Enhancing the Applicability of Randomized Response Techniques

Fabiola Reiber

Dissertation
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Science (Dr. rer. nat.)
in the DFG Research Training Group "Statistical Modeling in Psychology"
at the University of Tübingen

Thesis Defense:
Month dd, 2021

1st Supervisor:
Prof. Dr. Rolf Ulrich

2nd Supervisor:
Prof. Dr. Edgar Erdfelder

Dean of the Faculty of Science:
Prof. Dr. Thilo Stehle

Thesis Reviewers:
Prof. Dr. Rolf Ulrich
Prof. Dr. Edgar Erdfelder

Defense Committee:
TBA

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 contibutions

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. (2021). *Self-protecting responses in randomized response designs: A survey on intimate partner violence during the COVID-19 pandemic.* Manuscript submitted for publication.

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

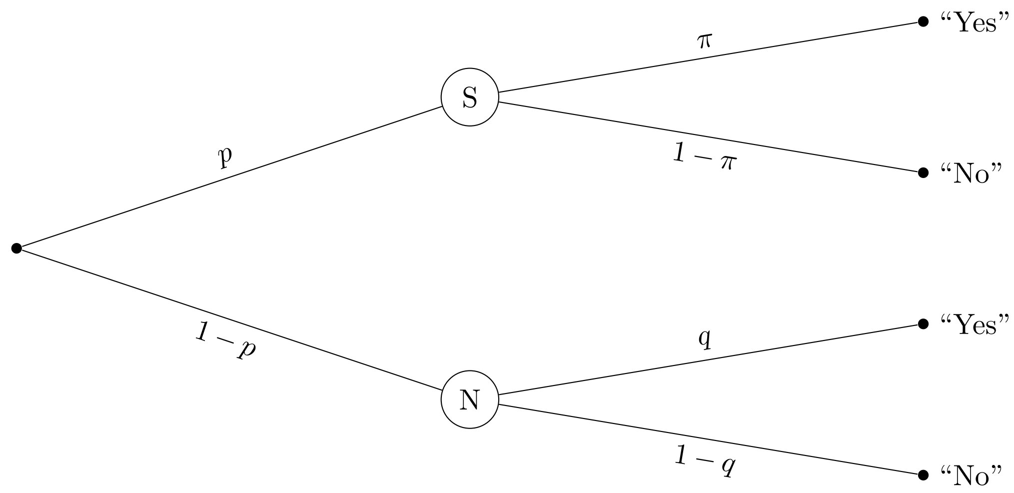

Figure 1: Probability tree of the UQM. Respondents are randomly allocated to respond to the sensitive question S or the neutral question N with probability p and $1-p$, respectively. The probabilities of responding "Yes" and "No" to the neutral question N are q and $1-q$ and the probabilities of responding "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

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

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)

