Psychological Methods

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

Fabiola Reiber, Martin Schnuerch, and Rolf Ulrich

Online First Publication, September 10, 2020. http://dx.doi.org/10.1037/met0000353

CITATION

Reiber, F., Schnuerch, M., & Ulrich, R. (2020, September 10). Improving the Efficiency of Surveys With Randomized Response Models: A Sequential Approach Based on Curtailed Sampling. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000353

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2020 American Psychological Association
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)

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

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.

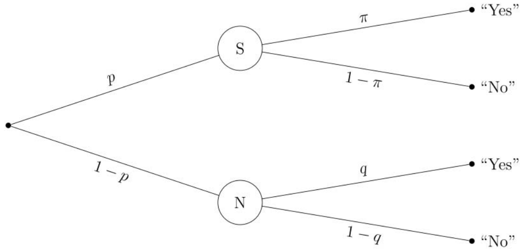

Figure 1. Probability tree of the UQM. The sensitive question S and the neutral question *N* are randomly received by respondents with probability p and $1 - p$, respectively. The probabilities of responding "Yes" and "No" to the neutral question *N* are q and $1 - q$ and the probabilities of responding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

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.

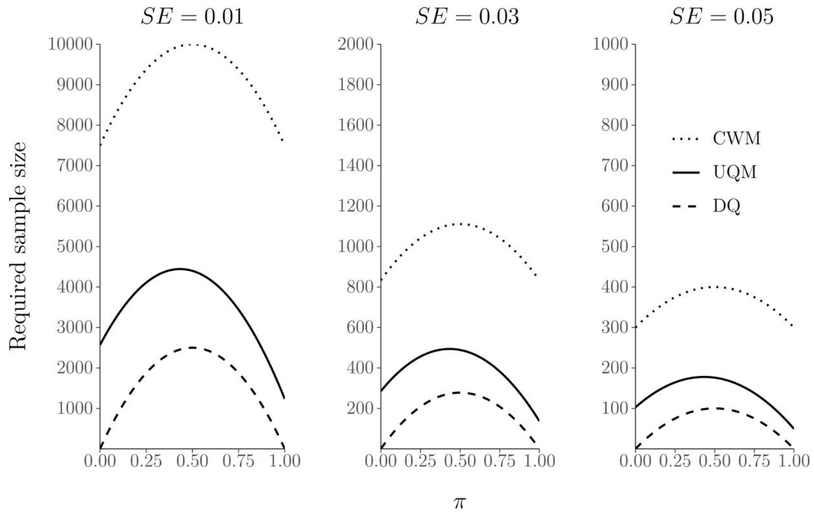

Figure 2. Required sample size depending on questioning type. Depicted is the required sample size as a function of the true prevalence π. The curves within a panel depict the questioning types: Direct question (DQ, dashed), unrelated question model (UQM, solid) and crosswise model (CWM, dotted). The 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.

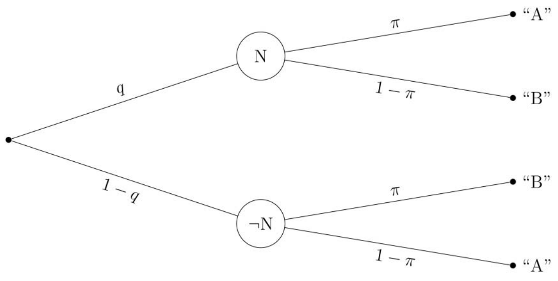

Figure 3. Probability tree of the CWM. Respondents are asked to respond to both questions S and N in one response, "A" or "B." Respondents carry the neutral attribute N with known probability q or do not carry it ¬N with probability $1 - q$. Carriers of the neutral attribute respond "A" with probability 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.

