thereby making use of the additivity property of χ^2-distributed values. The fit of the UQM and UQMC was compared using G^2 difference tests and the Akaike and Bayesian information criterion (AIC and BIC), which set the model fit in relation to model complexity using penalty terms depending on the number of free parameters.

Analysis of role conditions

To test the influence of the role manipulation within the UQMC, a full logistic model including baseline cheating and honest carrier prevalence parameters as well as parameters for the factor Role (victimization vs. perpetration), the factor Gender (male vs. female), and interaction terms was fitted by optimizing the G^2 statistic:

$$logit(\gamma) = \gamma_0 + \gamma_1 \cdot Gender + \gamma_2 \cdot Role + \gamma_3 \cdot Gender \cdot Role, \tag{5}$$

$$logit(\pi) = \pi_0 + \pi_1 \cdot Gender + \pi_2 \cdot Role + \pi_3 \cdot Gender \cdot Role. \tag{6}$$

The factor Role was dummy coded with victimization as reference category. Therefore, the effects of Role (γ_2 and π_2) can be interpreted as the difference in γ and π_{UQMC} between the victimization and perpetration conditions. Gender was included as control variable and was effect coded in order that the mean effects of Role across the levels of Gender could be estimated. This full model was compared to restricted models using G^2 difference tests. Specifically, we successively restricted the interaction effects of Role and Gender π_3 and γ_3 and the main effects of Role π_2 and γ_2 to be equal to 0.[8] We compared each resulting model to the previous more complex model with respect to G^2, AIC and BIC differences.

All analysis scripts and a preregistration of the study are on the Open Science Framework (OSF, https://osf.io/9bna3/).

Results

Estimation and model fit

Table 3 depicts UQM and UQMC parameter estimates and their standard errors for the four subsamples following from the allocated role condition and participant gender. Due to the beginning of the contact restrictions on March 23rd and the survey administration period from June 29th to July 15th, the estimates refer to 3 to 3.5 month IPV prevalences. The estimates for physical IPV without accounting for cheating, that is

$\hat{\pi}_{UQM}$, lie between 7.43 percent and 11.56 percent. Applying the UQMC, in three of the four subsamples cheating is estimated to be close to 0 and, correspondingly, the prevalence estimates differ only slightly between the UQM and the UQMC. Only the IPV prevalence estimates among female participants queried about IPV victimization differ strongly between the models, with an honest carrier prevalence estimate of $\hat{\pi}_{UQMC} =$ 17.56 percent and a cheating estimate of $\hat{\gamma} = 30.17$ percent.

The latter outcome is consistent with the results of the model comparison in Table 4. The model fit of the UQMC is better than that of the UQM with respect to all model comparison criteria only for this subsample. Within this subsample, the UQM's G^2 statistic is highly significant, indicating insufficient model fit. The G^2 statistic indicates sufficient model fit for both models in all other subsamples.

The last two rows in Table 4 depict the overall model fit of the UQM and the UQMC using the G^2 sums over the subsamples. The UQMC's G^2 test indicates a reasonable model fit, whereas the UQM's G^2 value is significant, indicating insufficient model fit. The significant G^2 difference test supports the superiority of including the UQMC's cheating parameter.

When cheating is taken into account, the plausible range of estimates for the prevalence of IPV is indicated by the interval $[\hat{\pi}_{UQMC}; \hat{\pi}_{UQMC} + \hat{\gamma}]$. As it is typical for RRTs, the standard errors of both bounds are quite high, despite the large sample size. To accommodate this uncertainty, the confidence intervals of the bounds need to be taken into account. Descriptively, the resulting range is highest in the subsample of female participants queried about victimization, with 95 percent CIs ranging from 12.15 to 62.88, and lowest in the subsample of female participants queried about perpetration, with 95 percent CIs ranging from 2.70 to 22.60. The subsamples of male participants are very similar with respect to the UQMC's estimates with 95 percent CIs ranging from 8.66 to 39.06 in the victimization condition and from 5.83 to 31.32 in the perpetration condition. Note that these intervals are relatively wide because they incorporate the cheating estimates. Thus, they do not only indicate unsystematic uncertainty in the estimates but

the systematic influence of a specific response style on the estimates. Therefore, despite being wide these confidence intervals are indicative of relevant information, which is ignored by the original UQM and most RRTs as well as direct questioning techniques.

The estimates do not indicate a clear gender effect. An effect of the role condition is more apparent in the UQMC's estimates, especially in the subsample of female participants. Consequently, the effects of the role condition on the UQMC's estimates and their interactions with gender were tested in a logistic model. The main effects of gender were not specifically tested because there were no founded expectations due to the inconclusive findings on the gender differences in IPV.

Analysis of the role condition

The results of testing the effects of the role condition on the IPV prevalence and cheating are in Table 5. Each row of this table includes G^2, AIC and BIC values of two models and their differences between both models. The parameter representing the respective effect is estimated freely in the "free" model and restricted to 0 in the "restricted" model. None of the fit statistics indicate that excluding an interaction term (π_3 or γ_3) for participant gender and role condition leads to a relevant decrease in model fit. Restricting the main effect of role condition on the honest carrier IPV prevalence π_2 to equal 0 does not lead to a decrease in model fit. Only the AIC favors the model allowing π_2 to differ from 0, and only if the restriction is introduced before the restriction on the main effect on cheating. The effect size estimate is $\hat{\pi}_2$ = -0.50 on the logit scale, which means that the odds of reporting IPV are estimated to be $e^{-0.50} = 0.61$ times as high for participants queried about perpetration as compared to victimization (i.e., taking the inverse, 1.64 times as high for participants queried about victimization). Restricting the main effect of role condition on the cheating prevalence γ_2 to equal 0 leads to a decrease in model fit according to the G^2 test, if it is restricted before the main effect on the IPV prevalence is restricted. However, this effect is not significant if multiple testing is taken into account using Holmes-Bonferroni corrections on the p-values. The AIC favors the unrestricted model both if the restriction of the main effect

on cheating is introduced before and after the main effect on the honest carrier IPV prevalence is restricted. The effect size estimate is $\hat{\gamma}_2$ = -2.31 on the logit scale, which means that the odds of cheating are estimated to be $e^{-2.31} = 0.10$ times as high for participants queried about perpetration as compared to victimization (i.e., taking the inverse, 10.11 times as high for participants queried about victimization).

To summarize, we found (a) no interaction of role condition and participants' gender and (b) no significant main effect of role condition on the honest carrier prevalence or cheating. Specifically, contrary to our expectations, the prevalence of cheaters γ is not estimated to be higher in the group queried about perpetration than in the group queried about victimization. Additionally, numerically, the effect of role condition on γ, indicated by a small AIC difference, even goes in the opposite direction (i.e., γ is estimated to be higher in the group queried about victimization). Moreover, the effect of the role condition on the prevalence of honest carriers π_{UQMC} is not complementary to the effect on cheating. An effect of role condition on π_{UQMC} is only indicated by a small AIC difference and, numerically, the effect goes in the same direction as the effect on cheating (i.e., π_{UQMC} is estimated to be higher in the group queried about victimization).

From the effect size estimates, separate predictions for the UQMC parameters for both role conditions can be derived. For IPV victimization, the predicted honest carrier prevalence is $\pi_{vict} = 0.14$ and the predicted cheating prevalence is $\gamma_{vict} = 0.05$, both pooled across gender. For IPV perpetration, the predicted honest carrier prevalence is $\pi_{perp} = 0.09$ and the predicted cheating prevalence is $\gamma_{perp} = 0.01$, both pooled across gender.

Discussion

The current study was conducted to assess the validity of the cheating extension of the unrelated question model (UQMC; Reiber, Pope, and Ulrich 2020) in an applied setting. To that end, we investigated intimate partner violence (IPV) in a UQMC design in an online survey. We assessed the fit of the UQMC and compared it to the fit of the

UQM not accounting for cheating. Additionally, respondents were either queried about IPV victimization or perpetration because we expected this manipulation of question sensitivity to influence cheating. In light of the inconclusive prior findings on gender differences we either conducted the analyses separately for male and female respondents or included gender as a control variable.

The overall model fit of the UQMC is acceptable and it is superior to the fit of the UQM, which cannot account for cheating. The biggest advantage is observable in the subsample of female participants queried about victimization of IPV. In this group the prevalence estimate of cheating is 30 percent. Thus, especially in this group of respondents, accounting for cheating allows responses to be more accurately described.

However, the effects of the IPV role condition manipulation are not as expected. Contrary to our expectations, cheating is estimated to be higher in the subsamples queried about victimization. Also in the logistic model, the observed effect of the IPV role condition on cheating is not as expected and numerically even opposite to our expectations. Theoretically, this could mean that perpetrators are less reluctant to report their behavior than victims[9], but this is not in line with previous literature, which showed that reporting of perpetration is stigmatized (e.g., Birkel and Guzy 2015; Franke et al. 2004) and associated even stronger with social desirability than victimization (Sugarman and Hotaling 1997). Moreover, if perpetrators were open to report their behavior and victims were reluctant to do so, the honest carrier prevalence of perpetration should be higher than that of victimization. Specifically, because the sample only consists of persons in an exclusive relationship, the true prevalence of IPV victimization and perpetration should be the same and, therefore, differences in the honest carrier prevalence should result from complementary differences in cheating. However, the prevalence estimate of honest carriers is not lower but numerically even higher in the subsample queried about victimization. In other words, the manipulation of the IPV role did not affect the parameters in different directions, indicating that the model's parameters are not complementary. This is not in line with the reasoning behind these parameters.

To summarize, although the general model fit is good (and therefore the model's assumptions seem to hold), we were not able to differentially manipulate the model parameters. In the following, three possible explanations for this inconsistency are outlined.

First, the inconsistency might be due to selective sampling. The expectations concerning the parameter relationship between role conditions are based on the assumption that the true prevalence of IPV perpetration and victimization in the assessed sample is the same. Yet, this does not necessarily have to be true. For example, IPV perpetrators could have decided to abort the survey more often than victims of IPV once they realized the content of the question. This would mean that the honest carrier and cheating prevalence are not complementary and explain how both can be higher in the victimization condition. However, the general dropout rates are not high enough to completely explain the inverted data pattern (see Section A of the online supplemental materials). Especially the dropout rates on the screen on which the queried role became apparent are very low (victimization: $N = 9$; perpetration: $N = 13$). A higher dropout among perpetrators before this point, which is independent of the role condition, could only have such a large impact on parameter estimates if the true prevalence of IPV was much higher than estimated in either of the conditions. Therefore, selective dropout is an unlikely explanation for the unexpected finding of model parameters being non-complementary. However, selective participation could still explain the results pattern, if IPV perpetrators were generally less likely to participate in surveys or be part of respondiAG's participant panel.

Second, there could be violations of the model assumptions which are mathematically consistent with the UQMC and thus not detectable merely by computational tests of model fit. For example, the UQMC inherited the assumption from the CDM that cheating is equally likely among respondents instructed to respond to the sensitive question and respondents instructed to respond to the neutral question. However, this need not be the case. Therefore, in the original presentation of the UQMC

(Reiber, Pope, and Ulrich 2020), the possibility of *partial cheating* was outlined. In this framework, in addition to the two categories of respondents defined in the UQMC, that is, honest respondents and cheaters, there is a third category termed partial cheaters. This group of respondents would respond honestly if directed to answer the neutral question but give a self-protecting "No"-response if directed to the sensitive question. Interestingly, following this logic, the estimation of the model parameters does not change. Specifically, the prevalence of cheating γ and the prevalence of honest carriers π_{UQMC} are estimated like in the UQMC that only allows for complete cheaters. The only thing that changes is the interpretation of the remainder category. In the UQMC, like in the CDM, the remainder category, $1 - \pi_{UQMC} - \gamma$, is interpreted as the prevalence of honest non-carriers. However, in the framework of partial cheating, this remainder category also entails the partial cheaters.[10] In light of this idea, the results of the study could be interpreted differently: There could be partial cheaters in the subsample queried about perpetration, who cannot be detected by the model but their presence would explain the unexpected differences between estimates in the perpetration and victimization conditions.

Third, following a more substantive line of reasoning, differences in the individual interpretations of IPV by the participants could account for the data pattern. The UQMC is only capable of detecting deliberate cheating. Therefore, the hypotheses depend on the assumption that not only the true prevalence of IPV victimization and perpetration is equal, but also the perceived prevalence. However, it has been proposed that perpetrators and victims judge the same instance of IPV differently (see, Follingstad and Rogers 2013). Specifically, the same situation can be reported as violent by the suspected victim but not by the suspected perpetrator. In such a case, a perpetrator not admitting to a violent act, which was perceived as violent by the victim, would not be a cheater in the sense of the UQMC. We decided to assess only physical IPV and provided specific examples in the instructions to minimize the likelihood of self-deception. However, it might still have played a role. This would explain why the lower estimated perpetration prevalence in the current study is not explainable by higher cheating.

Apart from these accounts there are limitations of the present study which might have influenced the results. On the one hand, it was crucial for the premises of our experimental manipulation that the participants were in a relationship with exactly one person. However, the relationship status in itself is a sensitive topic since in most social groups being in a committed relationship with one person still constitutes the norm. By only contrasting "being in a committed relationship with one person" to "not being in a relationship" or "being in more than one relationship of equal importance" in the respective screening question, we tried to minimize social desirability bias. However, it is still possible that some respondents chose to respond that they were in a committed relationship with one person although they were not. Nevertheless, this would only influence the results pattern if the likelihood of this response tendency differed strongly between perpetrators and victims of IPV.

On the other hand, there was a high proportion of respondents (27.60 percent) who did not respond correctly to the second training question. This calls into doubt that the instructions were sufficiently understood. Given that the probability to guess the correct response is 50 percent, this would in the worst case mean that another 30 percent did not fully understand the instructions. However, this seems unlikely because the rate of incorrect responses to the first training training question was much lower (18.03 percent). Instead, since respondents did not know that an incorrect response to the second training question would lead to an exclusion of their data, they might not have payed attention to this question after correctly answering the first one. Therefore, the high proportion of incorrect responses could be not as much indicative of a major problem with understanding the instructions but rather that this preregistered exclusion criterion was sub-optimal. Still, this exclusion criterion did not substantively influence the results pattern either, as indicated by the additional analyses in Section C of the online supplemental materials.

Whether any of these accounts is actually responsible for the observed inconsistencies in the data pattern is, of course, not testable using the given data.

However, the applied design enabled us to detect these inconsistencies and come up with plausible explanations. Surveys using direct questions or a simple RRT design are probably also affected by unexpected response patterns. In these cases, however, the inconsistencies do not become visible. Using the design applied in this study, we could, first, measure a specific type of instruction non-adherence, namely cheating, and the results indicate that especially among female participants queried about IPV victimization cheating is highly prevalent. Second, the unexpected effects of experimentally manipulating the queried IPV role indicated that additional factors influence the estimates. Although we can only speculate about these factors, detecting inconsistencies itself has important implications. It shows that the estimates need to be treated with caution - something that is arguably true for any survey on IPV.

All of the outlined explanations suggest that the IPV prevalence estimates in this study rather represent a lower limit to the true prevalence of IPV during the period of about three months starting with the initiation of the first contact restrictions due to the COVID-19 pandemic in Germany. However, even the lower limit estimates of about 10 percent are already very high for such a short time period. Therefore, although the exact numbers need to be interpreted carefully and, of course, a direct comparison to other time periods is not possible, the presented results are in line with the literature reporting alarmingly high numbers of IPV in the context of the COVID-19 pandemic and the related containment measures (e.g., Steinert and Ebert 2020).

Conclusion

The purpose of the current study was to validate the UQMC, an extension of the UQM, to account for self-protecting responses. To that end, we conducted an online survey on IPV during the first contact restrictions due to the COVID-19 pandemic in Germany. The UQMC provides a reasonable account of the data, which is superior to that of the UQM. The data indicate an alarmingly high prevalence of IPV, which is in line with the increase in IPV related to the COVID-19 pandemic reported by many other sources. Some unexpected data patterns emerged, highlighting once more the difficulty of

investigating sensitive research topics and the need for treating the respective estimates with caution. Nevertheless, testable RRT designs accounting for instruction non-adherence can provide more insight into the response process and, thereby, a better understanding of sensitive research topics.

Footnotes

[1]Note that the interpretation of π_{CDM} differs from that of π_{UQM} as it denotes the combined probability of being an honest respondent and a carrier of the sensitive attribute.

[2]The assumption that a higher probability to receive the sensitive question would lead to a lower proportion of honest admissions was tested by experimentally manipulating p in a UQM survey (Dietz et al. 2018). A difference in the expected direction was observed but it was not significant, possibly due to a lack of power.

[3]This would require estimating a model with separate cheating parameters for each subsample. Such a model, however, is underdetermined.

[4]Closed form equations for a UQMC implementation using only two subsamples are provided in Reiber, Pope, and Ulrich (2020). However, this approach does not allow for the assessment of model fit.

[5]Adapted from Moshagen, Musch, and Erdfelder (2012) and translated from German.

[6]Adapted from Moshagen, Musch, and Erdfelder (2012) and translated from German.

[7]We used R (Version 4.0.5; R Core Team 2021) and the R-packages dplyr (Version 1.0.5; Wickham et al. 2021), forcats (Version 0.5.0; Wickham 2020), ggplot2 (Version 3.3.2; Wickham 2016), kableExtra (Version 1.1.0; Zhu 2019), papaja (Version 0.1.0.9997; Aust and Barth 2020), purrr (Version 0.3.4; Henry and Wickham 2020), readr (Version 1.3.1; Wickham, Hester, and Francois 2018), stringr (Version 1.4.0; Wickham 2019), tibble (Version 3.1.0; Müller and Wickham 2021), tidyr (Version 1.1.3; Wickham 2021), and tidyverse (Version 1.3.0; Wickham et al. 2019) for all our analyses.

[8]To facilitate estimation we used the estimates from simpler models as starting values for the more complex models.

[9]A reviewer suggested they could even be boastful instead of ashamed about their controlling behavior.

[10]For an outline of the logic behind this conclusion see Reiber, Pope, and Ulrich (2020).

References

Archer, John. 2000. "Sex Differences in Aggression Between Heterosexual Partners: A Meta-Analytic Review." *Psychological Bulletin* 126:651–80. doi: 10.1037/0033-2909.126.5.651.

Aust, Frederik, and Marius Barth. 2020. *papaja: Create APA Manuscripts with R Markdown*.

Birkel, Christoph, and Nathalie Guzy. 2015. *Viktimisierungsbefragungen in Deutschland*. 47.1 ed. edited by R. Mischkowitz. Wiesbaden: Bundeskriminalamt.

Boruch, Robert F. 1971. "Assuring Confidentiality of Responses in Social Research: A Note on Strategies." *The American Sociologist* 6:308–11.

Böckenholt, Ulf, and Peter G. M. Van Der Heijden. 2007. "Item Randomized-Response Models for Measuring Noncompliance: Risk-return Perceptions, Social Influences, and Self-Protective Responses." *Psychometrica* 72:245–62. doi: 10.1007/s11336-005-1495-y.

Bradbury-Jones, Caroline, and Louise Isham. 2020. "The Pandemic Paradox: The Consequences of COVID-19 on Domestic Violence." *Journal of Clinical Nursing* 29:2047–49. doi: 10.1111/jocn.15296.

Bundeskriminalamt. 2020. *Partnerschaftsgewalt Kriminalstatistische Auswertung – Berichtsjahr 2019*.

Clark, S. J., and R. A. Desharnais. 1998. "Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model." *Psychological Methods* 3:160–68.

Devries, K. M., J. Y. T. Mak, C. García-Moreno, M. Petzold, J. C. Child, G. Falder, S. Lim, L. J. Bacchus, R. E. Engell, L. Rosenfeld, C. Pallitto, T. Vos,

N. Abrahams, and C. H. Watts. 2013. “The Global Prevalence of Intimate Partner Violence Against Women.” *Science* 340:1527–28. doi: 10.1126/science.1240937.

Dietz, P., A. Quermann, M. N. Maria van Poppel, H. Striegel, H. Schröter, R. Ulrich, and P. Simon. 2018. “Physical and Cognitive Doping in University Students Using the Unrelated Question Model (UQM): Assessing the Influence of the Probability of Receiving the Sensitive Question on Prevalence Estimation.” *PLoS ONE* 13. doi: 10.1371/journal.pone.0197270.

Elbe, Anne-Marie, and Werner Pitsch. 2018. “Doping Prevalence Among Danish Elite Athletes.” *Performance Enhancement and Health* 6:28–32. doi: 10.1016/j.peh.2018.01.001.

Ellsberg, Mary, Lori Heise, Rodolfo Peña, Sonia Agurto, and Anna Winkvist. 2001. “Researching Domestic Violence Against Women: Methodological and Ethical Considerations.” *Studies in Family Planning* 32:1–16.

Emundts, Corinna. 2020. “Gewalt in Beziehungen: "Die Zahlen sind schockierend".”

European Agency for Fundamental Rights. 2014. *Violence against women: an EU-wide survey.*

Follingstad, Diane R., and M. Jill Rogers. 2013. “Validity Concerns in the Measurement of Women’s and Men’s Report of Intimate Partner Violence.” *Sex Roles* 69:149–67. doi: 10.1007/s11199-013-0264-5.

