

RESEARCH ARTICLE

Refinement of the extended crosswise model with a number sequence randomizer: Evidence from three different studies in the UK

Khadiga H. A. Sayed^{1,2}*, Maarten J. L. F. Cruyff¹, Peter G. M. van der Heijden^{1,3}, Andrea Petróczi^{4,5}

1 Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands, **2** Department of Statistics, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt, **3** Department of Social Statistics and Demography, University of Southampton, Southampton, United Kingdom, **4** School of Life Sciences, Pharmacy and Chemistry, Kingston University, London, United Kingdom, **5** Department of Movement Sciences, KU Leuven, Leuven, Belgium

* These authors contributed equally to this work.

* k.h.a.sayed@uu.nl



OPEN ACCESS

Citation: Sayed KHA, Cruyff MJLF, van der Heijden PGM, Petróczi A (2022) Refinement of the extended crosswise model with a number sequence randomizer: Evidence from three different studies in the UK. *PLoS ONE* 17(12): e0279741. <https://doi.org/10.1371/journal.pone.0279741>

Editor: Werner Pitsch, Saarland University, GERMANY

Received: November 30, 2021

Accepted: December 14, 2022

Published: December 30, 2022

Copyright: © 2022 Sayed et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its [Supporting information](#) files. We provided tables with frequency counts in section "Data and methods" of the manuscript which meets PLOS ONE's minimal requirements of data set availability. Using these tables, one can reproduce the results described in our manuscript.

Funding: Partial funding for the collection was provided by Kingston University London Impact Funding, and Hampshire County Council as part of

Abstract

The Extended Crosswise Model (ECWM) is a randomized response model with neutral response categories, relatively simple instructions, and the availability of a goodness-of-fit test. This paper refines this model with a number sequence randomizer that virtually precludes the possibility to give evasive responses. The motivation for developing this model stems from a strategic priority of WADA (World Anti-Doping Agency) to monitor the prevalence of doping use by elite athletes. For this model we derived a maximum likelihood estimator that allows for binary logistic regression analysis. Three studies were conducted on online platforms with a total of over 6,000 respondents; two on controlled substance use and one on compliance with COVID-19 regulations in the UK during the first lockdown. The results of these studies are promising. The goodness-of-fit tests showed little to no evidence for response biases, and the ECWM yielded higher prevalence estimates than direct questions for sensitive questions, and similar ones for non-sensitive questions. Furthermore, the randomizer with the shortest number sequences yielded the smallest response error rates on a control question with known prevalence.

Introduction

Doping in elite sports has become a global challenge that threatens not only the fairness and justice of any sport competition, but also the health and careers of elite athletes [1]. Monitoring anti-doping activities worldwide and providing information on the prevalence of doping among elite athletes is one of the strategic priorities of the World Anti-Doping Agency (WADA). As custodian of the World-Anti-Doping Code 2021, WADA coordinates doping testing and publishes an annual list of prohibited substances and/or methods that defines

a separate project. There was no additional external funding received for this study. The funders had no role in any aspects of the study or publication. The funder provided support in the form of salaries for authors [AP and MC], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: AP chairs and MC is an ad hoc member of WADA's Working Group on Doping Prevalence. WADA supported the research aim but had no influence over the content, nor provided funding for the work in this project. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

doping in competitive sports [2]. Obtaining valid and reliable information on the prevalence of doping in competitive sports is essential to appraise the magnitude of the problem and to evaluate the effectiveness of global, national and regional anti-doping strategies [3]. Methods for the estimation of doping prevalence in elite sports include testing based on urine or blood analyses, the Athlete Biological Passport and survey methods [4]. A promising survey method for asking sensitive questions is using randomized response, because it is designed to eliminate evasive response biases by protecting the privacy of the respondents [5]. A meta-analysis has shown that randomized response designs provide more valid prevalence estimates than conventional direct questions (DQ) when questions are sensitive [6].

In 2011 WADA convened a working group to tackle the issue of doping prevalence and funded two studies for developing and testing a survey method for estimating doping prevalence at two international sport events: the 13th International Association of Athletics Federations World Championships in Athletics (WCA) in Daegu and the 12th Quadrennial Pan-Arab Games (PAG) in Doha [7]. The randomized response technique used in these surveys was the Unrelated Question Model (UQM) [8]. In the UQM, respondents are instructed to answer either a sensitive question or a non-sensitive (unrelated) question, depending on the outcome of a randomizer.

In [7], the UQM was implemented as follows. Respondents were instructed to think of the birthday of a person close to them, and to answer the unrelated question if the birthday fell in the first 10 days of the month, and to otherwise answer the question about doping use. The unrelated question was whether the birthday fell in the first six months of the year or not. The high prevalence estimates of 43.6% at WCA and 57.1% at PAG were unexpected, and raised doubts about the response validity. Several hypotheses about instruction noncompliance were considered, but none of these fully explained the high estimates. To gain insight in the response validity in future applications, a refinement of the methodology was recommended.

The UQM has been employed in several studies to estimate the prevalence of various sensitive topics such as induced abortion, doping and illicit drug use in fitness sports and elite athletes, rape victimization and tax evasion. However, the "Yes" answer of this model is incriminating, and respondents may choose to provide a self-protective "No" response regardless of their true status to avoid being categorized as carriers of the sensitive characteristic [9]. Furthermore, its statistical model is saturated, which makes it impossible to test for the presence of response biases in the data.