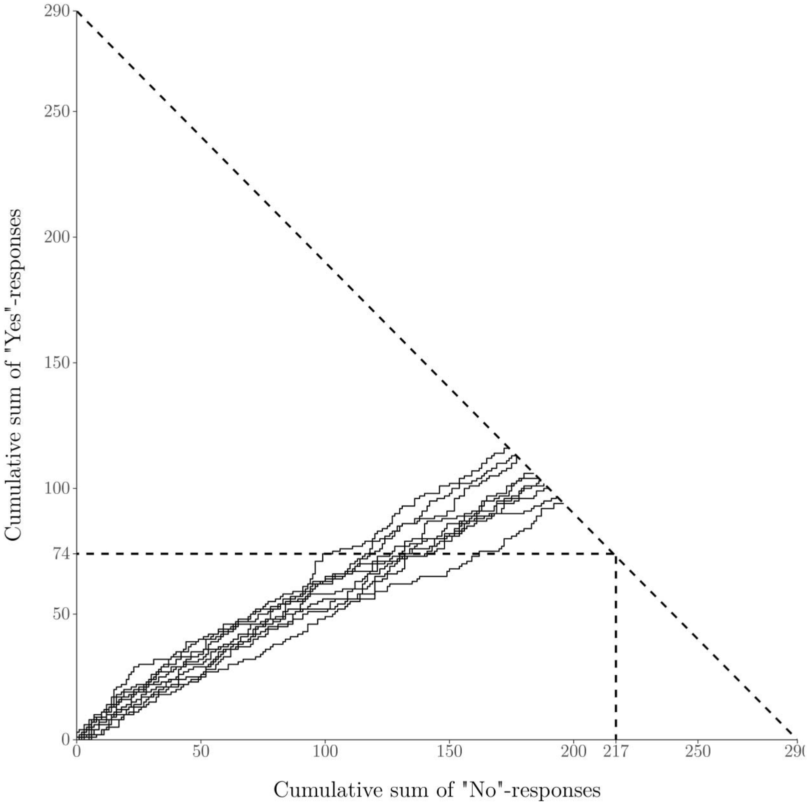

Figure 4. Sampling paths of simulated samples. The 10 samples were simulated as unrelated question model data with design parameters $p = .75$, $q = .70$, and true prevalence $\pi = .25$. The depicted bounds are (a) the maximum number of "Yes" responses $c_s = 74$ (horizontal); (b) the maximum number 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 $\pi = .25$.

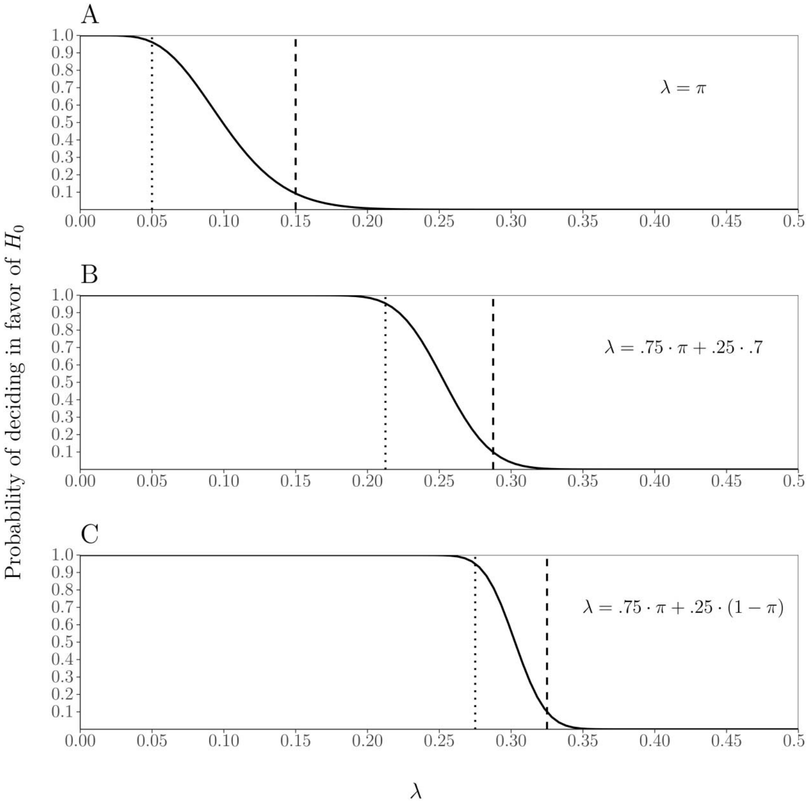

Figure 5. Operating characteristic curve. Depicted is the probability of deciding in favor of H_0 depending on the true probability of a "Yes" response λ in a curtailed sampling plan. All panels refer to the same hypothesis test concerning the prevalence π: $\pi_0 \leq .05$ (dotted line) versus $\pi_1 \geq .15$ (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