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

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

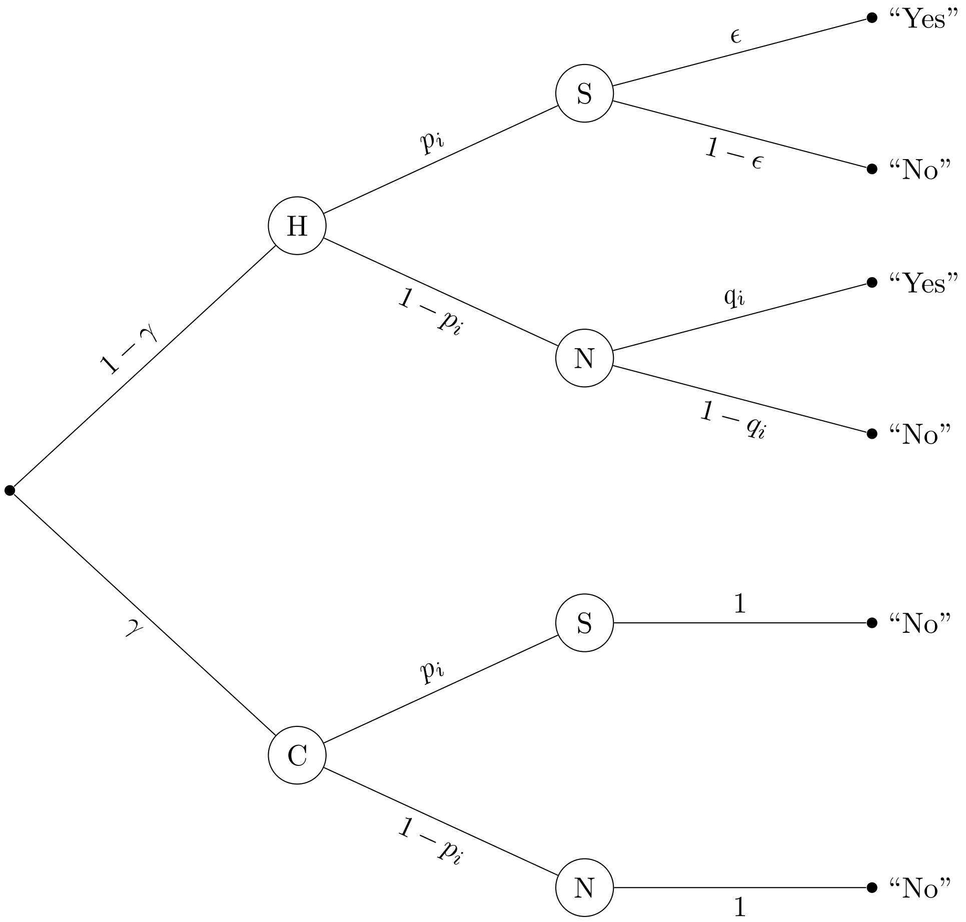

Figure 2: Probability tree of the UQMC. The prevalence of cheaters C is γ and the prevalence of honest participants H is $1-\gamma$. Both types of respondents are allocated to respond to the sensitive question S and the neutral question N with probability p_i and $1-p_i$, respectively. The model assumes 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. (2021). *Self-protecting responses in randomized response designs: A survey on intimate partner violence during the COVID-19 pandemic.* Manuscript submitted for publication.

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:

$$H_0 : \pi \leq \pi_0 = .10$$
$$H_1 : \pi \geq \pi_1 = .15$$

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.