Franke, B., D. Seifert, S. Anders, J. Schröer, and A. Heinemann. 2004. “Gewaltforschung zum Thema "häusliche Gewalt" aus kriminologischer Sicht.” *Rechtsmedizin* 14:193–98. doi: 10.1007/s00194-004-0263-5.

Garcia-Moreno, Claudia, Henrica A. F. M. Jansen, Mary Ellsberg, Lori Heise, and Charlotte H. Watts. 2006. "Prevalence of Intimate Partner Violence: Findings from the WHO Multi-Country Study on Women's Health and Domestic Violence." *The Lancet* 368:1260–69. doi: 10.1016/S0140-6736(06)69523-8.

Gracia, Enrique. 2004. "Unreported Cases of Domestic Violence Against Women: Towards an Epidemiology of Social Silence, Tolerance, and Inhibition." *Journal of Epidemiology and Community Health* 58:536–37. doi: 10.1136/jech.2003.019604.

Greenberg, Bernard G., Abdel-Latif A. Abul-Ela, Walt R. Simmons, and Daniel G. Horvitz. 1969. "The Unrelated Question Randomized Response Model: Theoretical Framework." *Journal of the American Statistical Association* 64:520–39.

Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*.

Hoffmann, Adrian, Birk Diedenhofen, Bruno Verschuere, and Jochen Musch. 2015. "A Strong Validation of the Crosswise Model Using Experimentally-Induced Cheating Behavior." *Experimental Psychology* 62:403–14. doi: 10.1027/1618-3169/a000304.

Hoffmann, Adrian, Berenike Waubert De Puiseau, Alexander F. Schmidt, and Jochen Musch. 2017. "On the Comprehensibility and Perceived Privacy Protection of Indirect Questioning Techniques." *Behavior Research Methods* 49:1470–83. doi: 10.3758/s13428-016-0804-3.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Measuring Voter Turnout by Using the Randomized Response Technique - Evidence Calling into Question the Methods´s Validity." *Public Opinion Quarterly* 74:328–43.

Horvitz, D. G., B. G. Greenberg, and J. R. Abernathy. 1976. "Randomized Response: A Data-Gathering Device for Sensitive Questions." *International*

Statistical Review 44:181–96.

Höglinger, Marc, and Andreas Diekmann. 2017. "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT." *Political Analysis* 25:131–37. doi: 10.1017/pan.2016.5.

Höglinger, Marc, and Ben Jann. 2018. "More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model." *PLOS ONE* 13:e0201770. doi: 10.1371/journal.pone.0201770.

Höglinger, Marc, Ben Jann, and Andreas Diekmann. 2016. "Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model." *Survey Research Methods* 10:171–87. doi: 10.18148/srm/2016.v10i3.6703.

Jarnecke, Amber M., and Julianne C. Flanagan. 2020. "Staying Safe During COVID-19: How a Pandemic Can Escalate Risk for Intimate Partner Violence and What Can Be Done to Provide Individuals with Resources and Support." *Psychological Trauma: Theory, Research, Practice, and Policy* 12:202–4. doi: 10.1037/tra0000688.

Johnson, Michael P. 2006. "Conflict and Control: Symmetry and Asymmetry in Domestic Violence." *Violence Against Women* 1003–18. doi: 10.1177/1077801206293328.

Kimmel, Michael S. 2002. ""Gender Symmetry" in Domestic Violence: A Substantive and Methodological Research Review." *Violence Against Woman* 8:1332–63. doi: 10.1177/107780102237407.

Leiner, D. J. 2020. "SoSci Survey (Version 3.2.12)."

Leiner, Dominik J. 2019. "Too Fast, Too Straight, Too Weird: Non-reactive Indicators for Meaningless Data in Internet Surveys." *Survey Research Methods* 13:229–48. doi: 10.18148/srm/2019.v13i3.7403.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. Van Der Heijden, and Cora J. M. Maas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-five Years of Validation." *Sociological Methods and Research* 33:319–48. doi: 10.1177/0049124104268664.

Meisters, Julia, Adrian Hoffmann, and Jochen Musch. 2020. "Can Detailed Instructions and Comprehension Checks Increase the Validity of Crosswise Model Estimates?" *PLoS ONE* 15:1–19. doi: 10.1371/journal.pone.0235403.

Moshagen, Morten, Jochen Musch, and Edgar Erdfelder. 2012. "A Stochastic Lie Detector." *Behavior Research Methods* 44:222–31. doi: 10.3758/s13428-011-0144-2.

Moshagen, Morten, Jochen Musch, Martin Ostapczuk, and Zengmei Zhao. 2010. "Reducing Socially Desirable Responses in Epidemiologic Surveys: An Extension of the Randomized-Response Technique." *Epidemiology* 21:379–82. doi: 10.1097/EDE.0b013e3181d61dbc.

Müller, Kirill, and Hadley Wickham. 2021. *Tibble: Simple Data Frames*.

Nelder, J. A., and R. Mead. 1965. "A Simplex Method for Function Minimization." *The Computer Journal* 7:308–13. doi: 10.1093/comjnl/7.4.308.

Ostapczuk, Martin. 2011. "Improving Self-Report Measures of Medication Non-Adherence Using a Cheating Detection Extension of the Randomised-Response-Technique." *Statistical Methods in Medical Research* 20:489–503. doi: 10.1177/0962280210372843.

Pitsch, Werner, Eike Emrich, and Markus Klein. 2007. "Doping in Elite Sports in Germany: Results of a Www Survey." *European Journal for Sport and Society* 4:89–102. doi: 10.1080/16138171.2007.11687797.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reiber, Fabiola, Harrison Pope, and Rolf Ulrich. 2020. "Cheater Detection Using the Unrelated Question Model." *Sociological Methods and Research*. doi: 10.1177/0049124120914919.

Schröter, Hannes, Beatrix Studzinski, Pavel Dietz, Rolf Ulrich, Heiko Striegel, and Perikles Simon. 2016. "A Comparison of the Cheater Detection and the Unrelated Question Models: A Randomized Response Survey on Physical and Cognitive Doping in Recreational Triathletes." *PLoS One* 11:e0155765. doi: 10.1371/journal.pone.0155765.

Smith, S. G., X. Zhang, K. C. Basile, M. T. Merrick, J. Wang, M. Kresnow, and J. Chen. 2018. *National Intimate Partner and Sexual Violence Survey (NISVS): 2015 Data Brief-Update Release*.

Statistisches Bundesamt. 2020. "Statistik der Geburten - Lebendgeborene: Deutschland, Monate, Geschlecht."

Steinert, Janina, and Cara Ebert. 2020. *Gewalt an Frauen und Kindern in Deutschland während COVID-19-bedingten Ausgangsbeschränkungen: Zusammenfassung der Ergebnisse*.

Sugarman, David B., and Gerald T. Hotaling. 1997. "Intimate Violence and Social Desirability: A Meta-Analytic Review." *Journal of Interpersonal Violence* 12:275–90.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859–83. doi: 10.1037/0033-2909.133.5.859.

Ulrich, Rolf, Harrison G. Pope, Léa Cléret, Andrea Petróczi, Tamás Nepusz, Jay Schaffer, Gen Kanayama, R. Dawn Comstock, and Perikles Simon. 2018. "Doping in Two Elite Athletics Competitions Assessed by Randomized-Response Surveys." *Sports Medicine* 48:211–19. doi: 10.1007/s40279-017-0765-4.

Usher, Kim, Navjot Bhullar, Joanne Durkin, Naomi Gyamfi, and Debra Jackson. 2020. "Family Violence and COVID-19: Increased Vulnerability and Reduced Options for Support." *International Journal of Mental Health Nursing* 29:549–52. doi: 10.1111/inm.12735.

Van Der Heijden, Peter G. M., Ger van Gils, Jan Bouts, and Joop J. Hox. 2000. "A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning." *Sociological Methods & Research* 28:505–37. doi: 10.1177/0049124100028004005.

Waltermaurer, Eve. 2005. "Measuring Intimate Partner Violence (IPV) You May Only Get What You Ask For." *Journal of Interpersonal Violence* 20:501–6. doi: 10.1177/0886260504267760.

Warner, S. L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60:63–66.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, Hadley. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*.

Wickham, Hadley. 2020. *Forcats: Tools for Working with Categorical Variables (Factors)*.

Wickham, Hadley. 2021. *Tidyr: Tidy Messy Data*.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4(43):1686. doi: 10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*.

Wickham, Hadley, Jim Hester, and Romain Francois. 2018. *Readr: Read Rectangular Text Data*.

Wimbush, James C., and Dan R. Dalton. 1997. "Base Rate for Employee Theft: Convergence of Multiple Methods." *Journal of Applied Psychology* 82:756–63. doi: 10.1037//0021-9010.82.5.756.

World Health Organization. 2012. *Intimate Partner Violence*.

Zhu, Hao. 2019. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.

Table 1

Sample demographics

		Sample size (percentage)		
	Target quota in percent	All	Victimization	Perpetration
Gender				
Male	49.60	1553 (47.13)	752 (46.48)	801 (47.76)
Female	50.40	1732 (52.56)	862 (53.28)	870 (51.88)
Diverse	-	10 (0.30)	4 (0.25)	6 (0.36)
Age				
[18,30)	18.50	640 (19.42)	328 (20.27)	312 (18.60)
[30,40)	16.50	510 (15.48)	251 (15.51)	259 (15.44)
[40,50)	17.00	558 (16.93)	280 (17.31)	278 (16.58)
[50,60)	21.20	689 (20.91)	342 (21.14)	347 (20.69)
[60,80)	26.80	898 (27.25)	417 (25.77)	481 (28.68)
Highest educational achievement*				
Less than primary school	-	12 (0.36)	6 (0.37)	6 (0.36)
Primary/lower secondary education	-	485 (14.72)	220 (13.60)	265 (15.80)
→ low	32.90	497 (15.08)	226 (13.97)	271 (16.16)
Middle secondary education	-	1043 (31.65)	512 (31.64)	531 (31.66)
→ medium	32.40	1043 (31.65)	512 (31.64)	531 (31.66)
High secondary education	-	269 (8.16)	141 (8.71)	128 (7.63)
Apprenticeship	-	868 (26.34)	415 (25.65)	453 (27.01)
University degree	-	618 (18.76)	324 (20.02)	294 (17.53)
→ high	34.70	1755 (53.26)	880 (54.39)	875 (52.18)

Note. Displayed are the target quotas and acquired sub sample sizes (percentages in parantheses) for all participants, and those in the victimization and perpetration conditions separately, for gender, age and highest educational achievement.

*Target quotas for highest educational achievement referred to the summary categories “low”, “medium”, and “high”. The percentages in the raw categories and the summary categories each sum up to 100 percent. The raw categories of highest educational achievement are translated from the German categories “Kein Schulabschluss”, “Grund-/Hauptschulabschluss”, “Realschule (Mittlere Reife)”, “Gymnasium (Abitur)”, “Abgeschlossene Ausbildung”, and “(Fach-)Hochschulabschluss” in this order.

Table 2

Condition allocation

	Victimization		Perpetration	
$_p{}^q$	0.25	0.75	0.25	0.75
1/3	420	407	405	425
2/3	402	389	426	421

Note. Number of participants per combination of the factors role, p and q.

Table 3

Model Estimates

	N	$\hat{\pi}$ (SE)	$\hat{\gamma}$ (SE)	$\hat{\pi}+\hat{\gamma}$ (SE)
Victimization Male				
UQMC	752	.14 (.03)	.08 (.06)	.23 (.08)
UQM	752	.12 (.02)	-	-
Victimization Female				
UQMC	862	.18 (.03)	.30 (.06)	.48 (.08)
UQM	862	.07 (.02)	-	-
Perpetration Male				
UQMC	801	.11 (.03)	.05 (.05)	.16 (.08)
UQM	801	.09 (.02)	-	-
Perpetration Female				
UQMC	870	.08 (.03)	.00 (.06)	.08 (.08)
UQM	870	.08 (.02)	-	-

Note. Estimates for the prevalence of IPV (prevalence of honest carriers) π in the UQM (UQMC), the prevalence of cheating γ and the upper bound of the prevalence of IPV $\pi+\gamma$ in the UQMC along with their estimated standard errors in parentheses.

Table 4

Comparison of the UQM and the UQMC

		Model Fit			Model Comparison						
	N	G^2	df	p	ΔG^2	df	p	AIC	ΔAIC	BIC	ΔBIC
Victimization Male											
UQMC	752	5.61	2	.061				9.61		18.85	
UQM	752	6.61	3	.086	1.00	1	.317	8.61	-1.00	13.23	-5.62
Victimization Female											
UQMC	862	3.90	2	.142				7.90		17.42	
UQM	862	19.10	3	$< .001$	15.19	1	$< .001$	21.10	13.19	25.86	8.44
Perpetration Male											
UQMC	801	0.76	2	.683				4.76		14.13	
UQM	801	1.24	3	.745	0.47	1	.492	3.24	-1.53	7.92	-6.21
Perpetration Female											
UQMC	870	0.37	2	.830				4.37		13.91	
UQM	870	0.37	3	.946	0.00	1	$> .999$	2.37	-2.00	7.14	-6.77
Overall											
UQMC	3285	10.64	8	.223							

Table 4 continued

		Model Fit			Model Comparison						
	N	G^2	df	p	ΔG^2	df	p	AIC	ΔAIC	BIC	ΔBIC
UQM	3285	27.31	12	.007	16.67	4	.002				

Note. Model fit and model comparison of UQMC and the UQM using G^2-tests and the Akaike and Bayesian Information Criterion (AIC and BIC).

Table 5

Tests for differences between the role conditions

	G^2						AIC			BIC		
	free	restr.	Δ	*df*	*p*	p_{corr}	free	restr.	Δ	free	restr.	Δ
Role × Gender interaction effects												
π_3 (restricted first)	10.68	11.58	0.89	1	.345	> .999	26.68	25.58	-1.11	75.46	68.26	-7.21
γ_3 (restricted after π_3)	11.58	13.13	1.55	1	.213	.852	25.58	25.13	-0.45	68.26	61.71	-6.55
γ_3 (restricted first)	10.68	11.13	0.44	1	.505	> .999	26.68	25.13	-1.56	75.46	67.81	-7.65
π_3 (restricted after γ_3)	11.13	13.13	2.00	1	.158	.788	25.13	25.13	0.00	67.81	61.71	-6.10
Role effects												
π_2 (restricted first)	13.13	15.71	2.59	1	.108	.647	25.13	25.71	0.59	61.71	56.20	-5.51
γ_2 (restricted after π_2)	15.71	19.54	3.83	1	.050	.352	25.71	27.54	1.83	56.20	51.93	-4.27
γ_2 (restricted first)	13.13	19.47	6.34	1	.012	.094	25.13	29.47	4.34	61.71	59.95	-1.76
π_2 (restricted after γ_2)	19.47	19.54	0.08	1	.784	.784	29.47	27.54	-1.92	59.95	51.93	-8.02

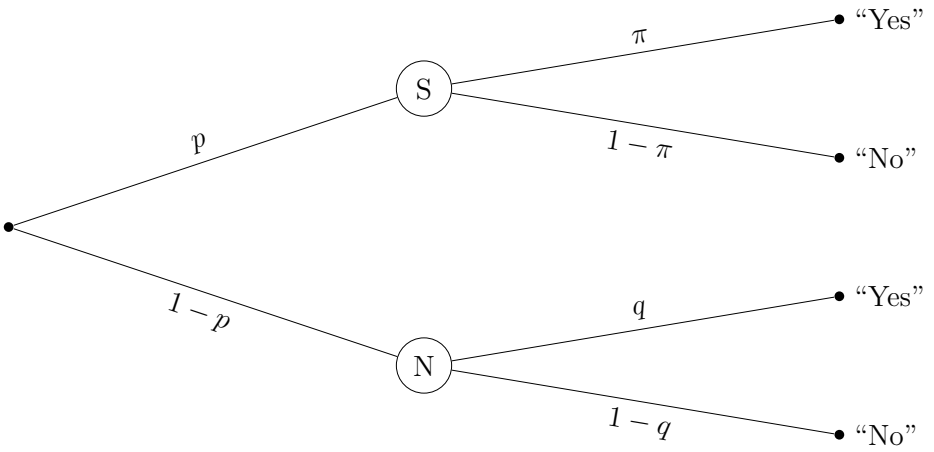
Table 5 continued

	G^2						AIC			BIC		
	free	restr.	Δ	*df*	*p*	p_{corr}	free	restr.	Δ	free	restr.	Δ

Note. N = 3285. The interaction effects of Role and Gender and the main effects of Role on cheating γ and the honest carrier prevalence π are tested using using G^2-difference tests and Akaike and Bayesian Information Criterion (AIC and BIC) differences. By row, single parameters are successively restricted to 0. Each resulting restricted ("restr.") model is compared to a more complex model, in which the respective parameter is estimated freely ("free"). p_{corr} refers to p-values adjusted for multiple testing using the Holmes-Bonferroni correction. The models for testing the interaction effects are derived from the full logistic model: $logit(\gamma) = \gamma_0 + \gamma_1 \cdot Gender + \gamma_2 \cdot Role + \gamma_3 \cdot Gender \cdot Role$; $logit(\pi) = \pi_0 + \pi_1 \cdot Gender + \pi_2 \cdot Role + \pi_3 \cdot Gender \cdot Role$. The models for testing the Role effects are derived from the main effects model: $logit(\gamma) = \gamma_0 + \gamma_1 \cdot Gender + \gamma_2 \cdot Role$; $logit(\pi) = \pi_0 + \pi_1 \cdot Gender + \pi_2 \cdot Role$

Figure 1

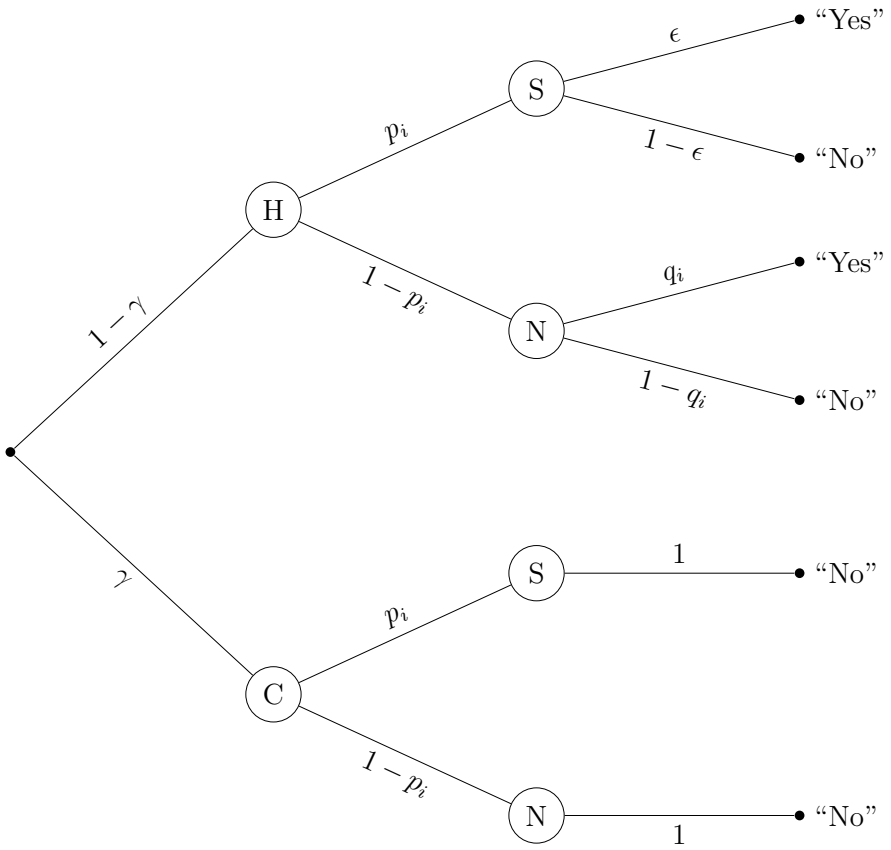
Probability tree of the UQM



Note. The sensitive question S and the neutral question N are randomly received by respondents with probability p and $1 - p$, respectively. The probabilities of responding "Yes" and "No" to the neutral question N are q and $1 - q$ and the probabilities of responding "Yes" and "No" to the sensitive question S are π and $1 - \pi$. Adapted from "Cheater detection using the unrelated question model" by F. Reiber, H. Pope, and R. Ulrich, 2020, *Sociological Methods and Research*, advance online publication, p. 3, https://doi.org/10.1177/0049124120914919 published by SAGE Publishing under the terms of Creative Commons Attribution 4.0.

Figure 2

Probability tree of the UQMC



Note. The prevalence of cheaters C is γ and the prevalence of honest participants H is $1 - \gamma$. In both cases, the sensitive question S and the neutral question N are received by participants with probability p_i and $1 - p_i$, respectively. The model assumes that cheaters always say “No” regardless of the question received. Honest participants respond “Yes” with probability q_i and “No” with probability $1 - q_i$ if instructed to answer the neutral question N. They answer “Yes” with probability ϵ and “No” with probability $1 - \epsilon$, if instructed to answer the sensitive question S. Thus, there are three groups of participants: (a) honest participants who are carriers of the sensitive attribute, who will respond “Yes” with probability $(1 - \gamma) \cdot \epsilon = \pi$ if they receive S; (b) honest non-carriers of this attribute who will respond “No” with probability $(1 - \gamma) \cdot (1 - \epsilon)$ if they receive S; and (c) cheaters, who will respond “No” with probability γ regardless of whether they receive S or N. Adapted from “Cheater detection using the unrelated question model” by F. Reiber, H. Pope, and R. Ulrich, 2020, *Sociological Methods and Research*, advance online publication, p. 8, https://doi.org/10.1177/0049124120914919 published by SAGE Publishing under the terms of Creative Commons Attribution 4.0.