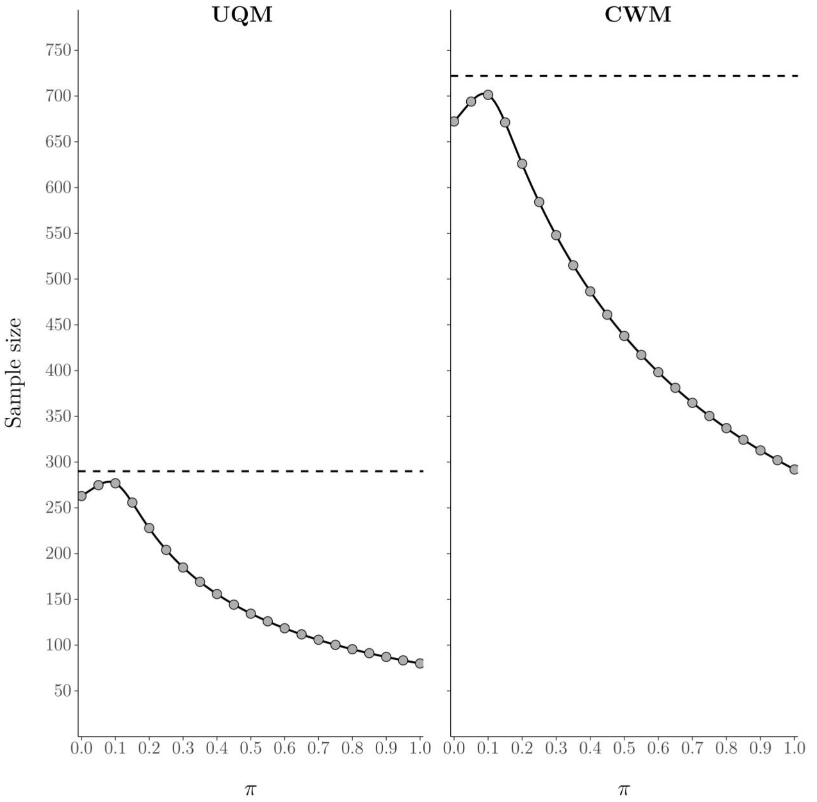

Figure 6. Average sample number curves. The solid curves depict the expectation of the sample size N when reaching one of the bounds of a curtailed sampling plan as a function of the prevalence π. The dashed lines depict the maximum sample size N_{max}. The gray dots depict the mean sample size $\bar{N}$ 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

$$\hat{\lambda} = \frac{N - c_f}{N - 1}. \tag{8}$$

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.

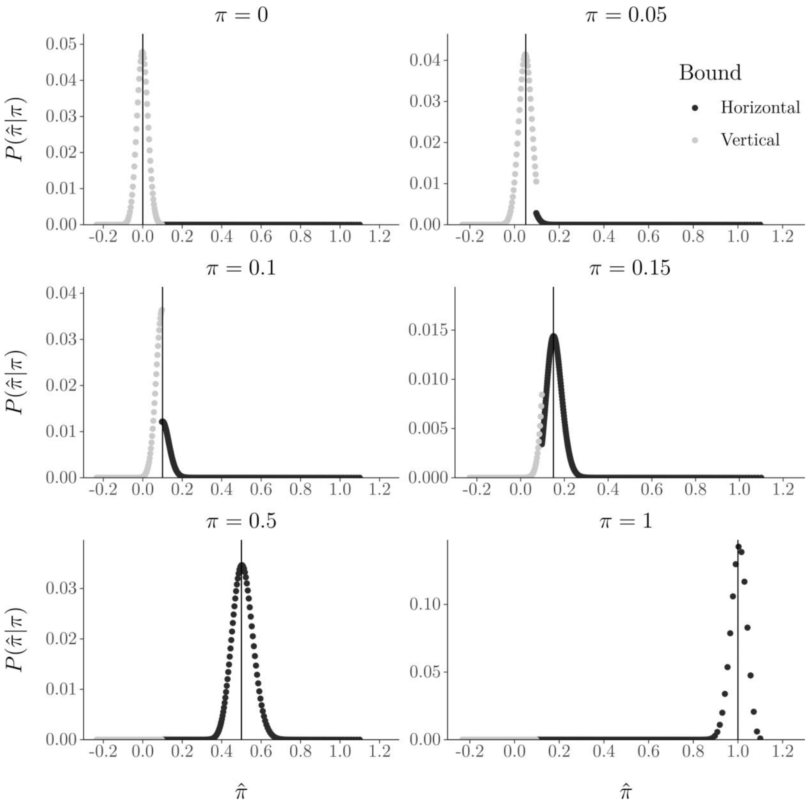

Figure 7. Theoretical sampling distributions of $\hat{\pi}|\pi$. Depicted are the probabilities of obtaining a certain estimate after testing the hypotheses $\pi_0 = .05$ and $\pi_1 = .15$ with $\alpha = .05$ and $\beta = .10$ in a curtailed sampling plan using the UQM with design parameters $p = .75$, $q = .70$. 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