There are various indirect question techniques that avoid the use of incriminating responses, such as the Unmatched List technique [10], the Single Sample Count [11, 12], and the Crosswise Model (CWM) [13]. The CWM has gained sizable popularity since its inception in 2008, and thus offers ample empirical evidence for its performance in obtaining information on a range of socially sensitive issues [14–25]. Like the UQM, the CWM employs a sensitive and an unrelated question, but the respondent is instructed to answer both questions simultaneously with "I have TWO 'Yes' or TWO 'No' answers" or "I have ONE 'Yes' answer". Since neither answer has an obvious incriminating connotation, both the urge and the potential for giving evasive responses is expected to be reduced. The statistical model, however, is still saturated and therefore unable to detect response biases. The Extended Crosswise Model (ECWM) extends the CWM by splitting the sample in two sub-samples, each with a different probability of answering "Yes" to the unrelated question. A diagram of the ECWM with two sub-samples is depicted in Fig 1, with π denoting the prevalence of the sensitive characteristic, and p_1 and p_2 the probability of answering "Yes" to the unrelated question in the subsequent sub-samples, for $p_1 = 1 - p_2$. The split in two samples with different randomization probabilities generates the necessary degree of freedom to conduct a goodness-of-fit test. Because of this property,

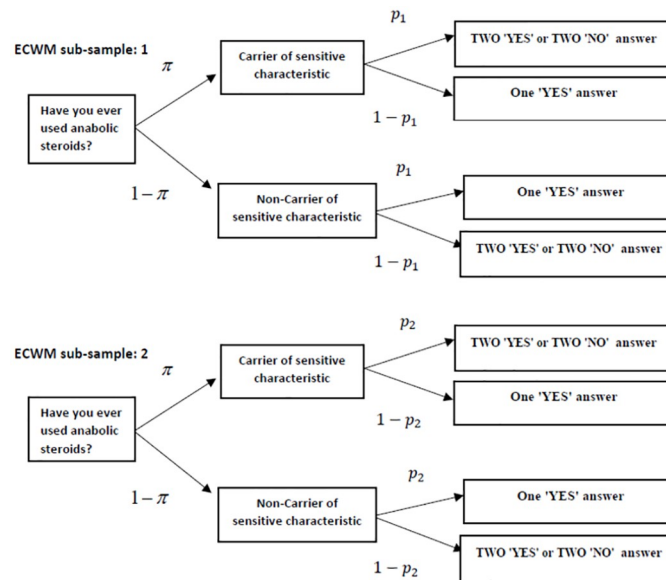


Fig 1. Diagram of the ECWM [26], with π denoting the prevalence of the sensitive characteristic, and p_1 and p_2 , for $p_1 + p_2 = 1$, the probabilities of answering “Yes” to the unrelated question in sub-samples 1 and 2.

<https://doi.org/10.1371/journal.pone.0279741.g001>

and its neutral response categories, the ECWM was chosen. Since its introduction by [26], the ECWM has been applied in studies by [23, 27, 28].

Addressing some of the concerns highlighted in the systematic review on CWM applications [29], we proposed a refinement of the ECWM with a number sequence randomizer rather than the birthday randomizer. A potential pitfall of the birthday randomizer is that birthdays may not be uniformly distributed over the months of the year, which in turn results in potentially biased probabilities to answer “Yes” to the unrelated question. This bias may be aggravated by the fact that in some countries birthdays are often unknown and arbitrarily set to January 1 (for an anecdotal example, see the Washington Post article “In Afghanistan, January 1 is everybody’s birthday” [30]). With the number sequence randomizer, the respondent is simply asked to memorize one number from a sequence of numbers, and the unrelated question is whether the memorized number is present or not in a second sequence of numbers. Since the probability that the memorized number reappears in the second sequence is exactly known, the potential biases of the birthday randomizer are avoided. The number sequence randomizer is discussed in detail in Section “Data and methods”, and for an example from Study II we refer to Fig 2. The ECWM by [26] in combination with our number sequence randomizer addresses these pitfalls.

The idea of splitting the sample in two sub-samples was originally proposed [31] in Cheater Detection Model (CDM) to estimate the prevalence of so-called ‘cheaters’, i.e. respondents who give the non-incriminating “No” answer when an incriminating “Yes” was required. Instead of cheaters we prefer the term ‘self-protective no-sayers’ as introduced by [32], because it provides a more accurate description of this type of respondents. The CDM has been used in studies to estimate the prevalence of doping, and yielded estimates of self-protective no-sayers ranging from 29.9% [33] to 22.9% and 30.6% [34]. The ECWM, however, is not capable of detecting self-protective no-saying. This is not as much because the answer “No” is not used, but because of the fact the incriminating response is not the same in both sub-samples. As can be inferred from Fig 1, carriers of the sensitive characteristic answer “I have TWO ‘Yes’ or

First, memorise ONE number from the list:									
19	51	43	93	46	78	21	67	47	84

↓

Read questions A and B carefully:									
<ul style="list-style-type: none"> • Question A: Did you meet people socially outside your household during the lockdown period between March 23rd and May 13th ? • Question B: Is the number you selected included in the set below? 									
62	59	44	64	93	74	8	26	27	19

↓



Answer – Select one from:	
	
Select this option if you have two 'yes' answers OR if you have two 'no' answers	Select this option if you have only ONE 'yes' answer

Fig 2. Example of the number sequence randomizer as used in Study II.