Supplemental Section A

Participant dropout

Figure A1 depicts the dropout of participants per screen of the questionnaire and the median processing time of each screen. Only the data of participants who completed the whole survey were included in the data analysis.

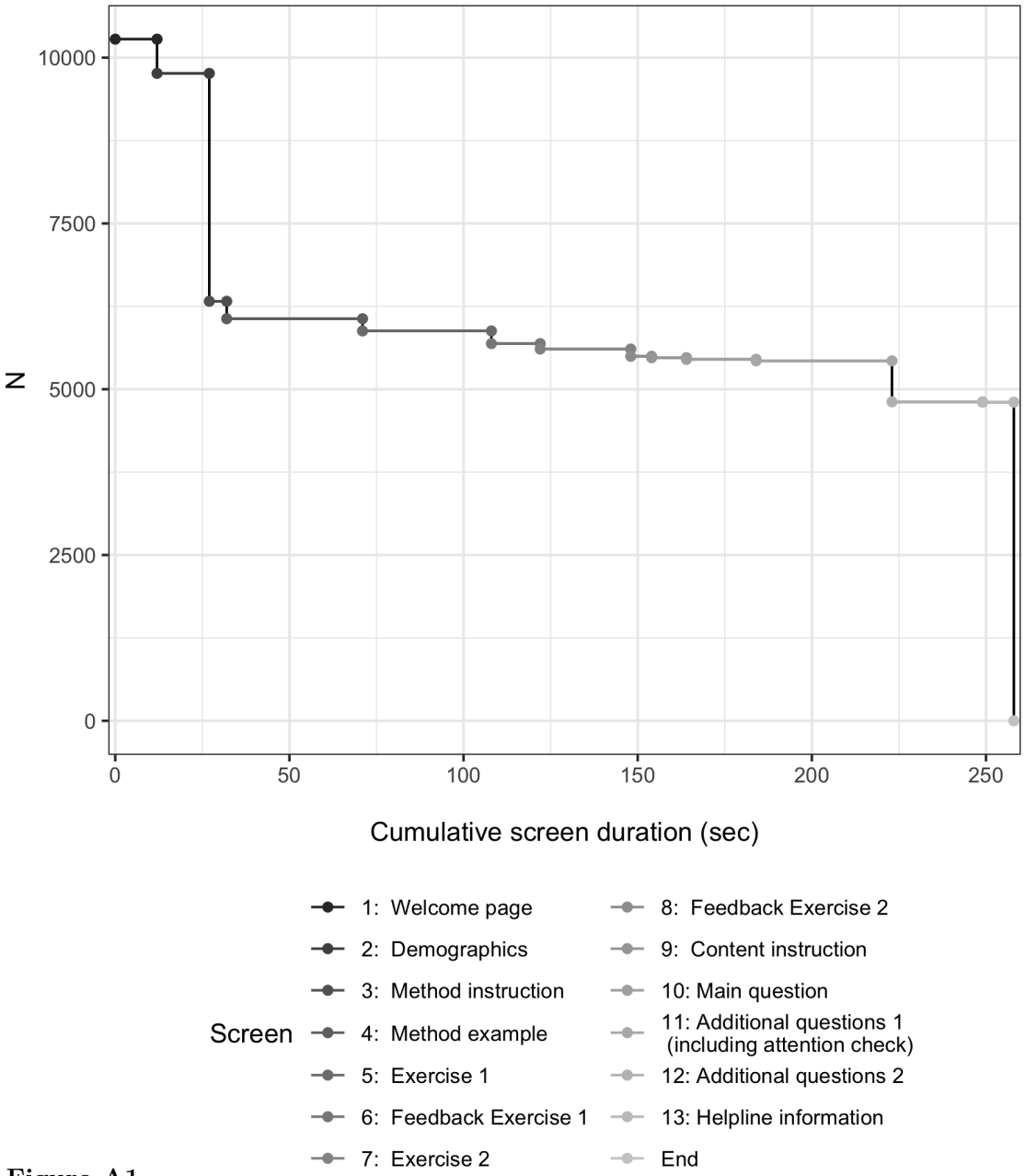


Figure A1

Participant dropout per screen

Supplemental Section B

Additional questions

1. How many rooms does your flat/house have (excluding kitchen and bathroom)?
2. How many persons live in the household in addition to you? (Total/Children)
3. Do you live together with your partner? (Yes / No)
4. Which city is not located in Germany? (Berlin, Hamburg, Cologne, London, Frankfurt, Munich)
5. Mark all alternatives that apply to you:
    - I am employed (incl. self-employed).
    - I mainly work from home.
    - I am in *Kurzarbeit* (short-time work, furlough scheme: forced reduction of working time to avoid bankruptcies or layoffs; part of the salary is compensated by the state)
    - I am in full-time education.
6. Do or did you feel threatened by unemployment? (Yes / No / I was unemployed before March 23rd / I became unemployed after March 23rd)
7. Did you have regular personal contact with persons (e.g., friends, family) not living in your household? (Yes - more than twice a week / Yes - at least once a week / Yes - less than once a week / No)
8. Did you adhere to the contract restrictions that apply to you? (Yes / Mostly / No)
9. How anxious do you currently feel? (1 - not anxious at all / 2 - a bit anxious / 3 - intermediate anxious / 4 - very anxious / 5 - extremely anxious)

Supplemental Section C

Analyses including participants who did not respond correctly to the second traing question

Of the sample including those participants who did not respond correctly to the second training question, 2212 (49.35%) answered the question on victimization of IPV. They did not differ from those who answered the question on perpetration with respect to age, $t(4480) = 0.98$, $p = .325$, gender, $p = .875$, Fisher's exact test, or highest educational achievement, $\chi^2(5) = 5.03$, $p = .413$.

Table C1

Sample demographics - Including participants incorrectly responding to the second training question

		Sample size (percentage)		
	Target quota in percent	All	Victimization	Perpetration
Gender				
Male	49.60	2173 (48.48)	1069 (48.33)	1104 (48.63)
Female	50.40	2297 (51.25)	1138 (51.45)	1159 (51.06)
Diverse	-	12 (0.27)	5 (0.23)	7 (0.31)
Age				
[18,30)	18.50	781 (17.43)	395 (17.86)	386 (17.00)
[30,40)	16.50	715 (15.95)	355 (16.05)	360 (15.86)
[40,50)	17.00	743 (16.58)	370 (16.73)	373 (16.43)
[50,60)	21.20	974 (21.73)	478 (21.61)	496 (21.85)
[60,80)	26.80	1269 (28.31)	614 (27.76)	655 (28.85)
Highest educational achievement*				
Less than primary school	-	16 (0.36)	9 (0.41)	7 (0.31)
Primary/lower secondary education	-	731 (16.31)	350 (15.82)	381 (16.78)
→ low	32.90	747 (16.67)	359 (16.23)	388 (17.09)
Middle secondary education	-	1468 (32.75)	731 (33.05)	737 (32.47)
→ medium	32.40	1468 (32.75)	731 (33.05)	737 (32.47)
High secondary education	-	329 (7.34)	173 (7.82)	156 (6.87)
Apprenticeship	-	1164 (25.97)	553 (25.00)	611 (26.92)
University degree	-	774 (17.27)	396 (17.90)	378 (16.65)
→ high	34.70	2267 (50.58)	1122 (50.72)	1145 (50.44)

Note. Displayed are the target quotas and acquired sub sample sizes (percentages in parantheses) for all participants, and those in the victimization and perpetration conditions separately, for gender, age and highest educational achievement.

*Target quotas for highest educational achievement referred to the summary categories “low”, “medium”, and “high”. The percentages in the raw categories and the summary categories each sum up to 100 percent. The raw categories of highest educational achievement are translated from the German categories “Kein Schulabschluss”, “Grund-/Hauptschulabschluss”, “Realschule (Mittlere Reife)”, “Gymnasium (Abitur)”, “Abgeschlossene Ausbildung”, and “(Fach-)Hochschulabschluss” in this order.

Table C2

Condition allocation - Including participants incorrectly responding to the second training question

	Victimization		Perpetration	
$_p{}^q$	0.25	0.75	0.25	0.75
1/3	555	545	557	590
2/3	580	532	545	578

Note. Number of participants per combination of the factors role, p and q.

Table C3

Model Estimates - Including participants incorrectly responding to the second training question

	N	$\hat{\pi}$ (SE)	$\hat{\gamma}$ (SE)	$\hat{\pi}+\hat{\gamma}$ (SE)
Victimization Male				
UQMC	1069	.23 (.03)	.11 (.05)	.34 (.07)
UQM	1069	.19 (.02)	-	-
Victimization Female				
UQMC	1138	.21 (.02)	.29 (.05)	.50 (.07)
UQM	1138	.11 (.02)	-	-
Perpetration Male				
UQMC	1104	.11 (.02)	.03 (.05)	.14 (.06)
UQM	1104	.09 (.02)	-	-
Perpetration Female				
UQMC	1159	.11 (.02)	.01 (.05)	.12 (.07)
UQM	1159	.11 (.02)	-	-

Note. Estimates for the prevalence of IPV (prevalence of honest carriers) π in the UQM (UQMC), the prevalence of cheating γ and the upper bound of the prevalence of IPV $\pi+\gamma$ in the UQMC along with their standard errors in parentheses.

Table C4

Comparison of the UQM and the UQMC - Including participants incorrectly responding to the second training question

		Model Fit			Model Comparison						
	N	G^2	df	p	ΔG^2	df	p	AIC	ΔAIC	BIC	ΔBIC
Victimization Male											
UQMC	1069	13.54	2	.001				17.54		27.49	
UQM	1069	15.99	3	.001	2.45	1	.117	17.99	0.45	22.97	-4.52
Victimization Female											
UQMC	1138	5.97	2	.051				9.97		20.04	
UQM	1138	24.01	3	$< .001$	18.04	1	$< .001$	26.01	16.04	31.05	11.01
Perpetration Male											
UQMC	1104	4.00	2	.135				8.00		18.01	
UQM	1104	4.25	3	.236	0.25	1	.617	6.25	-1.75	11.26	-6.76
Perpetration Female											
UQMC	1159	2.10	2	.350				6.10		16.21	
UQM	1159	2.12	3	.548	0.02	1	.886	4.12	-1.98	9.18	-7.03
Overall											

Table C4 continued

		Model Fit			Model Comparison						
	N	G^2	df	p	ΔG^2	df	p	AIC	ΔAIC	BIC	ΔBIC
UQMC	4470	25.60	8	.001							
UQM	4470	46.37	12	$< .001$	20.77	4	$< .001$				

Note. Model fit and model comparison using G^2-tests and the Akaike and Bayesian Information Criterion (AIC and BIC).

Table C5

Tests for differences between the role conditions

	G^2						AIC			BIC		
	free	restr.	Δ	*df*	*p*	p_{corr}	free	restr.	Δ	free	restr.	Δ
Role × Gender interaction effects												
π_3 (restricted first)	25.60	26.06	0.45	1	.501	> .999	41.60	40.06	-1.55	92.85	84.89	-7.95
γ_3 (restricted after π_3)	25.79	26.46	0.67	1	.414	> .999	39.79	38.46	-1.33	84.63	76.89	-7.74
γ_3 (restricted first)	26.06	26.46	0.40	1	.526	> .999	40.06	38.46	-1.60	84.89	76.89	-8.00
π_3 (restricted after γ_3)	39.54	41.97	2.43	1	.119	.595	49.54	49.97	0.43	81.56	75.59	-5.98
Role effects												
π_2 (restricted first)	38.62	41.97	3.35	1	.067	.404	48.62	49.97	1.35	80.64	75.59	-5.06
γ_2 (restricted after π_2)	26.46	38.62	12.16	1	< .001	.003	38.46	48.62	10.16	76.89	80.64	3.75
γ_2 (restricted first)	26.46	39.54	13.08	1	< .001	.002	38.46	49.54	11.08	76.89	81.56	4.67
π_2 (restricted after γ_2)	25.60	25.79	0.19	1	.664	.664	41.60	39.79	-1.81	92.85	84.63	-8.22

Table C5 continued

	G^2						AIC			BIC		
	free	restr.	Δ	df	p	p_{corr}	free	restr.	Δ	free	restr.	Δ

Note. N = 4470. The interaction effects of Role and Gender and the main effects of Role on cheating γ and the honest carrier prevalence π are tested using using G^2-difference tests and Akaike and Bayesian Information Criterion (AIC and BIC) differences. By row, single parameters are successively restricted to 0. Each resulting restricted ("restr.") model is compared to a more complex model, in which the respective parameter is estimated freely ("free"). p_{corr} refers to p-values adjusted for multiple testing using the Holmes-Bonferroni correction. The models for testing the interaction effects are derived from the full logistic model: $logit(\gamma) = \gamma_0 + \gamma_1 \cdot Gender + \gamma_2 \cdot Role + \gamma_3 \cdot Gender \cdot Role$;
$logit(\pi) = \pi_0 + \pi_1 \cdot Gender + \pi_2 \cdot Role + \pi_3 \cdot Gender \cdot Role$. The models for testing the Role effects are derived from the main effects model: $logit(\gamma) = \gamma_0 + \gamma_1 \cdot Gender + \gamma_2 \cdot Role$; $logit(\pi) = \pi_0 + \pi_1 \cdot Gender + \pi_2 \cdot Role$

Role effect on $\boldsymbol{\pi_{UQMC}}$: $\hat{\pi}_2$ = -0.81 on the logit scale, which means that the odds of reporting IPV are estimated to be $e^{-0.81} = 0.44$ times as high for participants inquired about perpetration as compared to victimization (i.e., taking the inverse, 2.25 times as high for participants inquired about victimization).

Role effect on $\boldsymbol{\gamma}$: $\hat{\gamma}_2$ = -2.84 on the logit scale, which means that the odds of cheating are estimated to be $e^{-2.84} = 0.06$ times as high for participants inquired about perpetration as compared to victimization (i.e., taking the inverse, 17.20 times as high for participants inquired about victimization).

From the effect size estimates, separate predictions for the UQMC parameters for both role conditions can be derived. For IPV victimization, the predicted honest carrier prevalence is $\pi_{vict} = 0.21$ and the predicted cheating prevalence is $\gamma_{vict} = 0.16$, both pooled across gender. For IPV perpetration, the predicted honest carrier prevalence is $\pi_{vict} = 0.11$ and the predicted cheating prevalence is $\gamma_{vict} = 0.01$, both pooled across gender.

B.3 Article III

Official citation:
Reiber, F., Schnuerch, M., & Ulrich, R. (2020). Improving the efficiency of surveys with randomized response models: A sequential approach based on curtailed sampling. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000353

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

http://dx.doi.org/10.1037/met0000353

Improving the Efficiency of Surveys With Randomized Response Models: A Sequential Approach Based on Curtailed Sampling

Fabiola Reiber
University of Tübingen

Martin Schnuerch
University of Mannheim

Rolf Ulrich
University of Tübingen

Abstract

Randomized response models (RRMs) aim at increasing the validity of measuring sensitive attributes by eliciting more honest responses through anonymity protection of respondents. This anonymity protection is achieved by implementing randomization in the questioning procedure. On the other hand, this randomization increases the sampling variance and, therefore, increases sample size requirements. The present work aims at countering this drawback by combining RRMs with curtailed sampling, a sequential sampling design in which sampling is terminated as soon as sufficient information to decide on a hypothesis is collected. In contrast to nontruncated sequential designs, the curtailed sampling plan includes the definition of a maximum sample size and subsequent prevalence estimation is easy to conduct. Using this approach, resources can be saved such that the application of RRMs becomes more feasible. An R Shiny web application is provided for simplified application of the proposed procedures.

Translational Abstract

Survey data are often subject to response biases, especially when sensitive (e.g., socially undesirable) characteristics are studied. However, protecting the respondents' anonymity can facilitate honest responding. Randomized response models (RRMs) achieve this goal by encrypting responses via random noise. Unfortunately, this noise increases uncertainty in the data and, therefore, large samples are required for sufficiently informative inference. To remedy this disadvantage, we propose to combine RRMs with a simple sequential testing procedure, that is, curtailed sampling. Following this approach, sample size requirements are reduced while still controlling statistical error probabilities. This way, resources can be saved such that the application of RRMs becomes more feasible. In this article, we describe how a curtailed sampling plan for RRM applications can be devised and how the respective data can be analyzed. We illustrate the procedure by means of simulations and reanalysis of empirical data. Additionally, we provide an easy-to-use R Shiny web application for simple implementation of the described procedures.

Keywords: sensitive questions, randomized response technique, sequential testing, curtailed sampling

Supplemental materials: http://dx.doi.org/10.1037/met0000353.supp

A large amount of findings in the human sciences is derived from studies relying on self-reports as the only available data source. However, self-reports are subject to biases, like the social desirability bias (Paulhus, 1991). This problem becomes especially pronounced, when the characteristic of interest is sensitive, that is, socially, morally, or even legally incriminating (see Tourangeau & Yan, 2007), such as environmental littering, endorsement of racist beliefs, drug abuse, or domestic violence. Survey respondents and interviewees are reluctant to disclose such incriminating information about themselves even

Fabiola Reiber, Department of Psychology, University of Tübingen; Martin Schnuerch, Department of Psychology, University of Mannheim; Rolf Ulrich, Department of Psychology, University of Tübingen.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG), Grant 2277, Research Training Group "Statistical Modeling in Psychology" (SMiP). We thank Jeff Miller for helpful comments on a first draft of this article, and David Izydorczyk for major contributions to the R Shiny web application for the procedures presented within this article. Parts of this article were presented at the 2019 annual meeting of the European Mathematical Psychology Group in Heidelberg, Germany. Simulated raw data and all simulation and analysis scripts are available on the Open Science Framework (https://osf.io/7kteu/).

Correspondence concerning this article should be addressed to Fabiola Reiber, Department of Psychology, University of Tübingen, Schleichstraße 4, 72076 Tübingen, Germany. E-mail: fabiola.reiber@uni-tuebingen.de

when they are assured confidentiality. Instead, responses to such questions are susceptible to selective nonresponding or dishonest responding (Tourangeau, Rips, & Rasinski, 2000). These self-protecting response tendencies do not only pose a problem in research focusing on the individual but also in research focusing on population characteristics. Specifically, these individual response tendencies distort inferences on the prevalence of the assessed characteristic.

Randomized response models (RRMs) are a class of questioning designs built to overcome this problem of self-protecting responses. RRMs assure anonymity protection of respondents by encrypting responses via a randomization process. They were originally developed (Warner, 1965) for investigating the prevalence of binary sensitive characteristics, like, for example, having consumed illicit drugs or not. In such cases, as explained before, a conventional prevalence estimate using the proportion of affirmative responses to a direct question is prone to be biased and likely underestimate the true prevalence because of self-protecting responses (see Krumpal, 2013; Tourangeau & Yan, 2007). In RRMs, in contrast, a randomization process involved in the questioning makes single responses inconclusive with respect to the individual manifestation of the sensitive characteristic. Therefore, the individual respondent's anonymity is protected. Nevertheless, drawing inferences on a group level is still possible knowing the probability underlying the randomization. This way, RRMs reduce the urge to give self-protecting responses and therefore enable a more valid assessment of the prevalence of sensitive attributes. RRMs have been applied in psychology and related fields to investigate prevalences of various sensitive topics; for examples, see Table 1. Readers interested in a comprehensive review of RRM applications are referred to Fox (2016).

Unfortunately, the validity increase in RRMs comes at a cost: The randomization, which is the key element of RRMs, induces additional noise. Compensating for this drawback requires large sample sizes—often more than 1,000 respondents—to allow for sufficiently powered inference (Ulrich, Schröter, Striegel, & Simon, 2012). Trying to reduce this demand on sample size by adjusting the inherent parameters of the design is always at the cost of anonymity protection, which would sabotage the intended purpose of RRMs.

The original RRM was followed by a large number of further developments (see Chaudhuri & Christofides, 2013; Fox, 2016, for overviews). Some developments focused on increasing validity by increasing the psychological acceptability of the questioning design. Others aimed at increasing efficiency by decreasing sampling variance through design adjustments. However, all RRMs use random encryption for creating anonymity and thus inherit, to various extents, both the validity advantages and efficiency disadvantages of the original RRM. This inevitable tradeoff is arguably one of the main reasons to restrain from applying RRMs.