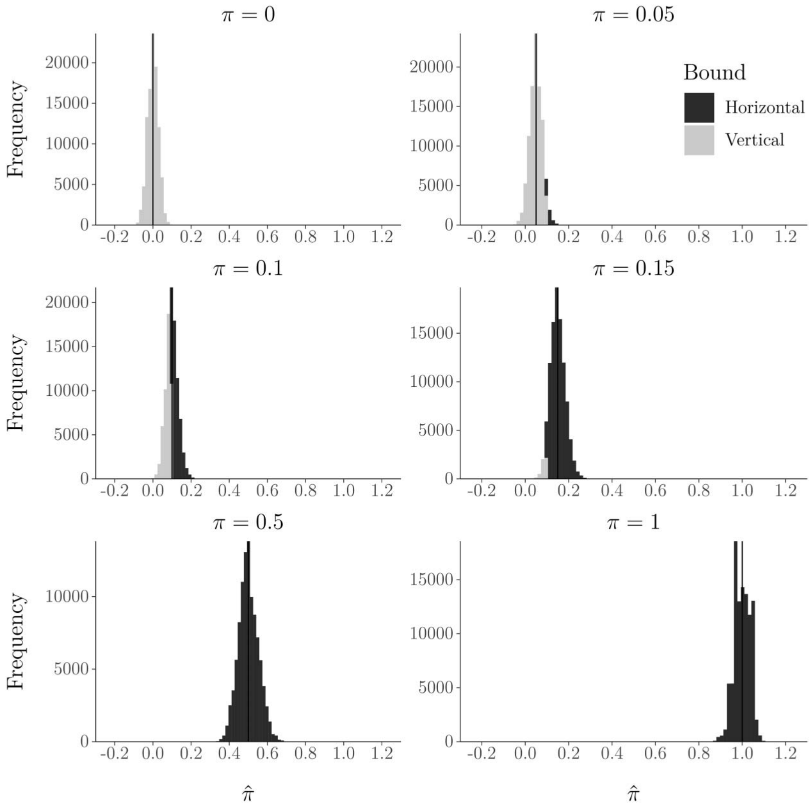

Figure 8. Simulated sampling distributions of $\hat{\pi} \mid \pi$. Depicted are frequency distributions of prevalence estimates $\hat{\pi}$ calculated from simulated samples using the information available in the moment when sampling would have been stopped in a curtailed sampling plan for testing the 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.

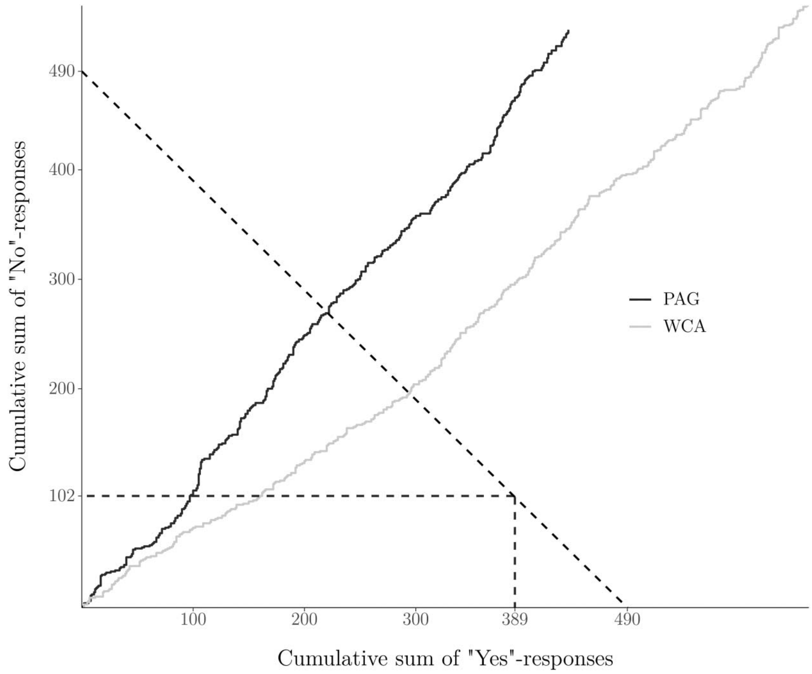

Figure 9. Sampling paths of the two samples in the doping study. The underlying design parameters are p = .50, q = .67. The depicted bounds are c_s = 102 (horizontal), c_f = 388 (vertical), and N_{max} = 490 (diagonal) and are based on the hypotheses π_0 = .02 and π_1 = .10 with α = .05 and β = .10. 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 ■