<https://doi.org/10.1371/journal.pone.0279741.g002>

TWO 'No' answers" with probability p_1 in sub-sample 1, and with probability p_2 in sub-sample 2, so that the self-protective responses are "I have One 'Yes' answer" in sub-sample 1, and "I have TWO 'Yes' or TWO 'No' answers" in sub-sample 2 assuming that $p_1 > p_2$. We call this kind of response behavior 'informed self-protection', because it requires an understanding of the connotation between the responses and the sensitive characteristic. As shown in Section "Statistical model", the presence of informed self-protection cannot be detected because its effects on the observed response probabilities 'cancel out'; as the fraction of misreported answers "I have TWO 'Yes' or TWO 'No' answers" in one sub-sample is expected to equal the fraction of misreported answers "I have ONE 'Yes' answer" in the other sub-sample. This does not render the goodness-of-fit test useless because it will be able to detect other response biases, as long as their effects on the observed response probabilities do not cancel out.

The aim of this study is to evaluate the performance of the ECWM in combination with the number sequence randomizer in three subsequent studies with socially sensitive behavioral choices to each; Study I and III concentrate on controlled substance use and Study II focuses on non-compliance with COVID-19 regulations. The performance of the ECWM is evaluated in several ways. Firstly, in line with the "more-is-better" criterion [6, 29, 35], we expect the ECWM to yield the same prevalence estimate as DQ when the question is non-sensitive, and higher estimates as the sensitivity of the question increases in the sense that the more sensitive the question, the more underreporting will occur when respondents are asked to answer the question directly. Secondly, we expect the goodness-of-fit tests for the ECWM questions to be non-significant. Additionally, we included a non-sensitive control question with a known prevalence of 100% to determine the effect of the length of the number sequences of our randomizer on the response error rate. Furthermore, by conducting logistic regressions analyses with the time spent to complete the survey and the perceived difficulty of the ECWM answer format we hope to gain some insight in the reasons for these response errors. Finally, we manipulated the format of the sensitive questions to investigate their effects on the prevalence estimates. In Study I we compare questions versus statements, and in Study II a factual versus judgmental questions. We hypothesized that the non-incriminating wording of the question in the factual format would encourage respondents to admit non-compliance with COVID-19 regulations, and thus yield higher prevalence estimates than the incriminating wording of the questions in the judgmental format.

The paper is structured as follows. In section “Statistical model”, we derive a maximum likelihood estimator for ECWM that also allows for the formulation of a regression model and for conducting the goodness-of-fit test. Section “Data and methods” provides information about participants, questionnaires, the procedure and measures of the three studies. Section “Results” evaluates the performance of the ECWM. Finally, the concluding remarks, limitations and directions for future research are presented in Sections “Discussion” and “Conclusion”.

Statistical model

This section reviews the moment estimator of [26], and uses it to demonstrate that the ECWM is not capable of detecting informed cheating. Next a binary logistic regression model is derived, which is used to conduct the goodness-of-fit test.

Moment estimation

A moment estimator is derived for the ECWM by [26] through fitting a separate model to each sub-sample separately. Let $\pi_{y|s}^*$ be the conditional probability of observing response y , for $y \in \{1 \equiv \text{“I have ONE ‘Yes’ answer”}, 2 \equiv \text{“I have TWO ‘Yes’ or TWO ‘No’ answers”}\}$ given membership of sub-sample s , for $s \in \{1, 2\}$. Let π_s denote the prevalence of the sensitive characteristic in sub-sample s , and p_s the probability of a “Yes” answer to the unrelated question in sub-sample s , for $p_1 = 1 - p_2$ and $p_1 \neq 0.5$. The model for the ECWM is given by

$$\pi_{y|s}^* = \begin{cases} p_s \pi_s + (1 - p_s)(1 - \pi_s) & \text{for } y = 2 \\ (1 - p_s)\pi_s + p_s(1 - \pi_s) & \text{for } y = 1 \end{cases} \tag{1}$$

Solving Eq (1) with respect to π_s for each sub-sample separately yields the moment estimators

$$\hat{\pi}_s = \frac{\hat{\pi}_{2|s}^* - 1 + p_s}{2p_s - 1}; \quad p_s \neq 0.5, \tag{2}$$

where $\hat{\pi}_{2|s}^*$ is estimated by the conditional proportion n_{2s}/n_s of “I have TWO ‘Yes’ or TWO ‘No’ answers” within sub-sample s . Note that the moment estimator Eq (2) yields a negative prevalence estimate $\hat{\pi}_s$ when in sub-sample s the numerator and denominator differ in sign. For example, this would happen when $p_1 = 0.8$ and the proportion of “I have TWO ‘Yes’ or TWO ‘No’ answers” in sub-sample 1 is below 20%, so that $\hat{\pi}_{2|1}^* < 0.2$.

Since under random assignment to the sub-samples $E(\hat{\pi}_1) = E(\hat{\pi}_2) = \pi$, the overall prevalence estimate is obtained as the weighted average

$$\hat{\pi} = \frac{n_1}{n} \hat{\pi}_1 + \frac{n_2}{n} \hat{\pi}_2, \tag{3}$$

where n_1 and n_2 are the respective sub-sample sizes, and $n = n_1 + n_2$ the total sample size. Substituting $1 - p_1$ for p_2 , the variance of $\hat{\pi}$ is given by

$$V(\hat{\pi}) = \frac{1}{n} \left[\pi(1 - \pi) + \frac{p_1(1 - p_1)}{(2p_1 - 1)^2} \right], \tag{4}$$

which is identical to the variance of $\hat{\pi}$ under the CWM.