However, altering the questioning design is not the only possibility to reduce sample size requirements. Indeed, there are procedures designed to make the sampling process itself more efficient, namely *sequential sampling* procedures (see, e.g., Wetherill, 1975). Instead of sampling a fixed number of observations, which is predefined based on power calculations, the data are monitored throughout the sampling process, and sampling is terminated as soon as a specified criterion is reached. As a consequence, if the data show a clear result, sampling can in many cases be stopped earlier and, thus, resources are saved. In this article, we demonstrate how RRMs can be incorporated in such a sequential sampling plan, namely *curtailed sampling* (see Wetherill, 1975), and how this can enhance the efficiency of RRM applications. First, we introduce two well-established RRMs to provide a better understanding of the mechanism driving the increase both in anonymity protection and in sampling variance. Second, we briefly outline the concept of sequential testing within curtailed sampling and how the two before described RRMs can be integrated in this sampling plan. Third, we describe how, following this procedure, unbiased prevalence estimates can be computed. Fourth, we demonstrate the efficiency of this curtailed RRM design by reanalyzing empirical data on physical doping. Finally, we discuss potential drawbacks and distinguish the present approach from other

Table 1
Exemplary RRM Applications in Psychology and Related Fields

Topic	Study	*N*
Induced abortion	Abernathy, Greenberg, & Horvitz, 1970	2,871
Rape victimization	Soeken & Damrosch, 1986	368*
Employee theft	Wimbush & Dalton, 1997	196
Job applicant faking	Donovan, Dwight, & Hurtz, 2003	221
Xenophobia	Ostapczuk, Musch, & Moshagen, 2009	606
Corruption	Gingerich, 2010	2,859
Dental hygiene	Moshagen, Musch, Ostapczuk, & Zhao, 2010	2,254
Poaching	Razafimanahaka et al., 2012	1,851
Cognitive enhancement	Dietz et al., 2013	2,557
Academic misconduct	Hejri, Zendehdel, Asghari, Fotouhi, & Rashidian, 2013	144
Organized crime	Wolter & Preisendörfer, 2013	333
Physical doping	Ulrich et al., 2018	2,168*
Prejudice against women leaders	Hoffmann & Musch, 2019	721

Note. This table contains exemplary studies applying RRM to investigate various sensitive topics. It serves to demonstrate the application range and does not comprise an exhaustive literature review. *N* = total size of the sample administered for the respective question using RRM.
* These samples consist of subsamples that were analyzed separately.

sequential procedures. In addition, we created a user-friendly R Shiny web application to apply the methods introduced in this article in substantive research.

Randomized Response Models

The Unrelated Question Model

In the first example, the unrelated question model (UQM; Greenberg, Abul-Ela, Simmons, & Horvitz, 1969), the sensitive question S of interest, for example, "Have you ever used illicit drugs?" is presented together with an unrelated neutral question N, for example "Is your mother's birthday between January and June inclusive?" Which of the two questions S and N a respondent has to answer depends on the outcome of a randomization device, like rolling a die. If, for example, the outcome is one, two, three, or four, the respondent is to answer the sensitive question S. By contrast, if the outcome is five or six, the respondent is to answer the neutral question N. Importantly, this outcome is known only to the respondent and only the response to either question is known to the interviewer. Therefore, the individual respondent's anonymity is protected because a "Yes" response can either mean "Yes, I have ever used illicit drugs" or "Yes, my mother's birthday is between January and June inclusive." The benefit of including a neutral question N is that any response is perceived as less stigmatizing because some responses have nothing to do with the sensitive topic. Figure 1 depicts the probabilities with which "Yes" or "No" responses are generated in the UQM. Clearly, a response can be generated without having to answer the sensitive question (lower branch). From this figure, the total probability λ of a "Yes" response is

$$\lambda_{UQM} = p \cdot \pi + (1 - p) \cdot q, \tag{1}$$

with probability p to receive the sensitive question S, prevalence π of the sensitive attribute and prevalence q of the neutral attribute. The neutral question N can be chosen such that q is known, like in the example above, where $q \approx .50$ under the assumption that birthdays are equally distributed over the year. The probability of a "Yes" response can be estimated from the proportion of "Yes" responses in a survey sample, leaving π the only unknown variable in Equation 1. Solving Equation 1 for π gives the estimator (see Greenberg et al., 1969)

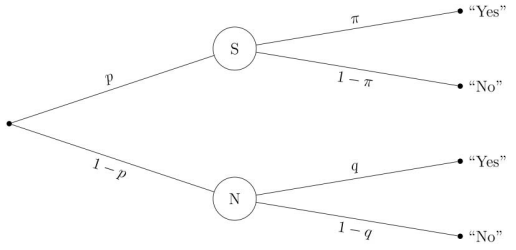


Figure 1. Probability tree of the UQM. The sensitive question S and the neutral question *N* are randomly received by respondents with probability p and $1 - p$, respectively. The probabilities of responding "Yes" and "No" to the neutral question *N* are q and $1 - q$ and the probabilities of responding "Yes" and "No" to the sensitive question S are π and $1 - \pi$.

$$\hat{\pi}_{UQM} = \frac{\hat{\lambda}_{UQM} - (1 - p) \cdot q}{p} \tag{2}$$

with sampling variance

$$Var(\hat{\pi}_{UQM}) = \frac{\lambda_{UQM} \cdot (1 - \lambda_{UQM})}{n \cdot p^2}. \tag{3}$$

As can be seen in Equation 3, the randomization procedure is reflected in the sampling variance through parameter p. In other words, the randomization adds variance to the sampling process and therefore impairs precision, leading to the above mentioned efficiency loss. To illustrate, the difference in required sample size between a direct question study and one that utilizes the UQM is depicted in the dashed and solid curves in Figure 2, respectively. This comparison is based on a common choice of UQM design parameters, that is, $p = .75$ and $q = .70$. Clearly, the required sample size is much larger in the UQM as compared with direct questioning. Especially in cases where a high precision is required ($SE = 0.01$) the difference becomes substantial and UQM applications are very costly compared with direct questioning.

The Crosswise Model

The second example is a newer development in the field of RRMs, the crosswise model (CWM; Yu, Tian, & Tang, 2008). It is a prominent model within a class of RRM developments labeled *nonrandomized response models*. They are named thus because no actual randomization device is part of the procedure although they make use of random encryption, anyway. In the CWM, like in the UQM, a sensitive question S is paired with a neutral question N, with known prevalence q. In this case q must not equal .50. In contrast to the UQM, respondents are not asked to respond to either of the questions based on the outcome of a randomization device but to give a combined response to both questions. As such, the answer categories are "A: My response to both questions is the same" (i.e., "Yes" to both or "No" to both) and "B: My response to both questions differs" (i.e., "Yes" to one and "No" to the other). In addition to evading the need for a randomization device, this procedure has the advantage of not asking for a confirming or dismissive response. Instead, the response categories themselves are neutral with respect to the sensitive attribute.[1] The response generating probabilities are depicted in Figure 3. From this figure, the probability of an "A" response can be derived as

$$\lambda_{CWM} = q \cdot \pi + (1 - q) \cdot (1 - \pi). \tag{4}$$

Clearly, any response can come from both a carrier and a noncarrier of the sensitive attribute, depending on that person's status on the neutral attribute. Because the latter status is not known, the individual respondent's anonymity is protected. However, because the probability of carrying the neutral attribute is known, the group-level prevalence can still be estimated by (see Yu et al., 2008)

[1] Of course, the probability of carrying the sensitive attribute is not the same given different responses. For $q > .50$, $P(C|\text{"A"}) > P(C|\text{"B"})$ and for $q < .50$, $P(C|\text{"A"}) < P(C|\text{"B"})$. For example, the odds of being a carrier are nine times larger given an "A" than given a "B" response for $q = .75$. However, it is unlikely that respondents' decisions are influenced by this.

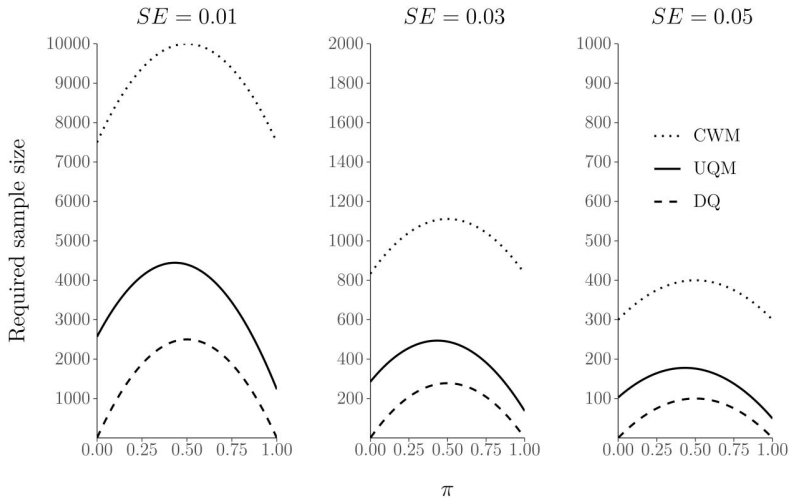


Figure 2. Required sample size depending on questioning type. Depicted is the required sample size as a function of the true prevalence π. The curves within a panel depict the questioning types: Direct question (DQ, dashed), unrelated question model (UQM, solid) and crosswise model (CWM, dotted). The design parameters are $p = .75$, $q = .70$ in the UQM and $q = .75$ in the CWM. The panels differ in the estimate's standard error (*SE*) 0.01, 0.03, and 0.05 from left to right. Note the individual *y*-axis scaling of each panel.

$$\hat{\pi}_{CWM} = \frac{\hat{\lambda}_{CWM} - 1 + q}{2q - 1} \tag{5}$$

with sampling variance

$$Var(\hat{\pi}_{CWM}) = \frac{\lambda_{CWM} \cdot (1 - \lambda_{CWM})}{n \cdot (2q - 1)^2}. \tag{6}$$

The increase in variance induced in this procedure is even higher than in the UQM as is visible in the dotted curves in Figure 2. Like for the UQM, the CWM specification used in this demonstration represents a common choice of design parameters, that is, $q = .75$. Thus, despite the high face-validity of the CWM, its applicability is impaired by its excessive costs in sample size.

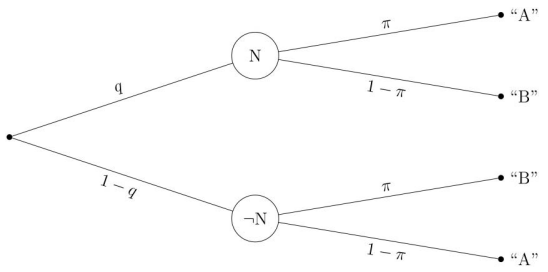


Figure 3. Probability tree of the CWM. Respondents are asked to respond to both questions S and N in one response, "A" or "B." Respondents carry the neutral attribute N with known probability q or do not carry it ¬N with probability $1 - q$. Carriers of the neutral attribute respond "A" with probability π because they carry the sensitive attribute and thus their response to both questions is the same. They respond "B" with probability $1 - \pi$ because they do not carry the sensitive attribute and thus their response to both questions differs. Noncarriers of the neutral attribute respond "B" with probability π because they carry the sensitive attribute and thus their response to both questions differs. They respond "A" with probability $1 - \pi$ because they do not carry the sensitive attribute and thus their response to both questions is the same. Note that the order of the two questions in the tree is arbitrary and is not meant to imply a sequential process. Instead, respondents answer both questions simultaneously and the order in the tree could just as well be reversed.

In conclusion, RRMs ensure individual anonymity protection by design and therefore provide researchers with a tool to acquire estimates less distorted by self-protecting responses. However, the application of these procedures is impaired by high sample size requirements because of the additional variance induced by randomization. This is especially problematic whenever respondents are difficult to recruit. Such difficulties arise, for example, when a special population is investigated (e.g., elite athletes in a survey of physical doping) or when taking part in the survey involves obstacles, such as fear of being stigmatized (e.g., for being addicted to drugs). Both these scenarios are not unlikely in research on sensitive topics, which is the field of applications of RRMs.

Hypothesis Testing With Randomized Response Models

This problem of high sample size requirements is relevant in studies focusing on prevalence estimation as well as in those focusing on hypothesis testing. There has been a general debate on the justification of hypothesis testing as compared with parameter estimation in psychology. Specifically, some authors argue that parameter estimation provides more informative results and should become the standard data analysis procedure (Cumming, 2014). However, others argue that "[n]either hypothesis testing nor estimation is more informative than the other; rather, they answer different questions" and "hypothesis testing, not estimation, is necessary for testing the quantitative predictions of theories" (Morey, Rouder, Verhagen, & Wagenmakers, 2014, p. 1290; see also, Anderson, 2019). Thus, the choice between estimation and hypothesis testing should be based on the research question.

In fact, the RRM literature features many studies addressing research questions that conform to hypothesis tests. For example, many studies investigating the validity advantage of RRMs make use of the *more-is-better assumption*, that is, they investigate whether prevalences of sensitive attributes are inferred to be higher when assessed using RRM questioning as compared with direct questioning (see Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). Although in most studies following this assumption prevalence estimates are compared between the questioning designs (e.g., Nordlund, Holme, & Tamsfoss, 1994; Wimbush & Dalton, 1997; Wolter & Preisendörfer, 2013), this is actually a research question that calls for a hypothesis test with decision error control (as in Hoffmann & Musch, 2016).

Likewise, in substantively motivated RRM applications, there are research questions that are best addressed using hypothesis tests. As such, an application of RRMs is often motivated by the question: Is a certain sensitive attribute really as small as one concludes from conventionally collected data? This question is reasonable whenever estimates from direct questioning or other commonly used data sources are surprisingly low.[2] The straightforward statistical approach to such a question is a statistical test of the hypothesis that the RRM estimate is higher than the conventional estimate.

Apart from being the theoretically suitable approach for certain research questions, hypothesis tests require smaller samples than precise estimation. However, the required RRM samples are still very large, as compared with hypothesis tests in the context of direct questions (Ulrich et al., 2012). To address this drawback, RRMs can be incorporated in a sampling framework that is designed to be economic in terms of sample size, namely, sequential testing or, more precisely, curtailed sampling.

Curtailed Sampling

When testing hypotheses about whether the prevalence of an attribute lies in a certain range, classical data collection requires the definition of a fixed sample size to achieve the requested statistical power. In RRM studies, this usually leads to very large sample size requirements, as explained before. A group of procedures aimed at minimizing sample size requirements are sequential tests. As stated above, the general idea of sequential tests is to terminate sampling as soon as sufficient support for the hypotheses is allocated, instead of continuing until some predefined sample size is reached. The rationale for this procedure is that in some cases, sufficient certainty for a decision might be present at an earlier stage and, thus, further sampling would constitute a waste of resources. Detecting this early support requires, as the name indicates, sequential testing throughout the sampling process. The challenge is, certainly, to design a sampling plan such that Type-1 and Type-2 decision errors are still controlled for.

There exists a variety of sequential sampling procedures. Among these, one basic procedure, applicable to binomial data, is curtailed sampling (see Wetherill, 1975). In curtailed sampling, data collection is terminated corresponding to stopping rules that apply when sufficient evidence for making a decision is obtained. The stopping rules are defined by the maximum sample size N_{max} and a bound c_s, denoting the amount of observed successes required to reject the null hypothesis. These parameters are equal to the fixed sample size and the critical value in a Neyman-Pearson test with (upper bound) Type-1 and Type-2 decision error probabilities α and β, respectively (Wetherill, 1975). In the classical Neyman-Pearson test, an a priori defined number of observations $N = N_{max}$ is sampled. If the number of successes among these observations exceeds the critical value c_s, the null hypothesis is rejected. Otherwise, it is maintained. The rationale of the curtailed sequential test is that if c_s successes are observed at any point before reaching N_{max}, the test will always reject the null hypothesis at N_{max}. Therefore, in contrast to the Neyman-Pearson test, instead of continuing the sampling process until $N = N_{max}$ is reached, it can be terminated as soon as c_s successes have been observed, thus rejecting the null hypothesis. In the same vein, if $c_f = N_{max} - c_s + 1$ failures are observed during the sampling process, the test will always maintain the null hypothesis at N_{max}, because the critical value c_s of successes cannot longer be reached. Hence, it can be terminated already at this point, thereby rejecting the alternative hypothesis.

The horizontal and vertical lines in Figure 4 display these two bounds, while the diagonal line denotes the maximum sample size N_{max}, for an exemplary UQM sampling plan described in more detail later. This diagonal line also represents the fixed sample size of a corresponding Neyman-Pearson test. In the context of the UQM (CWM), successes are defined as "Yes" responses and failures as "No" responses ("A" and "B" responses). Thus, N_{max} is the maximum number of all responses before sampling is stopped and c_s is the minimum number of "Yes" responses ("A" responses) required before rejecting Hypothesis H_0.

The parameters N_{max} and c_s depend on the hypotheses about the prevalence π of the sensitive attribute and the specified error probabilities. If, for example, one wants to construct a sampling plan that tests the Hypothesis H_0 that a sensitive attribute has a prevalence of at most $\pi_0 = .05$ against the Hypothesis H_1 that the prevalence is at least $\pi_1 = .15$ with error probabilities $\alpha = .05$ and $\beta = .10$, the following needs to be considered. The probability of deciding in favor of H_0 should be $1-\alpha = .95$ at $\pi = .05$ and $\beta = .10$ at $\pi = .15$. In the area between π_0 and π_1, termed the zone of indifference (Wetherill, 1975), no clear preference for a decision in favor of one of the two hypotheses exists. The resulting probabilities of a correspondingly constructed curtailed sampling procedure for deciding in favor of H_0 for all possible values of π are illustrated by the operating characteristic (OC) curve in Figure 5. The curve in Panel A depicts the straightforward case in which the probability of an affirmative response equals the prevalence π, that is in direct questioning.

However, in RRMs the probability of an affirmative response is not π but λ, which is a linear transformation of π and depends on the design parameters of the RRM. For example, in case of the UQM the probability of a "Yes" response, λ_{UQM}, can be computed from π using Equation 1. The curve in Panel B of Figure 5 depicts the resulting probabilities for deciding in favor of H_0, now with respect to λ_{UQM}. This demonstrates how the UQM influences the sampling plan requirements: The zone of indifference becomes narrower and, therefore, the differentiation between the competing hypotheses becomes more difficult. Specifically, larger N_{max} and

[2] The study presented in the section Sequential Reanalysis of Empirical Data later in this article is an example for such a case.

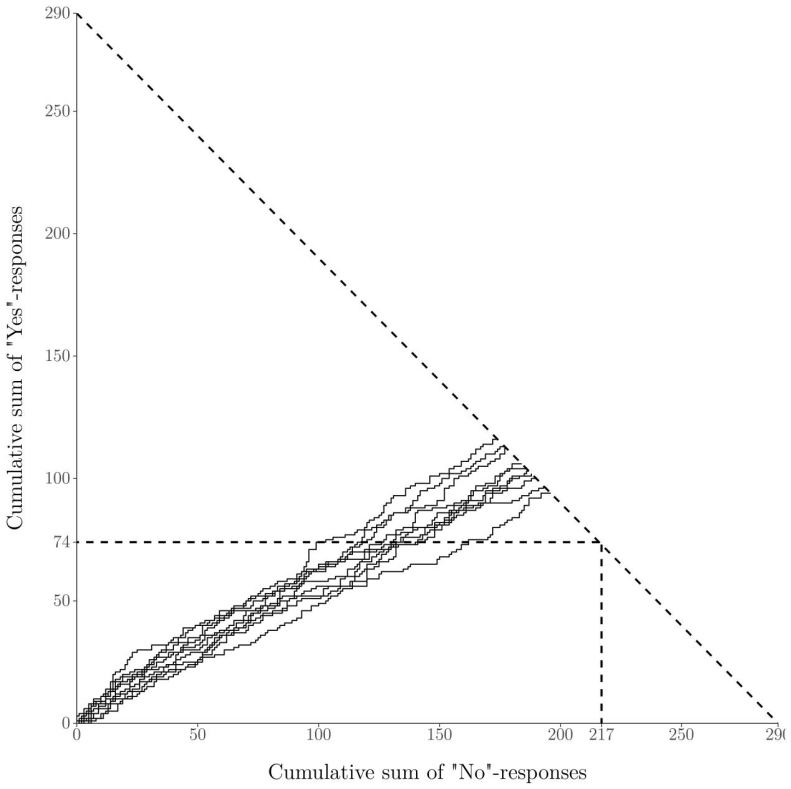


Figure 4. Sampling paths of simulated samples. The 10 samples were simulated as unrelated question model data with design parameters $p = .75$, $q = .70$, and true prevalence $\pi = .25$. The depicted bounds are (a) the maximum number of "Yes" responses $c_s = 74$ (horizontal); (b) the maximum number of "No" responses $c_f = 217$ (vertical); and (c) the maximum total number of responses $N_{max} = 290$ equaling the fixed sample size of a Neyman-Pearson test (diagonal). They are based on the hypotheses $\pi_0 = .05$ and $\pi_1 = .15$ with $\alpha = .05$ and $\beta = .10$.

c_s are required such that the error probability requirements are fulfilled for these stricter hypotheses.

Determination of the Sampling Plan Parameters

To determine N_{max} and c_s, a priori power analyses need to be conducted. Exact values such that the resulting error probabilities are closest to, but never larger than, α and β can be determined by a numerical search algorithm. This algorithm searches for the smallest N_{max} which, in combination with a corresponding c_s, meets the requirements. Specifically, it iteratively searches all possible values for N_{max} along the lines of the following four steps:

1. Starting with an initial N_{max}, a c_s is derived by computing the inverse of the CDF specified by the current N_{max} and λ_0 for the cumulative probability $1 - \alpha$.

2. The CDF specified by the current N_{max} and λ_1 is evaluated at the current c_s.

3. As long as the resulting cumulative probability is larger than β, N_{max} is increased by $+1$ and the procedure is repeated.