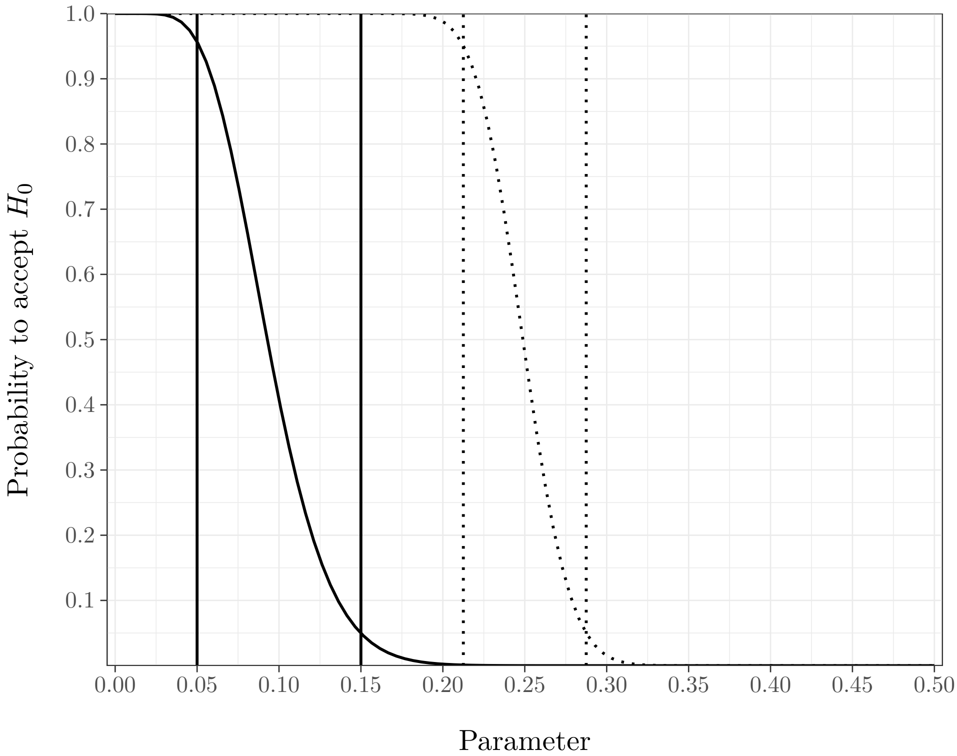

Figure 3: Operating characteristic curve of a curtailed sampling plan. The curves depict the probability of accepting the null hypothesis as a function of the true parameter value for a test of the hypotheses $H_0 : \pi \leq \pi_0 = .10$ vs. $H_1 : \pi \geq \pi_1 = .15$ with $\alpha = \beta = .05$. The 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:

$$H_0\colon \pi \leq \pi_0 = .01$$

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

$$H_1\colon \pi \geq \pi_1 = .15$$

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

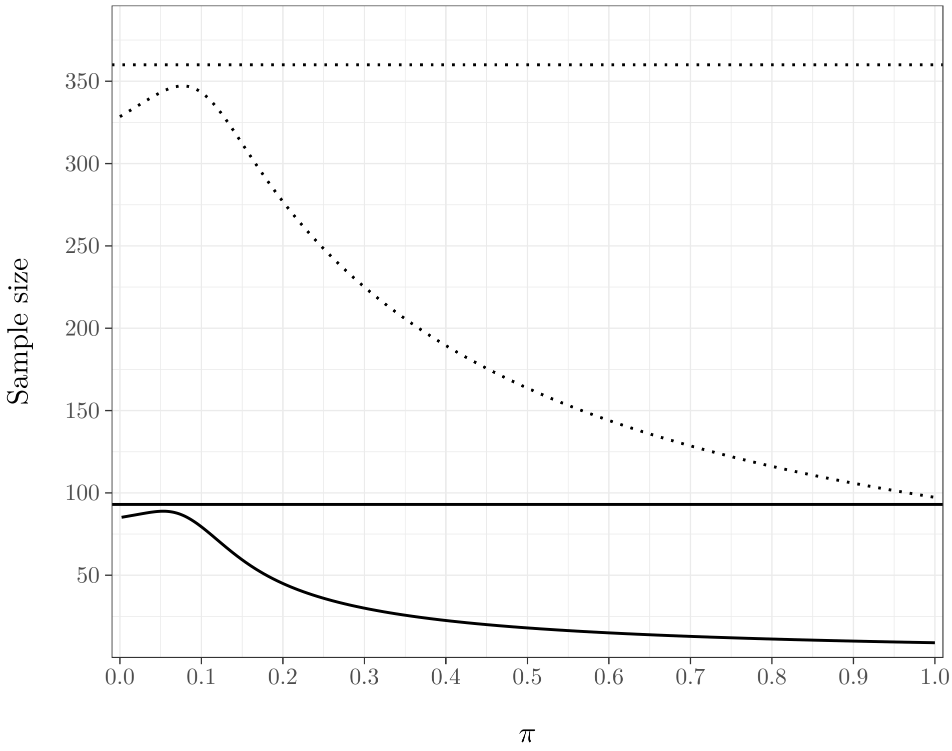

Figure 4: Expected sample size in a curtailed sampling plan. The curves depict the expected sample size in a curtailed sampling plan as a function of the true prevalence π for a test of the hypotheses $H_0 : \pi \leq \pi_0 = .10$ vs. $H_1 : \pi \geq \pi_1 = .15$ with $\alpha = \beta = .05$. The solid 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,

$$H_0\colon \pi \le \pi_0 = .70$$
Voting is not overreported.
$$H_1\colon \pi \ge \pi_1 = .80$$
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 prevalence, 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.