We now present an alternative moment estimator that estimates π directly instead estimating π_1 and π_2 . In matrix notation the model is given by

$$\boldsymbol{\pi}_{y|s}^* = \mathbf{P}\boldsymbol{\pi}, \tag{5}$$

where $\boldsymbol{\pi}_{y|s}^* = (\pi_{2|1}^*, \pi_{1|1}^*, \pi_{2|2}^*, \pi_{1|2}^*)'$, $\boldsymbol{\pi} = (\pi, 1 - \pi)'$, and

$$\mathbf{P} = \begin{pmatrix} p_1 & 1 - p_1 \\ 1 - p_1 & p_1 \\ p_2 & 1 - p_2 \\ 1 - p_2 & p_2 \end{pmatrix} = \begin{pmatrix} p_1 & 1 - p_1 \\ 1 - p_1 & p_1 \\ 1 - p_1 & p_1 \\ p_1 & 1 - p_1 \end{pmatrix}, \tag{6}$$

is a 4×2 transition matrix with the randomization probabilities of both sub-samples. In the right-hand side of Eq (6) we have substituted $1 - p_1$ for p_2 .

The moment estimator of π is then given by the single expression

$$\hat{\pi} = 2\mathbf{P}^{-1}\hat{\boldsymbol{\pi}}_{ys}^*, \tag{7}$$

where \mathbf{P}^{-1} is the Moore-Penrose generalized inverse of \mathbf{P} , and $\hat{\boldsymbol{\pi}}_{ys}^*$ the vector with the unconditional response probabilities as estimated by n_{ys}/n . The estimated variance of $\hat{\pi}$ is given by

$$\hat{V}(\hat{\pi}) = 4n^{-1}\mathbf{P}^{-1}(\mathbf{Diag}(\hat{\boldsymbol{\pi}}_{ys}^*) - \hat{\boldsymbol{\pi}}_{ys}^*\hat{\boldsymbol{\pi}}_{ys}^{*t})(\mathbf{P}^{-1})^t. \tag{8}$$

Detection of informed self-protection

We now show that the statistical model of the ECWM accounting for informed self-protection is not identified. Let us first consider the situation without informed cheating under model Eq (5). From the equality of the rows 1 and 4 and rows 2 and 3 of the matrix \mathbf{P} in Eq (6), it follows that $\pi_{2|1}^* = \pi_{1|2}^*$ and $\pi_{1|1}^* = \pi_{2|2}^*$. In the absence of response biases, we therefore expect to observe the equality of the response proportions $\hat{\pi}_{2|1}^* = \hat{\pi}_{1|2}^*$ and $\hat{\pi}_{1|1}^* = \hat{\pi}_{2|2}^*$ in our sample.

Now assume that carriers and non-carriers give informed self-protective answers with probabilities θ_c and θ_{nc} . For $p_1 > p_2$, the sensitive response in sub-sample 1 is $y = 2$, so that here the respective probabilities p_1 and $1 - p_1$ that carriers and non-carriers answer $y = 2$ are reduced by the respective factors $1 - \theta_c$ and $1 - \theta_{nc}$. In sub-sample 2 where $y = 1$ is the sensitive response, the effect of informed self-protection is in the opposite direction. This situation is described by the model $\tilde{\boldsymbol{\pi}}_{y|s}^* = \mathbf{Q}\boldsymbol{\pi}$, where

$$\mathbf{Q} = \begin{pmatrix} p_1(1 - \theta_c) & (1 - \theta_{nc})(1 - p_1) \\ 1 - p_1(1 + \theta_c) & p_1 + \theta_{nc}(1 - p_1) \\ 1 - p_1(1 + \theta_c) & p_1 + \theta_{nc}(1 - p_1) \\ p_1(1 - \theta_c) & (1 - \theta_{nc})(1 - p_1) \end{pmatrix}, \tag{9}$$

is the transition matrix accounting for informed self-protection, and $\tilde{\boldsymbol{\pi}}_{y|s}^*$ the vector with the conditional response probabilities. It is obvious that $\tilde{\boldsymbol{\pi}}_{y|s}^* \neq \boldsymbol{\pi}_{y|s}^*$ for $\theta_c > 0$ and/or $\theta_{nc} > 0$. The equality relations $\tilde{\pi}_{2|1}^* = \tilde{\pi}_{1|2}^*$ and $\tilde{\pi}_{1|1}^* = \tilde{\pi}_{2|2}^*$ however are not affected by informed self-protection, since \mathbf{Q} has the same equality of the rows 1 and 4 and the rows 2 and 3 as \mathbf{P} . It follows

that $\tilde{\pi}_{y|s}^*$ can also be explained by model $\tilde{\pi}_{y|s}^* = P\pi$, and that therefore both the parameters θ_c and θ_{nc} of the model $\tilde{\pi}_{y|s}^* = Q\pi$ are unidentified. And since $P \neq Q$, the estimator $\hat{\pi}$ of model (5) is biased in the presence of informed self-protection.