4. As soon as the resulting cumulative probability is smaller or equal to β, the search is terminated and the algorithm returns the current instantiations of N_{max} and c_s as suitable sampling plan parameters.

The respective pseudocode can be obtained from Section A of the online supplemental materials.

In the above mentioned UQM example (see Figure 4) with the design parameters $p = .75$ and $q = .70$, the parameters defined by an exact power analysis are $N_{max} = 290$ and $c_s = 74$, for testing the hypotheses $\pi_0 = .05$ and $\pi_1 = .15$ with $\alpha = .05$ and $\beta = .10$. Thus, the stopping rules in this case are defined as: Stop sampling if (a) the number of "Yes" responses reaches $c_s = 74$ or (b) the number of "No" responses reaches $c_f = 217$. It is possible that when either (a) or (b) is the case, the maximum number of responses $N_{max} = 290$ is reached, but it can never be exceeded.

Figure 4 depicts at what point the sampling paths of 10 simulated samples[3] reach one of the bounds. The mean sample size when a bound is reached is $\bar{N} = 204.00$ with $SD = 16.90$. The example demonstrates the advantages of a curtailed sampling design: The actual sample size N is no longer a fixed value but a random variable with maximum N_{max} and an expected value lower

[3] The simulation was conducted with the above described design parameters $p = .75$, $q = .70$, $\alpha = .05$, $\beta = .10$, $\pi_0 = .05$, $\pi_1 = .15$, the resulting bound-values $c_s = 74$ and $N_{max} = 290$ and true prevalence $\pi = .25$.

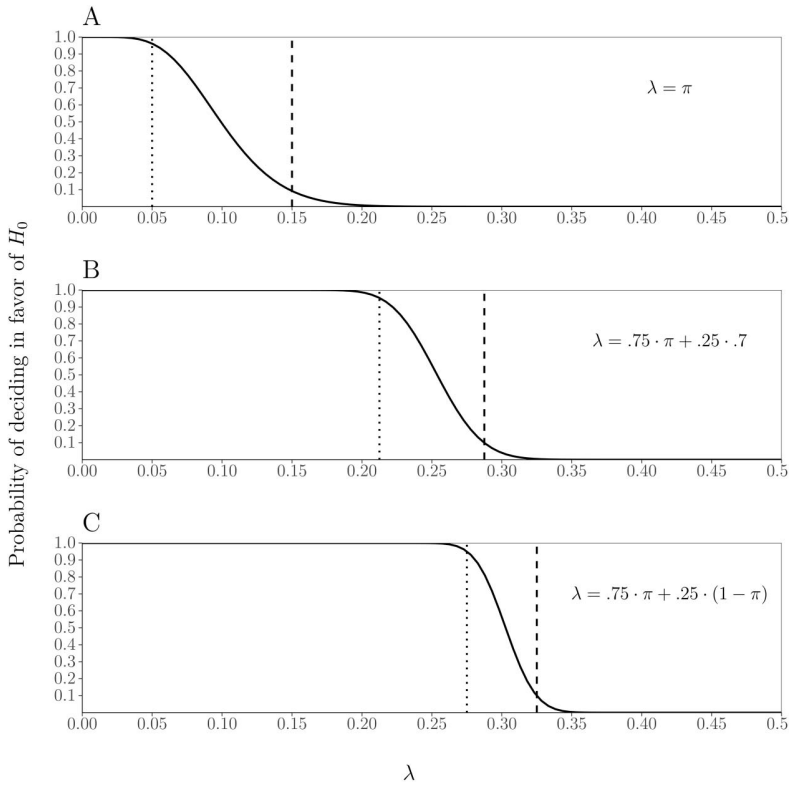


Figure 5. Operating characteristic curve. Depicted is the probability of deciding in favor of H_0 depending on the true probability of a "Yes" response λ in a curtailed sampling plan. All panels refer to the same hypothesis test concerning the prevalence π: $\pi_0 \leq .05$ (dotted line) versus $\pi_1 \geq .15$ (dashed line) with $\alpha = .05$ and $\beta = .10$. The panels differ with respect to the questioning design. Panel A depicts the case of direct questioning, such that the probability of a "Yes" response λ equals the prevalence π. Panel B depicts the case of the UQM, such that the probability of a "Yes"-response λ is a transformation of π using Equation 1, in this example with design parameters $p = .75$ and $q = .70$. Panel C depicts the case of the CWM, such that the probability of a "Yes" response λ is a transformation of π using Equation 4, in this example with design parameter $q = .75$.

than N_{max}. Specifically, in this example, the mean sample size saving is $N_{max} - \bar{N} = 290 - 204 = 86$. In other words, one can come to a conclusion earlier and, therefore, less resources are needed.

Efficiency of the Curtailed Sampling Plan

The extent of this advantage can be illustrated by the average sample number (ASN) curve in Figure 6. It depicts the expected sample size when reaching either of the bounds as a function of the true parameter value π. The average sample size per value of π is calculated by the possible sample sizes *N* weighted by their probability of occurrence. Specifically, the ASN curve of the above example (left panel) has its maximum of $N = 278.53$ at $\pi = .081$. For $\pi > .26$ the expected sample number drops below 200 and for $\pi > .75$ below 100. As a comparison, the necessary sample size for a classical test would always be N_{max}, that is, 290. Thus, the expected *N* in the curtailed sampling plan is always smaller than the sample size required by classical analyses. Especially if π is notably smaller or larger than the decision relevant values π_0 and π_1, respectively, the sample size saving is substantial. What is more, the sample size required by the curtailed design can never exceed that of the classical analysis. Additional ASN curves for varying UQM design specifications are provided in Section B of the online supplemental materials.

The same line of reasoning applies to any other RRM. The only difference between varying models lies in the beforehand transformation of the prevalence π to the actual response probability λ. In case of the CWM, this is done using Equation 4 and gives the probability λ_{CWM} of "A" responses. The curve in Panel C of Figure 5 depicts how this affects the testable hypotheses derived from the same hypotheses concerning π as in the example above ($\pi_0 = .05$ and $\pi_1 = .15$ with $\alpha = .05$ and $\beta = .10$). Clearly, the zone of indifference is even smaller than in the UQM, which is in line with the larger sampling variance of the CWM. In a CWM design with $q = .75$ this test requires a curtailed sampling plan with $N_{max} = 722$ and $c_s = 219$. Again, the impact on expected sample size manifests in the ASN curve in Figure 6 (right panel). Not surprisingly, N_{max} and the expected sample size exceed the corresponding values in the UQM. However, compared to the sample size required by a classical test, the saving in sample size can, again, be substantial, especially if the true prevalence is far from the indifference zone. Additional ASN curves for varying CWM design

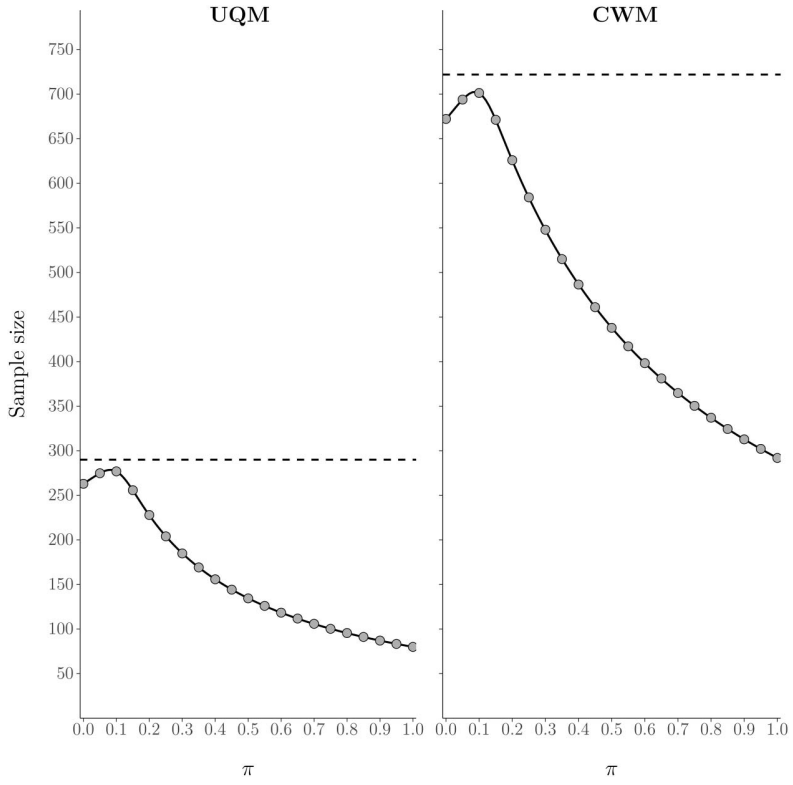


Figure 6. Average sample number curves. The solid curves depict the expectation of the sample size N when reaching one of the bounds of a curtailed sampling plan as a function of the prevalence π. The dashed lines depict the maximum sample size N_{max}. The gray dots depict the mean sample size $\bar{N}$ of 10,000 samples simulated for each prevalence value π. The two panels correspond to different sampling plans, one applying the unrelated question model (UQM, left panel) and the other one applying the crosswise model (CWM, right panel). Both sampling plans are defined with respect to the hypotheses $\pi_0 = .05$ and $\pi_1 = .10$ with $\alpha = .05$ and $\beta = .10$. For the UQM, the design parameters are $p = .75$, $q = .70$ and the resulting bound-values are $c_s = 74$ and $N_{max} = 290$. For the CWM, the design parameter is $q = .75$ and the resulting bound-values are $c_s = 219$ and $N_{max} = 722$.

specifications are also provided in Section B of the online supplemental materials.

Subsequent Estimation

Despite the previously discussed theoretical legitimization of hypothesis testing, subsequent prevalence estimation can be desirable after conducting the hypothesis test. Estimation following a statistical test is straightforward in a fixed-sample design, but sequential sampling may introduce considerable bias for conventional maximum-likelihood estimators (e.g., Whitehead, 1986).

Even though the same holds true for curtailed sampling procedures when relying on conventional estimators, unbiased estimation is feasible by using adjusted inverse binomial sampling estimation. In inverse binomial sampling, rather than sampling a predefined number of observations, evidence is collected until a certain number N_s of confirmative responses is obtained. The prevalence estimate then depends on the distribution of the total number of responses until this number is reached. Similarly, in curtailed sampling, estimation can be conducted depending on the distribution of the total number of responses when one of the bounds is reached.

Inverse binomial sampling follows a negative binomial distribution and, therefore,

$$\hat{\lambda} = \frac{N_s - 1}{N - 1} \tag{7}$$

is an unbiased estimator of the probability λ of a confirmative response (Haldane, 1945). This probability estimate $\hat{\lambda}$ refers to the probability of a "Yes" response or "A" response in the UQM or CWM, respectively, and can thus be transformed to the prevalence estimate $\hat{\pi}$ using Equations 2 and 5. The inverse binomial sampling estimator in Equation 7 can be applied to data assessed in a curtailed sampling plan, whenever sampling is stopped because the boundary c_s of confirmative responses is reached. In this case, $N_s = c_s$.

If, however, sampling is stopped because the bound c_f of dismissive responses is reached, the results do not follow a negative binomial distribution. Therefore, the estimator in Equation 7 is not an unbiased estimator of the probability of a confirmative response

in these cases. However, this requirement is fulfilled by the adjusted estimator

$$\hat{\lambda} = \frac{N - c_f}{N - 1}. \quad (8)$$

Indeed, the combination of both estimators yields a joint probability distribution of estimates with expectation equal to the true prevalence. Further details on the derivation of these estimators are provided in Section C of the online supplemental materials.

To illustrate the properties of the combined estimators, Figure 7 shows the theoretical sampling distributions of estimates attainable from the example curtailed sampling plan for the UQM introduced in the previous section (for testing the hypothesis whether $\pi \leq .05$ against $\pi \geq .15$). Specifically, each panel depicts the probabilities of attaining the possible prevalence estimates for a specific true prevalence value. The two estimators are marked by different shades of gray. Notably, there is a violation of normality at the transition point between the application ranges of the two different estimators. Nevertheless, the expectation of the combined estimators equals the true prevalence indicating unbiasedness.

In the same vein, Figure 8 shows the frequencies of prevalence estimates obtained from simulated samples under the same example sampling plan with the same true prevalence values as in Figure 7. Each panel depicts the frequency distribution of the subsequent prevalence estimates from 100,000 samples simulated with the respective true prevalence value. The mean of the estimates is close to the true values in all six cases with negligible bias, that is, $\overline{bias} = 0.00013$ with $SD = 0.00010$.

Given the non-normality of the estimates' probability distribution, the determination of confidence intervals using the sampling variance is not recommendable. Instead, the Clopper-Pearson interval (Clopper & Pearson, 1934) can be calculated.[4] The parameter coverage of the thus calculated 95% confidence intervals for the estimates of the simulated samples in Figure 8 is .954. The deviation from .95 is explainable by the discrete distribution which does not allow for exact cutoffs leading to a more conservative confidence interval.

The preceding analyses demonstrate that subsequent unbiased estimation following curtailed sampling is feasible within UQM. Importantly, the same holds for all types of RRMs. As such, equivalent probability distributions and parameter recovery distributions are obtainable for various curtailed sampling plans.

R-Scripts and a user guide for the application of the procedures described in the two preceding sections are available on the Open Science Framework (OSF; https://osf.io/7kteu/). The scripts provide functions for the determination of the sampling plan parameters N_{max} and c_s for given hypotheses, for plotting the OC and ASN curve for a given sampling plan and for analyzing and plotting curtailed sampling data. Additionally, an R Shiny web application (Chang, Cheng, Allaire, Xie, & McPherson, 2020), which requires no prior knowledge of R, is available on https://fabiolareiber.shinyapps.io/CurtailedRRT/ for easy application of these procedures. All reported simulation and analysis scripts are also available from the OSF.

Sequential Reanalysis of Empirical Data

The following reanalysis illustrates the benefit of curtailed sampling in the framework of the UQM. Ulrich et al. (2018) applied the UQM to assess the dark figure of doping at two international athletics competitions, namely the 13th International Association of Athletics Federations World Championships in Athletics (WCA) in Daegu, South Korea, and the 12th Quadrennial Pan-Arab Games (PAG) in Doha, Qatar, both in 2011. The application of the UQM elicited doping prevalence estimates that substantially exceed common estimates derived by direct questioning or biological testing, $\hat{\pi}_{WCA} = 43.6\%$ (95% CI [39.4 – 47.9]) and $\hat{\pi}_{PAG} = 57.1\%$ (95% CI [52.4 – 61.8]) as compared with estimates of 2% reported by the World Anti-Doping Agency (2012) for the same year 2011. These estimates were obtained with this level of precision on the basis of sample sizes of 1,203 and 965 at WCA and PAG, respectively, and were therefore associated with correspondingly high costs. However, what made these estimates so interesting was not their exact size, but that they were much higher than usual estimates. Importantly, as highlighted above, such a finding is attainable through hypothesis testing and it does not require precise estimation. Indeed, it is possible to conduct a sequential test with curtailed sampling, because the hypothesis concerning the prevalence is simple: Is the doping prevalence estimated with the UQM higher than usual estimates or not?

As prevalence estimates from official doping tests in elite athletics are very low (World Anti-Doping Agency, 2012), the following test seems reasonable. Hypothesis H_0 states that doping is virtually nonexistent, that is 2% like in the official testing figures, and Hypothesis H_1 states that the prevalence is above 10%. Thus, when $\pi = .02$, H_0 should be selected with at least probability $1 - \alpha = .95$ and when $\pi = .10$, H_0 should be selected with at most probability $\beta = .10$, to preserve sufficient decision error control. Given the design parameters $q = .50$ and $p = .67$ applied in the study, the minimal values for N_{max} and c_s of a curtailed sampling plan meeting the test's requirements can be calculated as 490 and 102, respectively.

When reanalyzing each of the two samples sequentially, in the order, in which they were assessed, a decision in favor of H_1 that the prevalence is equal to or above 10% would have been reached markedly ahead of time, with sample sizes of 262 in the WCA sample and 199 in the PAG sample, when reaching the bound of "Yes" responses $c_s = 102$. The corresponding sampling paths are depicted in Figure 9. In 1,000 random permutations of each sample the bound of "Yes" responses is reached in all cases. The mean sample size when reaching the bound is 222.60 and 186.31 in the WCA and PAG samples, respectively. In sum, sequential testing would have led to accepting the hypothesis that the doping prevalence is higher than suggested by official testing figures and thereby provided conclusions in the same direction as the original results with markedly lower sample size requirements and decision error control.

Following the sequential hypothesis test, the estimation procedure proposed in the previous section can be applied to the data. The estimates computed using the subsequent estimation procedure on the data available at the point in sampling, when the decision would have been made, are listed in Table 2 together with the conventionally computed original estimates. Both estimates are below the estimates calculated from the fixed samples but the

[4] Highest density intervals can be calculated as an alternative approach.

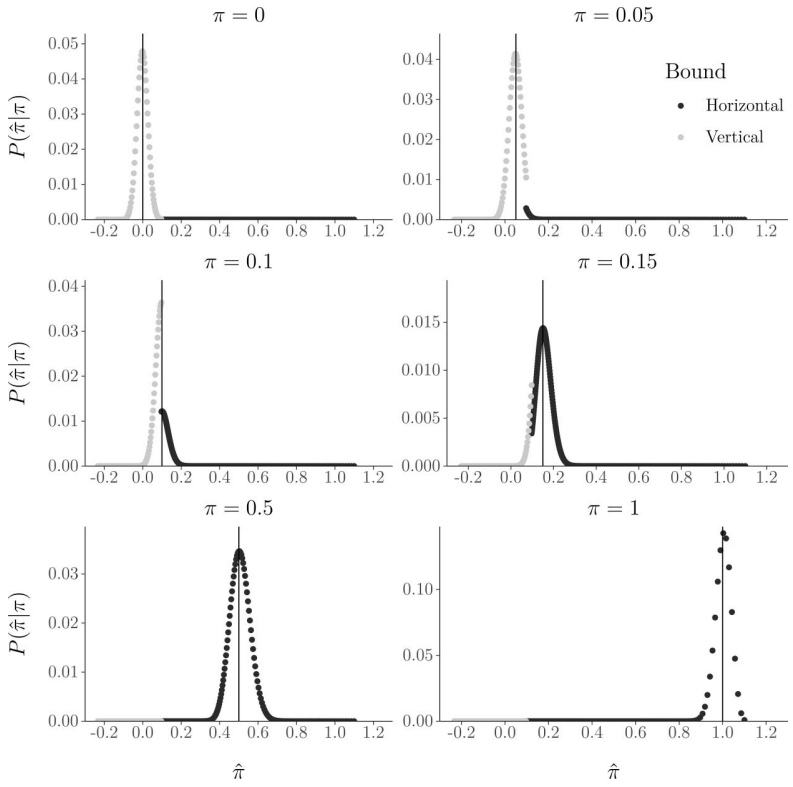


Figure 7. Theoretical sampling distributions of $\hat{\pi}|\pi$. Depicted are the probabilities of obtaining a certain estimate after testing the hypotheses $\pi_0 = .05$ and $\pi_1 = .15$ with $\alpha = .05$ and $\beta = .10$ in a curtailed sampling plan using the UQM with design parameters $p = .75$, $q = .70$. The panels differ with respect to the true prevalence value, from $\pi = 0$ in the top left panel to $\pi = 1$ in the bottom right panel, which is indicated by the vertical line in each panel. The estimates marked by black points are obtainable from samples in which the horizontal bound $c_s = 74$ is reached and are calculated using the estimator in Equation 7. The estimates marked by gray points are obtainable from samples in which the vertical bound $c_f = 217$ is reached and are calculated using the estimator in Equation 8.

confidence intervals overlap in both cases. Naturally, the confidence intervals of the sequential estimates are much larger than those calculated from the whole sample, as is to be expected from a study using a randomized response design with a sample size this small. Still, it is possible to narrow down the range of the prevalence estimates by this procedure at much lower cost than in classical fixed sample size studies. Importantly, the mean of the estimates computed from all 1,000 permutations of the data nearly equals the original estimates, 43.7 and 57.0 for WCA and PAG, respectively. This confirms that the deviation of the sequential estimate from the original estimate is due to random sampling error. Thus, subsequent estimation as an additional step after the sequential test could have served to acquire more precise information on the doping prevalence.

Discussion

Randomized response models provide means to increase the validity of estimates of sensitive attributes. Yet, their applicability is impaired by high demands on sample size due to random noise induced by the questioning design. Especially when aiming at sufficiently powered statistical inference, this can lead to very large required sample sizes. Combining RRMs with sequential testing by means of a curtailed sampling plan can ameliorate this drawback. Especially when the true prevalence is well outside the zone of indifference between the decision relevant values of the hypothesis test, considerable sample size savings are possible. In such cases, conclusions concerning the possible range of the prevalence of a sensitive attribute can be drawn with decision error control and at much lower cost than in classical fixed sample size studies. Additionally, subsequent estimation of the prevalence of interest using closed form estimators adjusted to the outcome of the hypothesis test can serve to acquire additional information. Reanalysis of data of a large scale UQM-study on the prevalence of doping in elite athletics (Ulrich et al., 2012) shows that results pointing in the same direction can be obtained at much lower cost using curtailed sampling and subsequent estimation.

However, comparing the results of the conventional estimation to those of the subsequent estimation within curtailed sampling highlights a limitation: The estimates are not as precise when sampling is conducted in a curtailed sampling plan. However, this is not surprising, as the goal of curtailed sampling is sample size reduction and estimates from a study with smaller sample size will

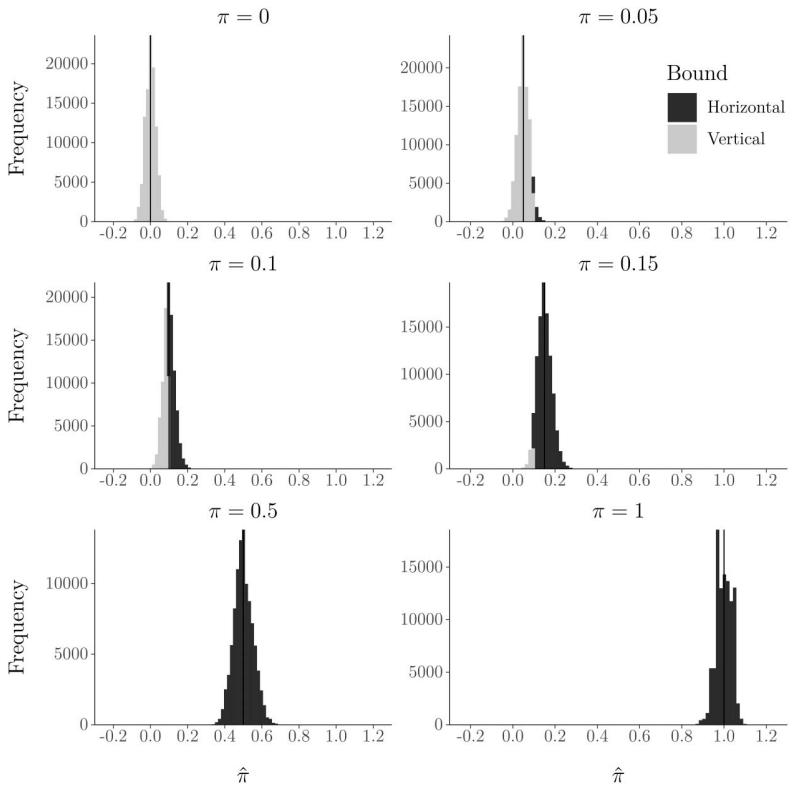


Figure 8. Simulated sampling distributions of $\hat{\pi}|\pi$. Depicted are frequency distributions of prevalence estimates $\hat{\pi}$ calculated from simulated samples using the information available in the moment when sampling would have been stopped in a curtailed sampling plan for testing the hypotheses $\pi_0 = .05$ and $\pi_1 = .15$ with $\alpha = .05$ and $\beta = .10$. Samples were simulated in a UQM design with design parameters $p = .75$, $q = .70$. The panels differ with respect to the true prevalence value, from $\pi = 0$ in the top left panel to $\pi = 1$ in the bottom right panel, which is indicated by the vertical line in each panel. Each panel includes a total of 100,000 simulated samples. The estimates from samples in which the horizontal bound $c_s = 74$ was reached are depicted in black and were calculated using the estimator in Equation 7. The estimates from samples in which the vertical bound $c_f = 217$ was reached are depicted in gray and were calculated using the estimator in Equation 8.

always be less precise. Moreover, this is not a real flaw of the method, because curtailed sampling is not designed for precise estimation but for hypothesis testing. Therefore, as stressed before, it should be applied only if the research question involves a test of sensible hypotheses. Specifically, this could include testing whether a prevalence is in a relevant range in a pilot study, testing whether estimates changed in a replication study, or testing whether RRM estimates differ from estimates derived using other methods in a validation study. In such cases, curtailed sampling can substantially increase efficiency and is recommendable.

There are also sequential methods developed specifically for estimation (e.g., Kelley, Darku, & Chattopadhyay, 2017). For instance, the basic rationale of Kelley, Darku, and Chattopadhyay (2017) is to sample until the confidence interval of the estimate is smaller or equal to a desired width. The advantage of this approach is that no assumptions on unknown parameters are necessary, as is the case when one determines the necessary fixed sample size for a sufficiently precise estimate beforehand. As discussed in the introduction of this article, both approaches have advantages and the choice between hypothesis testing with error control and precise parameter estimation should depend on the research question. In either case, a sequential design can increase sampling efficiency.

Within the hypothesis testing framework, it is important to keep in mind that the curtailed sampling plan only applies to the test of simple hypotheses, that is, hypotheses in which all parameters are either known or specified by the hypotheses. This is not the case, for example, if the RRM includes additional unknown parameters to account for cheating behavior (Clark & Desharnais, 1998; Reiber, Pope, & Ulrich, 2020). In the same vein, hypotheses on prevalence differences between groups are not simple, unless the concrete group prevalences are specified. Obviously, classical tests have the same limitation because a priori power analyses require specification of all parameters. However, as in classical analysis, it is possible to define a curtailed sampling plan for composite hypotheses based on conservative (i.e., extreme) assumptions about the unknown parameters. In this case, the error probabilities of the procedure denote upper limits which will hold for any parameter values less extreme than specified. Note, however, that a conservative assumption will result in a less efficient test.

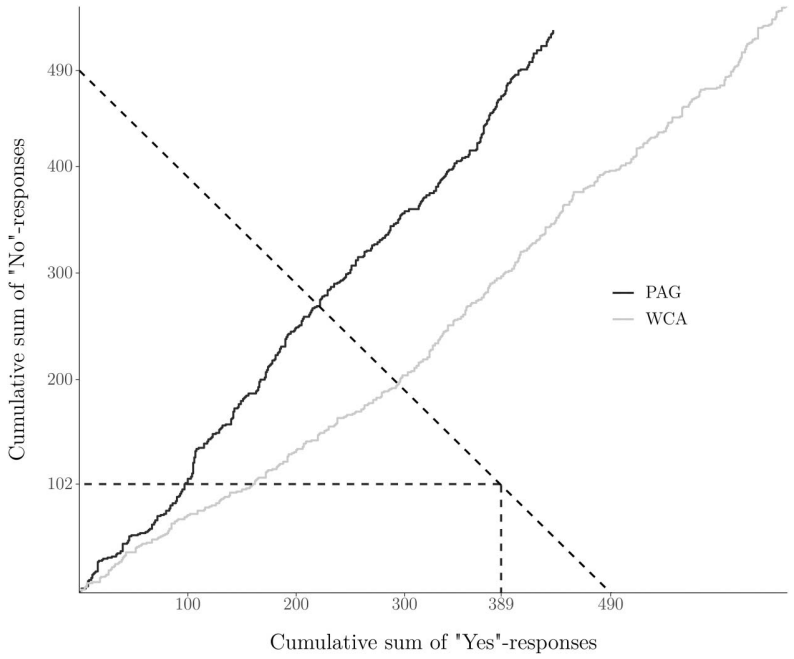


Figure 9. Sampling paths of the two samples in the doping study. The underlying design parameters are $p = .50$, $q = .67$. The depicted bounds are $c_s = 102$ (horizontal), $c_f = 388$ (vertical), and $N_{max} = 490$ (diagonal) and are based on the hypotheses $\pi_0 = .02$ and $\pi_1 = .10$ with $\alpha = .05$ and $\beta = .10$. The two samples were assessed at the World Championships in Athletics (WCA) in 2011 in Deagu, South Korea, and at the Pan-Arab Games (PAG) in 2011 in Doha, Qatar.

For simple hypotheses, we demonstrated that the curtailed sampling plan is more efficient than classical analysis. However, curtailed sampling is not the only sequential testing procedure. Another efficient procedure is the well-known sequential probability ratio test (SPRT; Wald, 1947). Here, the likelihood ratio of two competing hypotheses is continuously computed throughout sampling until it reaches one of two boundary values, which are based on predefined decision error probabilities. For the case of simple hypotheses, the SPRT has been proven to be the most efficient test procedure, that is, for given error rates no sequential test requires less observations on average (Wald & Wolfowitz, 1948). The SPRT has been applied to common test scenarios such as *t* tests (see Schnuerch & Erdfelder, 2020) and it is straightforward to apply it to RRM analysis, as well (Schnuerch, Erdfelder, & Heck, 2020).

A potential limitation of the SPRT is that it is a so-called *nontruncated* sequential procedure. That is, there is no definite upper sample size at or before which the test will reach a decision. Curtailed sampling, on the other hand, is a truncated procedure because a lower and upper bound for the sample size are known in advance. Therefore, potential costs and required resources are easier to calculate, which makes curtailed sampling more convenient to plan beforehand.

Moreover, curtailed sampling studies are straightforward to conduct. During sampling, one only has to count observed responses, whereas other sequential designs often require complex or tedious computations. And finally, as mentioned before, curtailed sampling enables simple, unbiased subsequent estimation of the unknown prevalence. Although estimates following a sequential test stopping early will be less precise than for fixed-sample procedures with larger samples, the estimator for the curtailed test presented herein is unbiased. Thus, curtailed sampling constitutes a compromise between the advantages of sequential tests (i.e., efficiency) and those of classical analysis (i.e., easy to plan and unbiased estimation).

In conclusion, curtailed sampling is a relatively easy to implement and practical tool for enhancing the efficiency of surveys applying RRMs. By reducing costs, it makes RRM applications more feasible for studies in which the approach usually would have been prevented by its excessive costs. Therefore, combining

Table 2
Conventional and Subsequent/Sequential Estimation of Doping Prevalence

	Conventional estimation*			Subsequent/sequential estimation		
Sample	*N*	Estimate	CI	*N*	Estimate	CI
WCA	1,203	43.6%	39.4, 47.9	262	33.2%	24.7, 42.4
PAG	965	57.1%	52.4, 61.8	199	51.5%	41.5, 62.5

Note. N = sample size, a random variable in case of sequential estimation: current sample size, when bound $c_s = 102$ "yes" responses were reached; CI = 95% confidence interval (Clopper-Pearson intervals for the subsequent estimates); WCA = 13th International Association of Athletics Federations World Championships in Athletics in Daegu, South Korea, 2011; PAG = 12th Quadrennial Pan-Arab Games in Doha, Qatar, 2011.
* Estimates and confidence intervals are adopted from Ulrich et al. (2018).

curtailed sampling with RRMs provides more valid assessment of sensitive attributes for a broader range of research questions.

References

Abernathy, J. R., Greenberg, B. G., & Horvitz, D. G. (1970). Estimates of induced abortion in urban North Carolina. *Demography, 7,* 19–29. http://dx.doi.org/10.2307/2060019

Anderson, S. F. (2019). Misinterpreting *p*: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000248

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). shiny: Web application framework for r (R package version 1.5.0) [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=shiny

Chaudhuri, A., & Christofides, T. C. (2013). *Indirect questioning in sample surveys*. Berlin, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-36276-7

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

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26,* 404–413. http://dx.doi.org/10.1093/biomet/26.4.404

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. http://dx.doi.org/10.1177/0956797613504966

Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., & Ulrich, R. (2013). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy, 33,* 44–50. http://dx.doi.org/10.1002/phar.1166

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the Randomized Response Technique. *Human Performance, 16,* 81–106. http://dx.doi.org/10.1207/S15327043HUP1601_4

Fox, J. A. (2016). *Randomized response and related methods: Surveying sensitive data* (2nd ed.). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781506300122

Gingerich, D. W. (2010). Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys. *Political Analysis, 18,* 349–380. http://dx.doi.org/10.1093/pan/mpq010

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64,* 520–539. http://dx.doi.org/10.1080/01621459.1969.10500991

Haldane, J. B. S. (1945). A labour-saving method of sampling. *Nature, 155,* 49–50. http://dx.doi.org/10.1038/155049b0

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

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A stochastic lie detector versus the crosswise model. *Behavior Research Methods, 48,* 1032–1046. http://dx.doi.org/10.3758/s13428-015-0628-6

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

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2017). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods, 23,* 226–243. http://dx.doi.org/10.1037/met0000127

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

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 and Research, 33,* 319–348. http://dx.doi.org/10.1177/0049124104268664

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E. J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science, 25,* 1289–1290. http://dx.doi.org/10.1177/0956797614525969

Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2010). Reducing socially desirable responses in epidemiologic surveys: An extension of the randomized-response technique. *Epidemiology, 21,* 379–382. http://dx.doi.org/10.1097/EDE.0b013e3181d61dbc

Nordlund, S., Holme, I., & Tamsfoss, S. (1994). Randomized response estimates for the purchase of smuggled liquor in Norway. *Addiction, 89,* 401–405. http://dx.doi.org/10.1111/j.1360-0443.1994.tb00913.x

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

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1. measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/B978-0-12-590241-0.50006-X

Razafimanahaka, J. H., Jenkins, R. K., Andriafidison, D., Randrianandrianina, F., Rakotomboavonjy, V., Keane, A., & Jones, J. P. (2012). Novel approach for quantifying illegal bushmeat consumption reveals high consumption of protected species in Madagascar. *Oryx, 46,* 584–592. http://dx.doi.org/10.1017/S0030605312000579

Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods and Research*. Advance online publication. http://dx.doi.org/10.1177/0049124120914919

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods, 25,* 206–226. http://dx.doi.org/10.1037/met0000234

Schnuerch, M., Erdfelder, E., & Heck, D. W. (2020). Sequential hypothesis tests for multinomial processing tree models. *Journal of Mathematical Psychology, 95,* 102326. http://dx.doi.org/10.1016/j.jmp.2020.102326

Soeken, K. L., & Damrosch, S. P. (1986). Randomized response technique: Applications to research on rape. *Psychology of Women Quarterly, 10,* 119–126. http://dx.doi.org/10.1111/j.1471-6402.1986.tb00740.x

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511819322

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

Ulrich, R., Pope, H. G., Cléret, L., Petróczi, A., Nepusz, T., Schaffer, J., . . . Simon, P. (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine, 48,* 211–219. http://dx.doi.org/10.1007/s40279-017-0765-4

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods, 17,* 623–641. http://dx.doi.org/10.1037/a0029314

Wald, A. (1947). *Sequential analysis*. New York, NY: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics, 19,* 326–339. http://dx.doi.org/10.1214/aoms/1177730197

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60,* 63–66. http://dx.doi.org/10.2307/2283137

Wetherill, G. B. (1975). *Sequential methods in statistics* (2nd edition). London, UK: Chapman and Hall Ltd.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika, 73,* 573–581. http://dx.doi.org/10.2307/2336521

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

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods and Research, 42,* 321–353. http://dx.doi.org/10.1177/0049124113500474

World Anti-Doping Agency. (2012). *2011 Laboratory testing figures*. Retrieved from https://www.wada-ama.org/en/resources/laboratories/anti-doping-testing-figures-report

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

Received November 6, 2019
Revision received July 6, 2020
Accepted July 20, 2020 ■

Supplemental Section A

Exact determination of the bounds of a curtailed sampling plan

The proposed algorithm searches for the smallest possible values of N_{max} and c_s such that, when combined, they meet the required properties of the hypothesis test. In other words, it searches for the smallest possible values of N_{max} and c_s for which $P(\text{“}H_0\text{”}|\lambda_0) \geq 1 - \alpha$ and $P(\text{“}H_0\text{”}|\lambda_1) \leq \beta$ holds.

Algorithm 1: Iterative search for N_{max} and c_s

Input:
$N_{low} \longleftarrow$ lower bound of search window for N_{max}
$N_{up} \longleftarrow$ upper bound of search window for N_{max}
$\lambda_0 \longleftarrow$ lower bound of λ indifference zone
$\lambda_1 \longleftarrow$ upper bound of λ indifference zone
$\alpha \longleftarrow$ acceptable type 1 error probability
$\beta \longleftarrow$ acceptable type 2 error probability
begin
  $N = N_{low}$
  while $N \leq N_{up}$ **and** $b > \beta$ **do**
    $c_s = CDF^{-1}(1 - \alpha, N, \lambda_0)$
    $b = CDF(c_s, N, \lambda_1)$
    $N = N + 1$
  if $b > \beta$ **then**
    warning
    “No N_{max} and c_s meeting the defined requirements in the defined range”
  else
    return
    $N - 1 \longrightarrow N_{max}$: maximum sample size before sampling is terminated
    c_s: maximum number of successes before sampling is terminated

Supplemental Section B

Expected sample size for varying RRM design specifications

Figures B1 and B2 depict the espected sample size and mean sample sizes of 10,000 simulations each for varying curtailed sampling plan and design specifications of the UQM and the CWM, respectively.

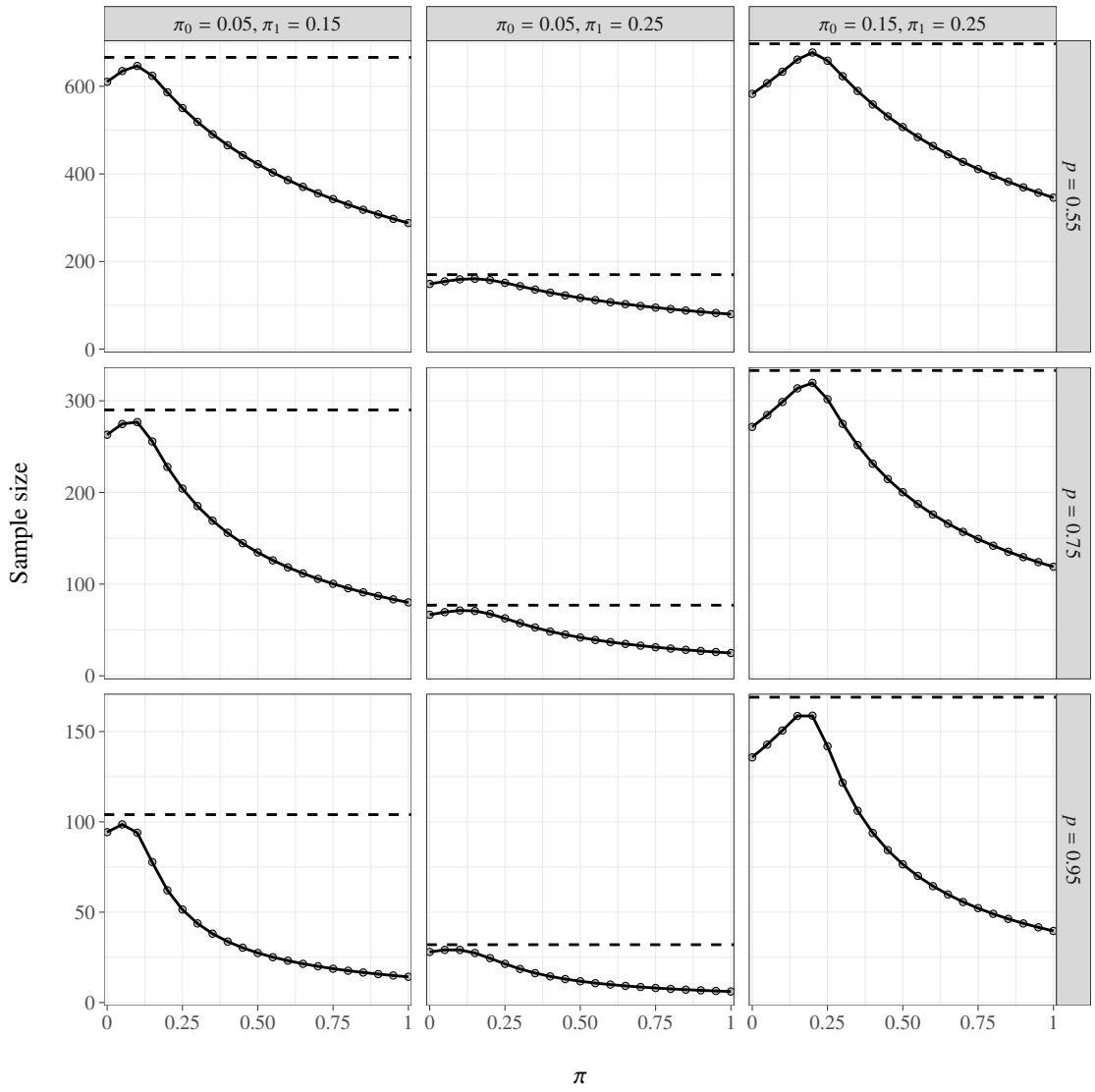


Figure B1. Average sample number curves for varying UQM design specifications. The solid curves depict the expectation of the sample size N when reaching one of the bounds of a curtailed sampling plan as a function of the prevalence π. The dashed lines depict the maximum sample size N_{max}. The gray dots depict the mean sample size $\overline{N}$ of 10,000 samples simulated for each prevalence value π. Each panel corresponds to a specific curtailed sampling plan using different UQM desing specifications. From top to bottom the panels differ with respect to the design parameter p, with p = .55, p = .75 and p = .95, respectively. From left to right the panels differ with respect to the specified hypotheses, with π_0 = .05 and π_1 = .15, π_0 = .05 and π_1 = .25, and π_0 = .15 and π_1 = .25, respectively. In all panels, α = .05 and β = .10 and q = .70.