Maximum likelihood estimation

In this section we derive a maximum likelihood estimator (MLE) for the ECWM. This MLE can be used to obtain a prevalence estimate of the sensitive characteristic, but can also be used to estimate the parameters of a logistic regression model [36, 37]. A goodness-of-fit test is performed by comparing the log likelihoods of the intercept-only model and the model with a covariate denoting sub-sample membership.

Let $\pi_{y|s}^*$ denote the probability of observing response y by individual i in sub-sample s , for $i \in \{1, \dots, n\}$, so that

$$\pi_{y|s}^* = \begin{cases} p_s \pi_i + (1 - p_s)(1 - \pi_i) & \text{for } y_{is} = 2 \\ (1 - p_s)\pi_i + p_s(1 - \pi_i) & \text{for } y_{is} = 1 \end{cases} \quad (10)$$

Model (10) is converted into a logistic regression model by the specification of the logistic function

$$\pi_i = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})}, \quad (11)$$

where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$ is the vector with covariates, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$ the corresponding vector with regression coefficients. The model is estimated by maximization of the log likelihood function

$$\ln \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{x}_i) = \sum_{i=1}^n \ln \pi_{y_{is}}^*. \quad (12)$$

The variance of the elements in $\hat{\boldsymbol{\beta}}$ are given by the diagonal elements of the inverse of the Hessian matrix. These variances are on the logit scale, but if so desired they can be transformed to the probability scale by use of the delta method.

The estimate of the population prevalence π is obtained by fitting the intercept-only model, and the point and interval estimates of the moment and maximum likelihood estimators are identical as long as the moment estimators are non-negative prevalence estimates [38]. The inclusion of a covariate denoting sub-sample membership yields the same prevalence estimates $\hat{\pi}_1$ and $\hat{\pi}_2$ as [26] moment estimator (2). The goodness-of-fit test statistic (likelihood ratio statistic) is then computed as twice the difference of log likelihood of the models with and without the sub-sample membership covariate, which is identical to the G^2 statistic given by $2 \sum_y \sum_s n_{ys} \log(n_{ys}/\hat{n}_{ys})$, where n_{ys} are the observed response frequencies, and \hat{n}_{ys} the fitted ones, unless the prevalence estimate is on the boundary of the parameter space (0, 1). If a boundary solution exists, the likelihood ratio statistic will be zero whereas the value of G^2 statistic will be greater than zero. The likelihood ratio statistic has an approximate chi-squared distribution on 1 degree of freedom, and a significant value provides evidence for instruction non-adherence.

The R package ECWM for moment and maximum likelihood estimation of the ECWM is available at the github page <https://github.com/Khadiga-S/ecwm>.

Data and methods

To evaluate the performance of the ECWM with the number sequence randomizer, three independent studies in non-athlete populations were conducted in 2020. Study II measures compliance with the UK Covid-19 lockdown rules between March 23rd and May 13th, 2020. Study I and Study III measure the respondents' experiences with dietary and herbal supplements, controlled stimulants and controlled drugs for performance and/or image enhancement. For a summary of the survey design and participants allocation to different survey formats in the three studies, see [S1 Table](#).

Participants

The participants in the three studies were recruited via the [Prolific Academic Platform](#). Prolific is an online research platform that enables researchers to recruit participants for research online in a cost-effective and reliable way. Participants are paid for their contribution at or above the minimum living wage, pro rata for the time estimated for survey completion. There is some evidence that the Prolific participants produce data quality higher than other online platforms [39]. Prior to their participation, all respondents consented to participate in our study and written informed consent was obtained. The studies did not involve minors and were subject to full ethical review prior to the data collection from which we obtained favorable ethical opinions. Ethical approval to perform the three studies was obtained from the Research Ethics Committee of the Faculty of Science, Engineering and Computing, Kingston University London. Ethical approval for the data analysis was also obtained from Utrecht University.

The target respondents in all studies were English-speaking people over the age of 18 years. Study II was restricted for those who were in the UK during the lockdown period. Of 2,424 participants who started study II, 6 (0.25%) persons were not in the UK during the lockdown and another 16 (0.62%) did not complete the survey. In Study I, the survey started with 1,518 individuals, 13 (0.85%) did not complete the questionnaire and answered only the demographic questions, while of 1,811 participating in Study III, 10 (0.55%) did not complete the survey. The Prolific platform was set to prevent the same person participating in the other two studies. The final samples consist of 1,505 respondents in Study I, 2,402 study II and 1,801 study III. Additional information about some background characteristics of the participants in the three studies and power curves to determine minimum sample size required for each CWM condition, can be found in [S1 Appendix](#) and [S1 Fig](#), respectively.

Questionnaires

The questionnaires of the three studies began with a brief introduction explaining the purpose and content of the survey. Participants in Study II were informed that they would be asked four questions related to their daily routines during the COVID19 lockdown, whereas in Study I and Study III they would be asked four questions about their experiences with dietary and herbal supplements, controlled stimulants and controlled drugs for performance- and/or image-enhancement, as well as some demographic questions. They were informed that the survey would take less than 5 minutes to complete, and that they could withdraw from the study at any time if they are unable or unwilling to complete the survey. Participants were informed that voluntary completion of the survey constitutes informed consent to participate in the study, and assured that the survey is completely anonymous, so it is totally safe to be honest and answer truthfully.

Procedures and measures

In each study, the survey began with a set of basic demographic questions about age, gender, highest level of education and the level of sport activity. Before progressing to the sensitive questions section, participants were asked to indicate their birth month or to select a month at random if they wish. It was explained to the participants that this information is only used to assign them to one of the experimental conditions with the same content, but without giving details on how this assignment was operationalized. Participants were blinded to the experimental condition allocation, which is detailed in [S2 Fig](#). In Study II and Study III the participants were randomly assigned to either a DQ condition or ECWM condition, whereas in Study I the DQ condition was not employed. In Study I, the sensitive behaviour was presented as question or statement, while only the question format was applied in Study III. In Study II, the questions were presented in two different formats, namely Factual and Judgmental. [Table 1](#) shows the observed response frequencies of the three studies in the DQ and the ECWM conditions. Participants in the DQ condition had to answer the sensitive questions with “Yes” or “No”, whereas in the ECWM condition they were asked to answer either “I have only one ‘Yes’ answer” without revealing which question was answered with “Yes”, or “I have two ‘Yes’ answers or none” without revealing whether it was two or none. The probabilities $1/5$ and $4/5$ of answering “Yes” to the unrelated question in the two ECWM conditions are indicated in brackets, and the answers “I have TWO ‘Yes’ or TWO ‘No’ answers” and “I have ONE ‘Yes’ answer” are denoted by 2 and 1, respectively. Respondents in the ECWM condition were asked an additional question about the perceived difficulty of the instructions. This question “How easy or difficult was it for you to answer this question format?” was scored on an easy to difficult scale from 0–100.

The surveys were hosted on a closed survey platform ([SurveyMonkey](#)) in all three studies. The number of eligible participants were 35,852 for Study I, 58,744 for Study II, and 28,368 for Study III. Data were collected on February 21st, May 22nd, and July 7th 2020, respectively. On average, participants were paid at the level of £11.57/hr, £9.87/hr £16.88/hr taking part in Studies I, II, and III, respectively.

The number sequence randomizer

In Studies I, II and III number sequence randomizers were used with respectively 15, 10 and 5 numbers. The first sequence consisted of randomly generated two-digit numbers, and the respondents were asked to memorize one of the numbers. In order to avoid selection bias, ‘lucky’ or easy-to-memorize numbers, the single-digit numbers 1 to 9, master numbers (11, 22, 33, 44, . . .) and multiples of ten (10, 20, 30, 40, . . .) were excluded from this sequence. Then a second number sequence of the same length was shown that potentially included all numbers from 1 to 99. Depending on the length of the first sequence and the sub-sample, 3, 2 or 1 numbers or 12, 8, or 4 numbers reappeared in the second sequence, so that the probability that the memorized number reappeared was $1/5$ in one sub-sample and $4/5$ in the other. These ratios were alternated between questions within each sub-sample. The respondents then had to answer the unrelated question “Is the number you selected included in the set below?” and the sensitive question. This procedure is illustrated in [Fig 2](#) for the Covid-F1 question of Study II. In this example, two of the ten numbers from the first sequence reappear in the second, so in this example the probability of answering “Yes” to the unrelated question B is $1/5$.

Since the first and second number sequence are not shown simultaneously, it makes it extremely difficult for the respondent to remember and keep track of how many numbers of the first sequence reappeared in the second one. Consequently, it is practically impossible for the respondent to infer the probability of a “Yes” answer to the unrelated question, and thus to

Table 1. Observed response frequencies of Studies I, II, and III.

	Experimental Conditions					
	DQ		ECWM (1/5)		ECWM (4/5)	
	Yes	No	2/0	1/1	2/0	1/1
Study I-Question						
Drug use1-Q1: Have you ever been drug tested, for any reason? (including breath alcohol tests)	-	-	249	112	150	223
Drug use1-Q2: Have you ever used nutritional or herbal supplements for reasons other than health? (for example to make you more alert, help with weight loss)	-	-	188	185	183	178
Drug use1-Q3: Have you ever used controlled psychoactive drugs? (e.g., cannabis, cocaine, ecstasy)	-	-	197	164	174	199
Drug use1-Q4: Have you ever used controlled sport drugs without medical need to enhance your athletic performance? (e.g., anabolic steroids, hormone)	-	-	275	98	120	241
Study I-Statement						
Drug use1-S1: I have been drug tested, at least once, in my life (including breath alcohol tests)	-	-	233	143	166	229
Drug use1-S2: I have used, at least once, nutritional or herbal supplements for reasons other than health (for example to make you more alert, help with weight loss)	-	-	197	198	171	205
Drug use1-S3: I have used, at least once, controlled psychoactive drugs (e.g., cannabis, cocaine, ecstasy)	-	-	187	189	208	187
Drug use1-S4: I have used controlled sport drugs without medical need, at least once, to enhance my athletic performance? (e.g., anabolic steroids, hormone)	-	-	304	91	120	256
Study II-Factual						
Covid-F1: Did you meet people socially outside your household during the lockdown period?	37	341	249	145	186	247
Covid-F2: Did you go out for shopping, errands or work more than it was necessary during the lockdown period	52	326	284	149	164	230
Covid-F3: Did you spend more than one hour outside your house exercising during the lockdown period?	128	250	233	161	181	252
Covid-F4: Did you use any form of transportation (including driving or cycling) to go for exercise outside your area during the lockdown period?	65	313	271	162	146	248
Study II-Judgmental						
Covid-J1: Did you break the social distancing rule during the lockdown period?	44	359	262	158	149	225
Covid-J2: Did you break the stay home rule during the lockdown period?	53	350	233	141	154	266
Covid-J3: Did you break the one-hour rule for outside exercise during the lockdown period?	103	300	237	183	151	223
Covid-J4: Did you break the stay in your area rule during the lockdown period?	28	375	210	164	155	265
Study III-Question						
Drug use2-Q1: Have you ever been drug tested, for any reason? (including breath alcohol tests)	121	469	438	163	232	378
Drug use2-Q2: Have you ever used nutritional or herbal supplements for reasons other than health? (for example to make you more alert, help with weight loss)	274	316	330	280	285	316
Drug use2-Q3: Have you ever used controlled psychoactive drugs? (e.g., cannabis, cocaine, ecstasy)	230	360	302	299	300	310
Drug use2-Q4: Have you ever used controlled sport drugs without medical need to enhance your athletic performance? (e.g., anabolic steroids, hormone)	26	564	478	132	138	463