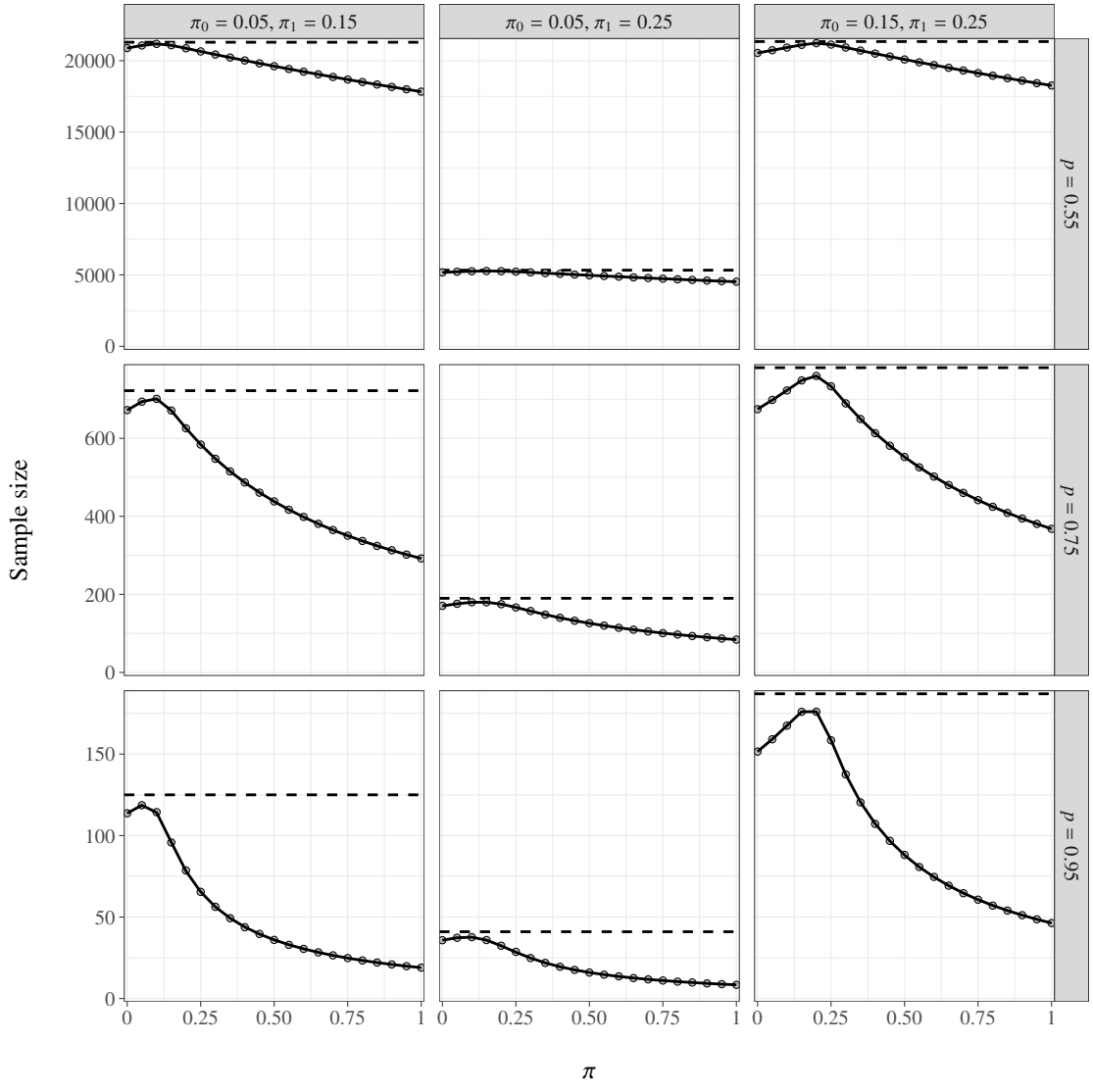


Figure B2. Average sample number curves for varying CWM design specifications. The solid curves depict the expectation of the sample size N when reaching one of the bounds of a curtailed sampling plan as a function of the prevalence π. The dashed lines depict the maximum sample size N_{max}. The gray dots depict the mean sample size $\overline{N}$ of 10,000 samples simulated for each prevalence value π. Each panel corresponds to a specific curtailed sampling plan using different CWM desing specifications. From top to bottom the panels differ with respect to the design parameter q, with $q = .55$, $q = .75$ and $q = .95$, respectively. From left to right the panels differ with respect to the specified hypotheses, with $\pi_0 = .05$ and $\pi_1 = .15$, $\pi_0 = .05$ and $\pi_1 = .25$, and $\pi_0 = .15$ and $\pi_1 = .25$, respectively. In all panels, $\alpha = .05$ and $\beta = .10$.

Supplemental Section C
Subsequent estimation - derivation

Parameter estimation following curtailed sampling must differ from usual prevalence estimation because sampling is stopped only because a certain number of confirmative (dismissive) responses was reached in the last trial. If sampling is always stopped because the bound c_s of confirmative responses was reached, the number of responses N follow a negative binomial distribution with probability mass function (see Wetherill, 1975)

$$P(N|\lambda) = \binom{N-1}{c_s-1} \cdot \lambda^{c_s} \cdot (1-\lambda)^{N-c_s} \tag{1}$$

because the last response was a confirmative response. Thus, the inverse binomial sampling estimator

$$\hat{\lambda} = \frac{c_s - 1}{N-1} \tag{2}$$

with the estimated variance

$$Var(\hat{\lambda}) = \frac{(c_s-1)(N-c_s)}{(N-1)^2(N-2)} \tag{3}$$

is an unbiased estimator of the prevalence in these cases.

If, however, sampling is always stopped because the bound c_f was reached, the number of responses N has probability mass function

$$P(N|\lambda) = \binom{N-1}{N-c_f} \cdot \lambda^{N-c_f} \cdot (1-\lambda)^{c_f} \tag{4}$$

because the last response was a dismissive response. Thus, the inverse binomial sampling estimator can be adjusted to

$$\hat{\lambda} = \frac{N - c_f}{N-1} \tag{5}$$

with the estimated variance

$$Var(\hat{\lambda}) = \frac{(N-c_f)[N-(N-c_f)-1]}{(N-1)^2(N-2)} \tag{6}$$

which is, again, unbiased for these cases.

Both estimators can be combined to cover all possible outcomes of the curtailed sampling procedure and their joint expectation is

$$\begin{aligned} E(\hat{\lambda}|\lambda) = &\sum_{N=c_s}^{N_{max}} \frac{c_s-1}{N-1} \cdot \binom{N-1}{c_s-1} \cdot \lambda^{c_s} \cdot (1-\lambda)^{N-c_s} \\ &+ \sum_{N=c_f}^{N_{max}} \frac{N-c_f}{N-1} \cdot \binom{N-1}{N-c_f} \cdot \lambda^{N-c_f} \cdot (1-\lambda)^{c_f}. \end{aligned} \tag{7}$$

It can be shown that $\hat{\lambda}$ is an unbiased estimator of λ (Girshick, Mosteller, & Savage, 1946). From this follows the joint variance

$$\begin{aligned} Var(\hat{\lambda}|\lambda) = &\sum_{N=c_s}^{N_{max}} \left(\frac{c_s-1}{N-1}\right)^2 \cdot \binom{N-1}{c_s-1} \cdot \lambda^{c_s} \cdot (1-\lambda)^{N-c_s} \\ &+ \sum_{N=c_f}^{N_{max}} \left(\frac{N-c_f}{N-1}\right)^2 \cdot \binom{N-1}{N-c_f} \cdot \lambda^{N-c_f} \cdot (1-\lambda)^{c_f} \\ &- E(\hat{\lambda}|\lambda)^2. \end{aligned} \tag{8}$$

References

Girshick, M. A., Mosteller, F., & Savage, L. J. (1946). Unbiased estimates for certain binomial sampling problems with applications. *The Annals of Mathematical Statistics*, *17*, 13–23. doi: 10.1214/aoms/1177731018

Wetherill, G. B. (1975). *Sequential methods in statistics* (2nd ed.; M. S. Bartlett & D. R. Cox, Eds.). London: Chapman and Hall Ltd.

C Unpublished study report: *The Influence of the Randomization Probability on the Perceived Privacy Protection in Randomized Response Techniques*

The Influence of the Randomization Probability on the Perceived Privacy Protection in Randomized Response Techniques

A Comparison of the Unrelated Question Model, Crosswise Model and Forced Response Model

Fabiola Reiber & Martin Schnuerch

March 29, 2021

The randomized response technique (RRT) is a survey tool, which was developed for the assessment of prevalences of sensitive characteristics (Warner, 1965). The validity of such assessments can be impaired by response biases such as the social desirability bias (Paulhus, 1991). The rationale of the RRT is that such biases can be counteracted by facilitating honest responding through anonymity protection. Specifically, a random element in the questioning design ecrypts individual responses such that they are not conclusive of the respondents' status. Nevertheless, inference on the aggregate level can be drawn given a sufficiently large sample.

Several ways to implement the randomization in the questioning design have been proposed (for overviews see, e.g., Christofides, 2013; Fox, 2016). Three commonly used RRT variants are the forced response model (FRM; Boruch, 1971), the unrelated question model (UQM; Greenberg et al., 1969) and the crosswise model (CWM; Yu, Tian, & Tang, 1991).

To illustrate the mechanism of these three techniques, imagine conducting a study on the prevalence of drug abuse. The straightforward direct questioning approach would be to administer the question "Did you ever take illicit drugs?" among a survey sample. However, this is a sensitive question, which might elicit self-protecting responses. The urge to give a self-protecting response can be decreased by increasing the sense of anonymity protection using one of the three named RRTs.

First, if the FRM is applied, the survey respondents are instructed to respond to the sensitive question only conditional on the outcome of a randomization device, such as a die. For example, they are instructed to respond honestly to the question on drug abuse if the die comes up one, two, or three. Else, they are instructed to give a *forced response*, which is not related to drug abuse, such as: say "yes" if the die comes up four or five and say "no" if it comes up six. Importantly, the outcome of the randomization is covert and only known to the respondent him- or herself. Therefore, a response is not conclusive of the responent's status, because it can either be a response to the sensitive question or a forced response. Nevertheless, it is possible to estimate the prevalence of drug abuse given a sufficiently large sample, because the probabilities underlying the randomization are known. The probability of a "yes"-response is

$$\lambda_{\text{FRM}} = p \cdot \pi + (1 - p) \cdot q \tag{1}$$

with randomization probability $p = 1/2$ to respond to the question on drug abuse (i.e., the die coming up one, two, or three), the conditional probability $q = 2/3$ to say "yes" given the response is a forced response (i.e., the die coming up four or five, out of four, five, and six) and the prevalence π of interest (i.e., drug abuse). The probability λ_{FRM} can be estimated from the proportion of "yes"-responses in the sample and the equation can be rearranged for the prevalence

$$\hat{\pi}_{\text{FRM}} = \frac{\hat{\lambda}_{\text{FRM}} - (1 - p) \cdot q}{p}. \tag{2}$$

Thus, although the individual status remains unknown the prevalence of drug abuse can be estimated.

Second and similarly, if the UQM is applied, the instruction to respond to the sensitive question on drug abuse is also conditional on the outcome of a randomization device. However, the alternative is not a forced response but an honest response to a second *unrelated question*, such as, "Is your birthday in the first 20 days of a month?". Specifically, the instuction could be to respond honestly to the question on drug abuse, if the die comes up one, two, or three and respond honestly to the birthday question if it comes up four, five, or six. Because the randomization is, again, covert it is not clear, whether a specific response is related to the question on drug abuse. Therefore, the individual respondent's status remains hidden. Like in the FRM the prevalence can be estimated from a sufficiently large sample, because the probabilities underlying the randomization are known. The probability of a "yes"-response is, indeed, mathematically equivalent to the FRM

$$\lambda_{\text{UQM}} = p \cdot \pi + (1 - p) \cdot q \tag{3}$$

with the distinction that $q \approx 2/3$ is the probability to respond "yes" to the unrelated birthday question. Therefore, the prevalence estimate can be computed using the same equation in the UQM as in the FRM,

$$\hat{\pi}_{\text{UQM}} = \frac{\hat{\lambda}_{\text{UQM}} - (1 - p) \cdot q}{p}. \tag{4}$$

Third, if the CWM is applied, the sensitive question "Did you ever take illicit drugs?" is also paired with an unrelated question, such as "Is your birthday in the first 20 days of a month?". However, respondents are not instructed to respond to either question, but to give a combined (*crosswise*) response. As such, they are instructed to respond "A" if their response to both questions is the same (i.e., "yes" to both or "no" to both) and "B" if their response to both questions differs (i.e., "yes" to one and "no" to the other). As long as the response to the unrelated question is unknown, the status on the sensitive attribute remains unknown, as well. For example, if a respondent answers "A", that could either mean that he or she has taken illicit drugs and his or her birthday is in the first 20 days of a month or that neither is true. In the CWM, the probability of an "A"-response is

$$\lambda_{\text{CWM}} = p \cdot \pi + (1 - p) \cdot (1 - \pi) \tag{5}$$

with randomization probability $p \approx 2/3$ to respond "yes" to the unrelated birthday question. Thus, the prevalence of drug abuse can be estimated as

$$\hat{\pi}_{\text{CWM}} = \frac{\hat{\lambda}_{\text{CWM}} - (1 - p)}{2p - 1} \tag{6}$$

from a sufficiently large sample.

In summary, these three RRT variants have in common that some sort of randomization leads to individual anonymity protection. Theoretically, the anonymity of respondents is protected irrespective of the size of the randomization probability p. Specifically, the status of a respondent cannot be inferred with certainty as long as $p \neq 0, 1$. However, it is possible that the size of this randomization probability p influences the *perceived* anonymity protection because the conditional probability of being a carrier of the sensitive attribute given a certain response depends on p. Importantly, it is the perceived anonymity protection that influences the validity of prevalence estimates. Indeed, it is a crucial mechanism of the RRT: Only if the randomization increases the perceived anonymity protection, participants are more inclined to answer honestly.

From a rational point of view the perceived anonymity protection should depend on the the informativity of a response with respect to the sensitive attribute. This differs as a function of p for RRTs, as can be seen in the odds-ratio (OR) of being a carrier given a "yes" ("A")- and given a "no"("B")-response

$$OR = \frac{\left(\dfrac{P(\text{carrier}|\{\text{yes,A}\})}{P(\text{non-carrier}|\{\text{yes,A}\})} \right)}{\left(\dfrac{P(\text{carrier}|\{\text{no,B}\})}{P(\text{non-carrier}|\{\text{no,B}\})} \right)} \tag{7}$$

with

$$P(\text{carrier}|\text{yes}) = \frac{[p + (1 - p) \cdot q] \cdot \pi}{p \cdot \pi + (1 - p) \cdot q} \tag{8}$$

and

$$P(\text{carrier}|\text{no}) = \frac{(1-p)\cdot(1-p)\cdot\pi}{1-(p\cdot\pi+(1-p)\cdot q)} \tag{9}$$

in the UQM and FRM, and

$$P(\text{carrier}|\text{A}) = \frac{p\cdot\pi}{p\cdot\pi+(1-p)\cdot(1-\pi)} \tag{10}$$

and

$$P(\text{carrier}|\text{B}) = \frac{(1-p)\cdot\pi}{\pi\cdot(1-p)+(1-\pi)\cdot p} \tag{11}$$

in the CWM.

Figure 1 depicts the absolute log OR as a function of p for the UQM (black curve), the FRM (black curve) and the CWM (gray curve). The UQM and FRM are subsumed in the same curve because they are mathematically equivalent. For the UQM and FRM the linetype differs with respect to different values of q. Rationally, a lower absolute log OR should correspond with higher perceived anonymity protection. Clearly, the influence of p on the predicted (i.e., rational) perceived anonymity protection is not equivalent in all models. Moreover, it is not to be expected that respondents have such a rational representation of anonymity protection.

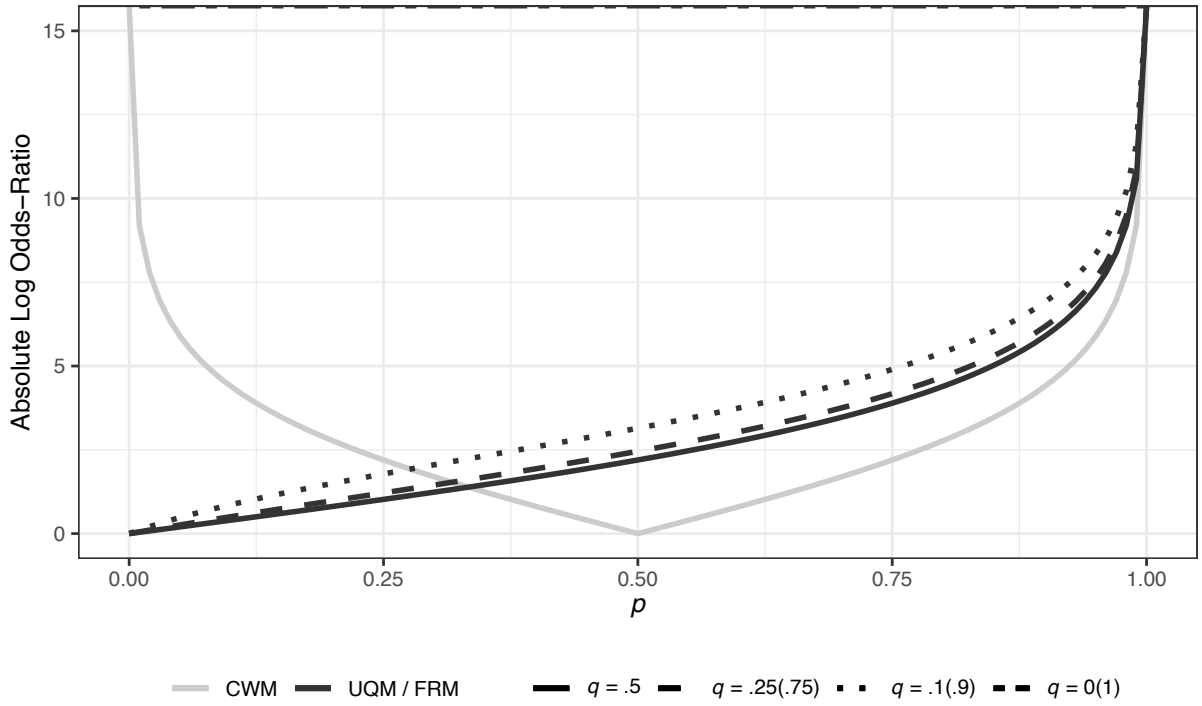


Figure 1
Influence of p on the informative content of responses. The true prevalence is fixed to $\pi = .3$.

The actual influence of p on the perceived anonymity protection of survey respondents is not known. This is, however, an important characteristic, as the design of RRT studies depends heavily on the choice of this parameter. One goal when choosing p is to maximize the perceived anonymity protection. However, the size of p has another important impact: The more randomness is induced through p, the higher the sampling variance becomes, making estimates less precise. Therefore, in terms of sampling efficiency, it is desirable to keep the randomness at a minimum. These two requirements drive p in opposite directions and it is therefore crucial to make a considerate choice on p. A prerequisite for such a choice is knowing the actual influence p has on the perceived anonymity protection.

1 Objective

Thus, the aim of this study was to empirically investigate the influence of the randomization probability p on the perceived anonymity protection and, as a consequence, the (validity of) prevalence estimates.

We investigated this influence within as well as across different RRTs.

2 Methods

Design. To that end, we conducted a large scale online survey with between subject variation of both the questioning design (UQM, FRM, CWM and direct questioning) and randomization probability p (.75, .95). For reasons of simplicity and because the expected influence is substantially smaller in the rational model we refrained from additionally varying q. Table 1 displays the resulting seven conditions.

The randomization probabiltiy p was implemented using the distribution of birthdates over a month. As the distribution of birthdates is approximately uniform, the probability of the birthdate being in a specific range can be calculated. In the UQM and FRM conditions, respondents were instructed to respond to the sensitive question if their mother's birthday lay within a specific range, which depended on the respective p condition. In the CWM conditions, the unrelated question A was whether the respondents' mother's birthday lay within the respective range. The respective ranges and resulting exact randomization probabilities p_{exact} for the two p conditions are depicted in bold face in Table 2.

Substantive question. To create a realistic scenario we implemented a highly sensitive question, that is, "Have you ever experienced intimate partner violence?". We defined intimate partner violence as any type of physical or sexual violence by an intimate partner, such as spouse or boyfriend/girlfriend, and administered the survey among women in Germany. The prevalence of female victimization of physical and sexual intimate partner violence in Germany was estimated to be 22% in a EU-wide survey on violence against women using direct interview questioning (FRA, 2014). Thus, we expected prevalence estimates of about this size (and higher in the RRT conditions) in our study.

Sample. 2201 female German speaking (native or fluent) participants, of legal age, resident in Germany, with at least intermediate-level education were recruited from the participant panel of the market research institute Respondi. The data of 2028 participants were analyzed after excluding respondents who responded, on average, twice as fast as the median respondent per screen of the survey. We aimed for equal distribution to the age groups "18 to 29", "30 to 39", "40 to 49" and "50 or older" and the education level groups "German 'Realschule' or equivalent", "German 'Abitur' or equivalent", "University degree or higher".

Procedure. After indicating their informed consent to participate in the study, participants were informed about the topic of the study and the applied definition of intimate partner violence. Then, participants were randomly assigned to the seven conditions. To maximize power for the various condition comparisons the distribution wase based on equal weights for the RRT conditions and half for the direct question condition (see Table 1). The sensitive question was presented in the respective questioning design with the respective randomization probability. To make sure that participants understood the RRT instructions, they first answered a training question in the same RRT design for a fictional person based on a vignette. Because the correct answer was known, feedback on the response could be provided before the respondents answered the question on intimate partner violence for themselves. In the next step, the perceived anonymity protection was assessed for the implemented randomization probability p in a follow-up question using a slider. The individual function of p on the perceived anonymity protection was assessed by a *magnitude estimation procedure* with the implemented p as reference. Specifically, respondents were asked to indicate how well they would have felt their anonymity was protected in four alternative scenarios with varying p (see Table 2 for the implementation using date ranges and resulting exact p), compared to the perceived anonymity protection in the implemented scenario, if the perceived anonymity protection had been 100 in the implemented scenario. To make sure that the magnitude estimation procedure was understood, respondents were first presented with a training question. Specifically, they were asked to rate the safety of jogging and skydiving as compared to road racing. Finally, participants were presented with an attention check, which asked them to mark the second lowest option of a selection of the numbers 1 through 5 (participants who responded incorrectly

Table 1
Conditions

p	UQM	FRM	CWM	DQ
0.75	2	2	2	
0.95	2	2	2	
1.00				1

Note. Cell weights for each condition; p: randomization probability; UQM: Unrelated Question Model; FRM: Forced Response Model; CWM: Crosswise Model; DQ: Direct questioning.

Table 2
Probabilities

p	Implementation as days in a month*	p_{exact}
0.05	30 - 31	0.047
0.25	24 - 31	0.244
0.51	16 - 31	0.507
0.75	**9 - 31**	**0.737**
0.95	**3 - 31**	**0.934**

Note. p: target randomization probability; p_{exact}: exact p resulting from the date range; * If the respondent's mother's birthay is within this range, (a) in the UQM and FRM conditions he or she should respond to the sensitive question or (b) in the CWM conditions he or she should respond 'Yes' to the neutral question A; The implemented scenarios are highlighted in bold face. Respondents receive either of the two and then judge the perceived anonymity protection in all four alternative scenarios relative to this anchor.

were excluded) before being provided helpline information for victims of intimate partner violence and being redirected to Respondi.

3 Results

3.1 Completion time

Figure 2 depicts the survey completion time per condition.

3.2 Demographics

Information on age and highest educational achievement of participants per condition is provided in Figure 3 and Table 3, respectively.

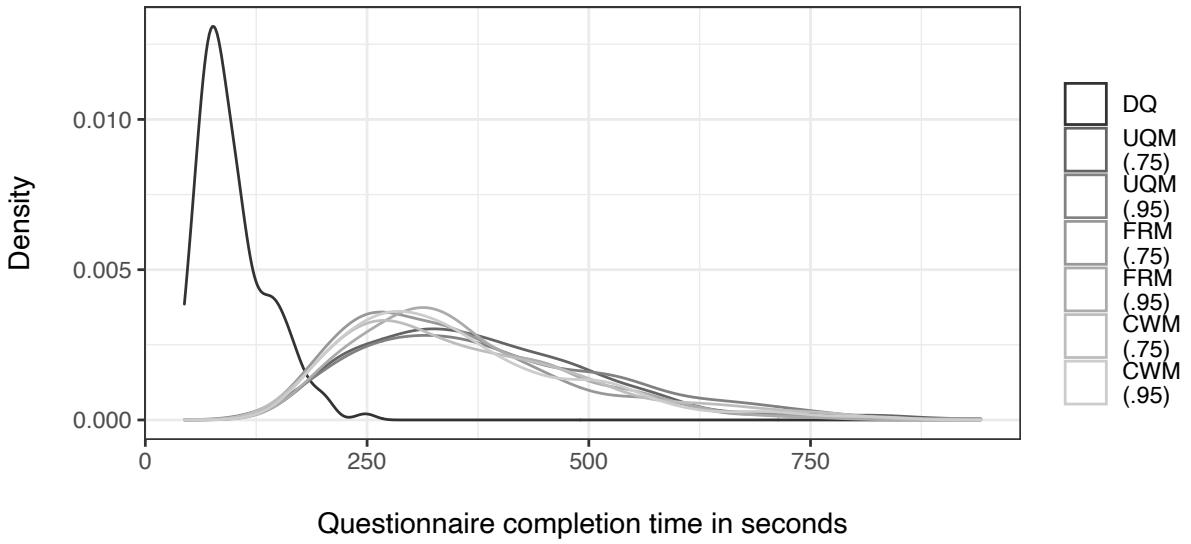


Figure 2
Distribution of completion time per condition

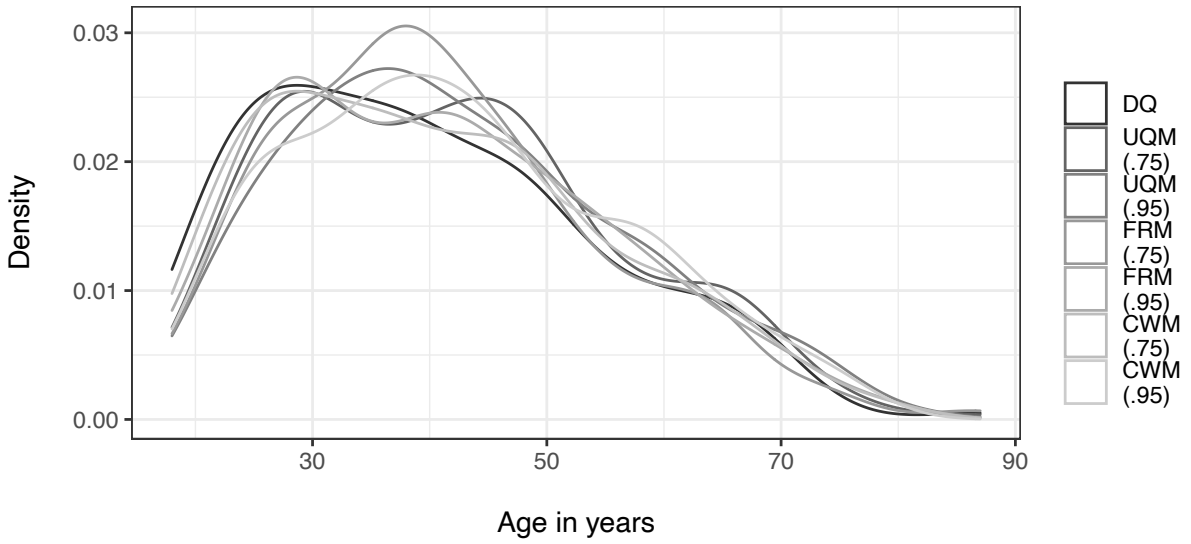


Figure 3
Age distribution per condition

Table 3
Education per condition

Condition	Realschule	Abitur	Hochschulabschluss
DQ	50	71	55
UQM (.75)	102	104	107
UQM (.95)	97	94	100
FRM (.75)	95	101	115
FRM (.95)	103	104	98
CWM (.75)	111	100	100
CWM (.95)	109	101	111

Note. Number of participants in each condition by highest educational achievement; conditions are, direct questioning (DQ), unrelated question model (UQM), forced response model (FRM), crosswise model (CWM) with $p = .75$ (75) or $p = .95$ (95).

Table 4
Proportion correct responses to control question

DQ	UQM (.75)	UQM (.95)	FRM (.75)	FRM (.95)	CWM (.75)	CWM (.95)
-	0.652	0.708	0.907	0.879	0.743	0.748

3.3 RRT

Descriptives on the RRT control question and estimates for intimate partner violence are in Tables 4 and 5 and in Figure 4.

3.4 Anonymity rating and magnitude estimation

Figure 5 depicts the initial perceived anonymity rating per condition. Figure 6 depicts the results of the training question for the magnitude estimation procedure. The proportion of unexpected results, that is, a rating below 100 for jogging and a rating above 100 for skydiving was 0.249 and 0.116, respectively. Figure 7 depicts the results of the magnitude estimation for each randomization probability p and Figure 8 summarizes these results over all probabilities. Table 6 depicts the proportion of ratings exactly equal to 100 for each randomization probability p.

Table 5
Prevalence estimates per condition

Condition	N	Estimate	SE	Lower CI	Upper CI
DQ	176	0.216	0.031	0.155	0.277
UQM (.75)	313	0.204	0.034	0.138	0.270
UQM (.95)	291	0.227	0.026	0.175	0.279
FRM (.75)	311	0.146	0.032	0.083	0.209
FRM (.95)	305	0.143	0.022	0.099	0.186
CWM (.75)	311	0.156	0.053	0.052	0.260
CWM (.95)	321	0.253	0.028	0.198	0.307

Note. Conditions are, direct questioning (DQ), unrelated question model (UQM), forced response model (FRM), crosswise model (CWM) with $p = .75$ or $p = .95$. N: Number of responses, SE: standard error, CI: 95 percent confidence interval.

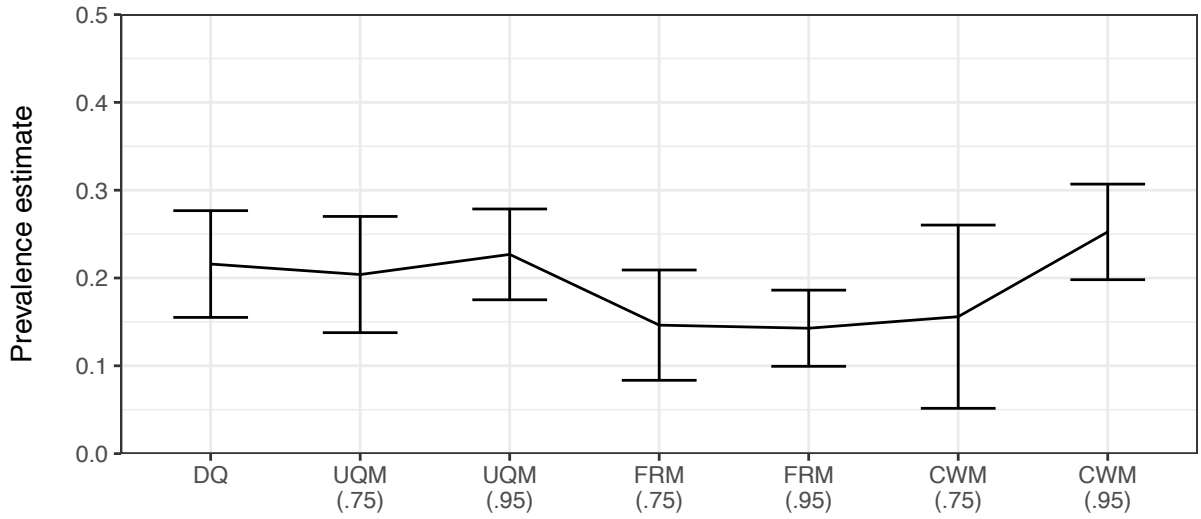


Figure 4
Prevalence estimates across conditions. Errorbars indicate 95 percent confidence intervals.

Table 6
Proportion of ratings equal 100 per level of magnitude estimation.

p	Proportion
0.05	0.407
0.25	0.414
0.55	0.444
0.75	0.453
0.95	0.387

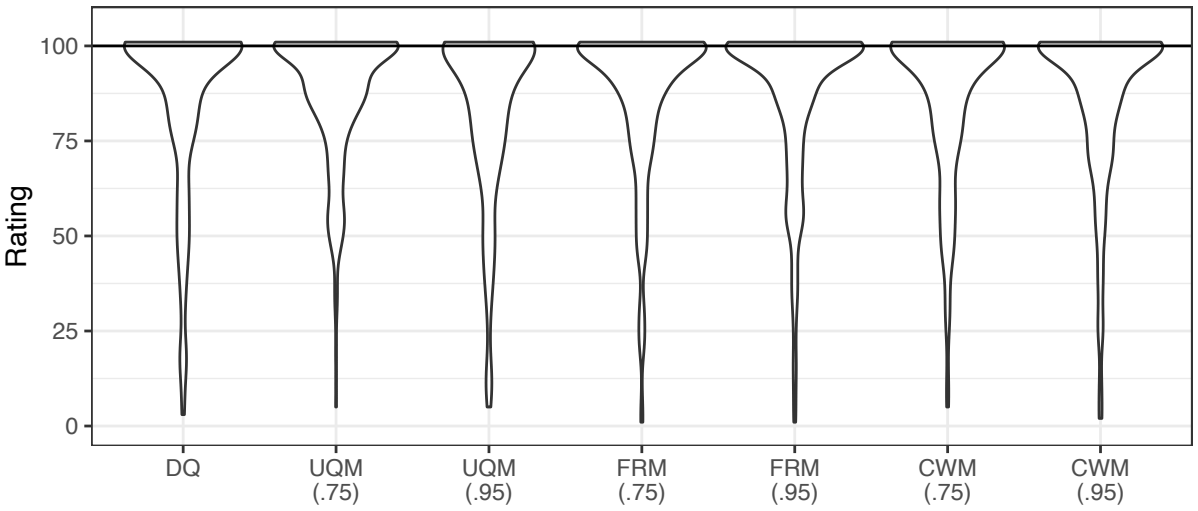


Figure 5
Initial rating per condition

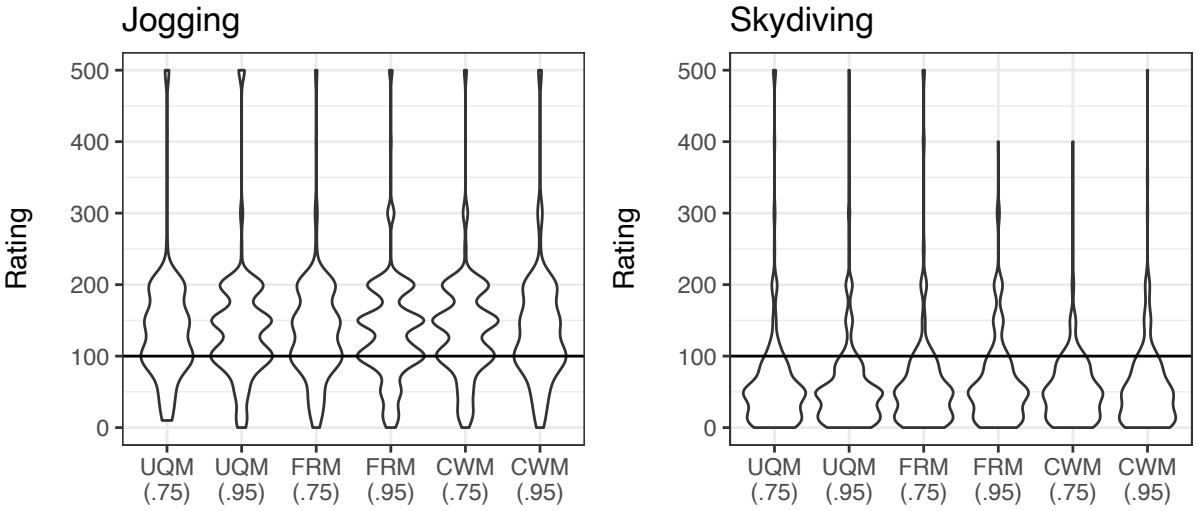


Figure 6
Magnitude estimation for control questions

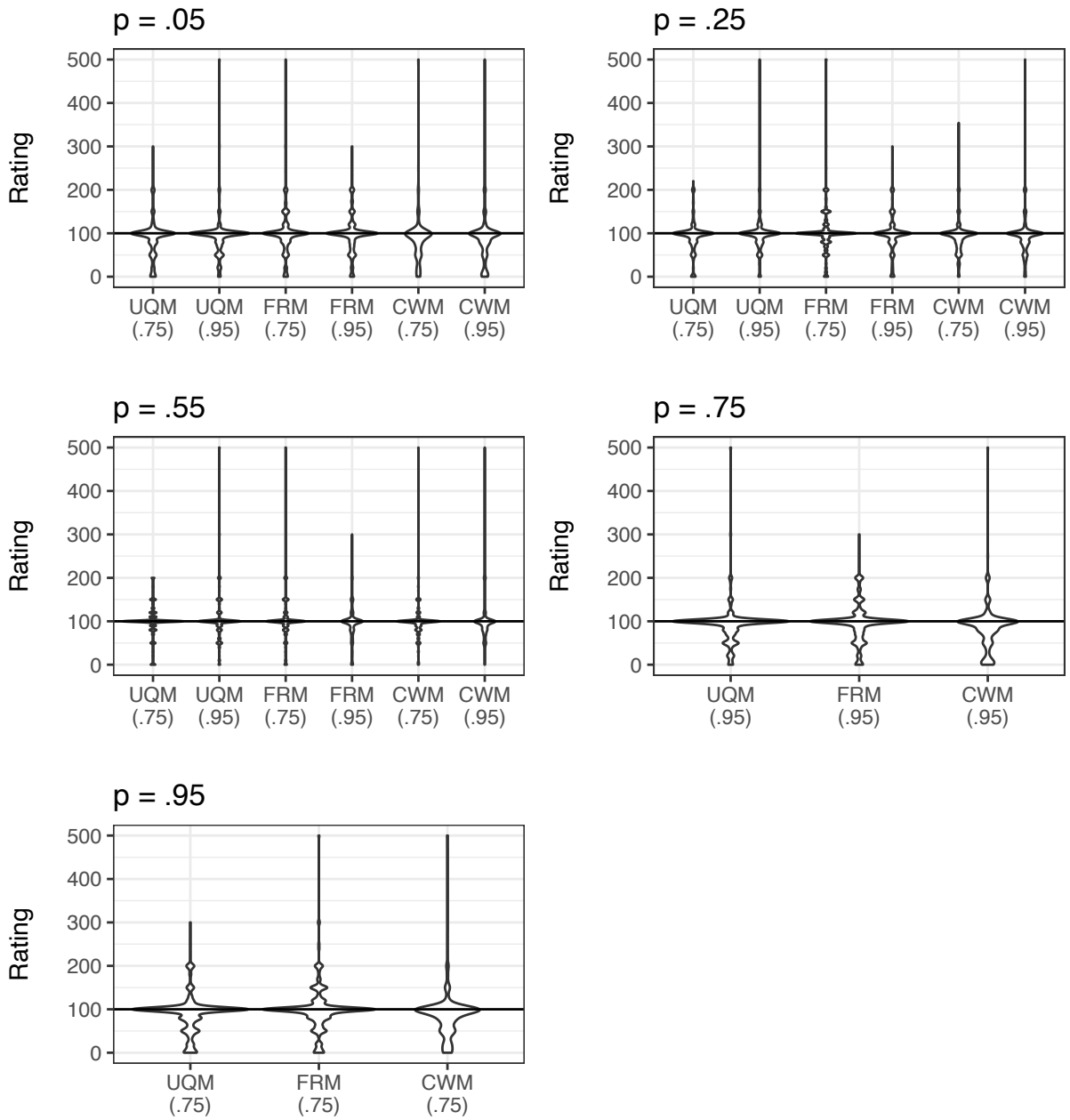


Figure 7
Magnitude estimation for each p. Ratings above 500 are excluded.

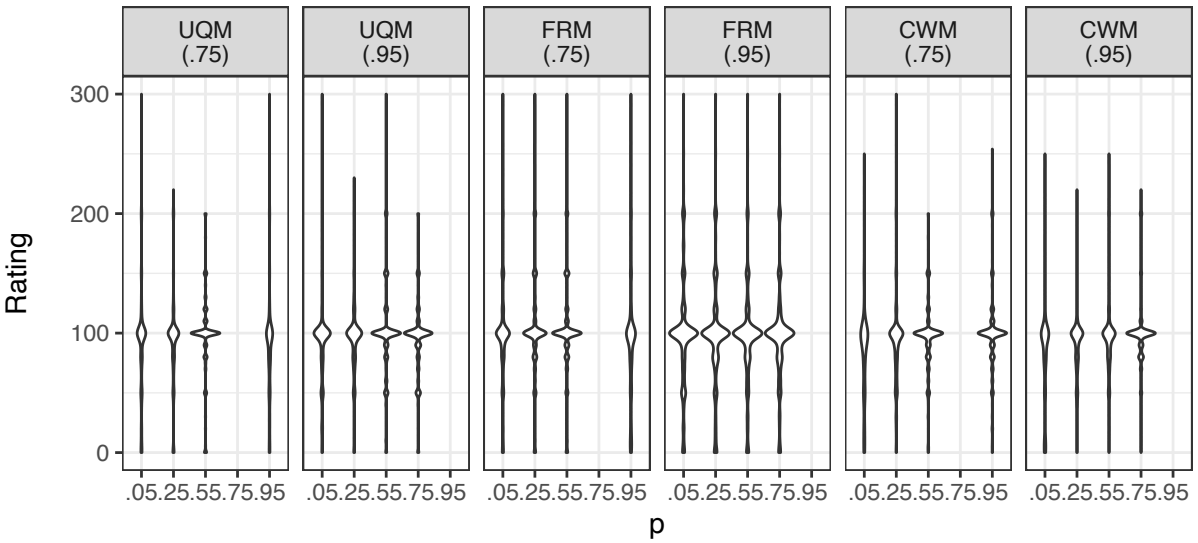


Figure 8
Magnitude estimation for all ps per condition.

This dissertation was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) to the Research Training Group “Statistical Modeling in Psychology” (GRK 2277).