<https://doi.org/10.1371/journal.pone.0279741.t001>

figure out which is the incriminating response. Obviously, this virtually rules out the possibility of informed self-protection.

Control question

Prolific participants receive payment for completing a survey, thus the more surveys they fill out, the higher their earnings. The reference to “earn more” means that the faster they complete a survey the better their hourly rate of pay is in the sense that the more time they have to participate in other surveys, posted by other researchers, completely independent of our study. Consequently, there is a risk of some form of survey ‘fatigue’ and/or ‘rush’, which may lead them to skip instructions or to careless answering of the questions. To assess the quality of respondents’ answers, the questionnaires of all studies included a control question with known prevalence, namely whether the participants were paid or not for participating in the survey.

In the DQ condition this question has an expected prevalence estimate of 100%. To ensure the same expectation in the ECWM condition, the probability of answering “Yes” to the corresponding unrelated question in Study I and Study II was fixed at 1 by letting all numbers of the first sequence reappear in a different order in the second one. This probability was set to be zero in Study III. Therefore, in the ECWM condition we expect to see 100% “I have TWO ‘Yes’ or TWO ‘No’ answers” in the first two studies and 100% “I have ONE ‘Yes’ answer” in Study III. This setup allows us to compare the error rates under DQ and the ECWM. In both conditions the response errors are attributable to survey fatigue, but in the ECWM condition there is the additional risk of non-adherence due to misunderstanding of the instructions or recall errors when using the number sequence randomizer. By comparing the error rates under DQ and the ECWM, we may get an estimate of the fraction of the error rate that is specifically attributable to the ECWM. Furthermore, by regressing the answers to the control question on variables assessing the time spent on the survey and the perceived difficulty of the ECWM instructions, we hope to gain some insight in the reasons for these response errors.

Results

This section presents the maximum likelihood prevalence estimates and goodness-of-fit statistics of all studies, followed by an analysis of the control question.

Prevalence estimates and goodness-of-fit

[Table 2](#) presents the prevalence estimates and goodness-of-fit statistics of the three studies. The last three columns present the difference of the prevalence estimates $\Delta\hat{\pi} = \hat{\pi}_{ECWM} - \hat{\pi}_{DQ}$ between the ECWM and the DQ for each question separately, the values of test statistic $z = \Delta\hat{\pi} / \sqrt{\text{var}(\Delta\hat{\pi})}$, with $\text{var}(\Delta\hat{\pi}) = \text{var}(\hat{\pi}_{ECWM}) + \text{var}(\hat{\pi}_{DQ})$, and the associated p -values. The results for Studies II and III with a DQ condition show that the prevalence estimates are in line with the “more-is-better” criterion; the ECWM yields significantly higher prevalence estimates than DQ when the questions are sensitive, and similar estimates when the questions are non-sensitive. The exceptions are Covid-F3, where the difference is not significant, and Drug use2-Q4, where DQ yields a (not significantly) higher prevalence estimate. Six goodness-of-fit tests are significant and show a lack of fit; three in Study I, two in Study II and one in Study III. Three of these only exceed the critical value of 3.84 by a small margin, and the largest G^2 statistics occur on the non-sensitive Drug use-Q1 questions, which suggests that the lack of fit is not due to self-protective response biases, but to a misunderstanding of the instructions on the first question.

[Table 3](#) presents the difference (*diff.*) of the ECWM prevalence estimates between the question and statement conditions of Study I, and the factual and judgmental conditions of Study II. The p -values of the test statistic z show that there are no significant differences between the corresponding questions, and in contrast to our expectation the factual condition did not yield consistently higher prevalence estimates than the judgmental condition.

Analysis of the control question

[Table 4](#) presents the results of the analysis of the control question. For this analysis, the data of the two format conditions of Study I and II were pooled. The percentages of incorrect answers on the ECWM control questions show a clear negative correlation with the length of the number sequence randomizer. Specifically, error rates in Study I using a 15-number sequence (with a probability of 1 of the unrelated question) are significantly higher than those in Study II using a 10-number sequence (with a probability of 1 of the unrelated question) ($X^2 = 4.14$,

Table 2. Prevalence estimates of Studies I, II, and III.

	DQ $\hat{\pi}$ (95% CI)	ECWM			ECWM–DQ		
		$\hat{\pi}$ (95% CI)	$G^2_{(1)}$	<i>p</i> -value	$\Delta\hat{\pi}$	<i>z</i> -value	<i>p</i> -value
Study I-Question							
Drug use1-Q1 *	-	26.2 (20.4, 32.0)	6.77	0.009	-	-	
Drug use1-Q2 *	-	50.2 (44.2, 56.3)	0.09	0.767	-	-	
Drug use1-Q3	-	43.4 (37.4, 49.4)	0.11	0.740	-	-	
Drug use1-Q4	-	16.2 (10.7, 21.7)	4.27	0.039	-	-	
Study I-Statement							
Drug use1-S1 *	-	33.5 (27.7, 39.2)	1.28	0.258	-	-	
Drug use1-S2 *	-	46.4 (40.6, 52.3)	1.67	0.197	-	-	
Drug use1-S3	-	52.5 (46.6, 58.4)	0.44	0.507	-	-	
Drug use1-S4	-	12.3 (7.0, 17.5)	7.65	0.006	-	-	
Study II-Factual							
Covid-F1	9.8 (6.8, 12.8)	33.4 (27.8, 38.9)	3.26	0.071	23.6	7.31	0.000
Covid-F2	13.8 (10.3, 17.2)	29.7 (24.2, 35.3)	4.56	0.033	16.0	4.80	0.000
Covid-F3	33.9 (29.1, 38.6)	35.6 (30.0, 41.2)	0.07	0.784	1.7	0.45	0.652
Covid-F4	17.2 (13.4, 21.0)	28.7 (23.2, 34.2)	0.01	0.915	11.5	3.39	0.001
Study II-Judgmental							
Covid-J1	10.9 (7.9, 14.0)	31.1 (25.5, 36.8)	0.41	0.521	20.3	6.17	0.000
Covid-J2	13.2 (9.9, 16.5)	28.6 (23.0, 34.2)	0.09	0.764	15.5	4.64	0.000
Covid-J3	25.6 (21.3, 29.8)	36.8 (31.1, 42.5)	0.83	0.362	11.3	3.07	0.002
Covid-J4	6.9 (4.5, 9.4)	33.6 (27.9, 39.3)	3.97	0.046	26.9	8.45	0.000
Study III-Question							
Drug use2-Q1 *	20.5 (17.3, 23.8)	21.0 (16.6, 25.4)	16.46	0.000	0.5	0.19	0.850
Drug use2-Q2 *	46.4 (42.4, 50.5)	44.4 (39.7, 49.1)	0.28	0.596	-2.0	-0.63	0.531
Drug use2-Q3	39.0 (35.1, 42.9)	49.1 (44.4, 53.8)	0.04	0.843	10.1	3.23	0.001
Drug use2-Q4	4.4 (2.8, 6.1)	3.8 (0.0, 7.7)	0.31	0.580	-0.6	-0.27	0.791

* Non-sensitive question.

<https://doi.org/10.1371/journal.pone.0279741.t002>

$df = 1, p = 0.021$). Additionally, the highest error rate of 13.2% was found in Study I using a 15-number sequence, and the lowest error rate of 6.9% was found in the Study III using a 5-number sequence ($X^2 = 25.09, df = 1, p < 0.001$). This result strongly suggests that part of the errors was due to overlooking the memorized number in the second sequence. These kind

Table 3. Comparison between question formats across ECWM condition.

	<i>diff.</i>	<i>SE(diff.)</i>	<i>z</i> -value	<i>p</i> -value
Question vs Statement				
Drug use1-Q1—Drug use1-S1	-7.3	4.2	-1.75	0.079
Drug use1-Q2—Drug use1-S2	3.8	4.3	0.88	0.377
Drug use1-Q3—Drug use1-S3	-9.1	4.6	-1.98	0.048
Drug use1-Q4—Drug use1-S4	3.9	3.9	1.00	0.317
Factual vs Judgmental				
Covid-F1—Covid-J1	2.3	4.0	0.56	0.575
Covid-F2—Covid-J2	1.2	4.0	0.29	0.773
Covid-F3—Covid-J3	-1.2	4.1	-0.29	0.772
Covid-F4—Covid-J4	-4.9	4.0	-1.21	0.225

<https://doi.org/10.1371/journal.pone.0279741.t003>

Table 4. Analysis of control question.

	error rate in %		parameter estimates	
	DQ	ECWM	<i>logtime</i>	<i>difficulty</i>
Study I (15-number sequence)	–	13.2	–0.001	0.125
Study II (10-number sequence)	3.8	10.7	–0.148	0.222**
Study III (5-number sequence)	3.5	6.9	0.083	0.458***

* $p < .05$,** $p < .01$,*** $p < .001$.
<https://doi.org/10.1371/journal.pone.0279741.t004>

of errors may also (partly) explain the significant G^2 statistics in Study I. The error rates of 3.8% and 3.5% in the DQ condition of Studies II and III suggest some form survey fatigue/rush among the respondents, but they also suggest that part of the errors in the ECWM are due to survey fatigue/rush. For the ECWM condition of Study III this would mean that about 3.5% of the errors are attributable survey fatigue, and about 3.4% to some kind of misinterpretation of the ECWM instructions.

In an effort to explain the incorrect answers to the control question in the ECWM conditions, we performed a series of standard logistic regressions with the standardized scores of the variables *logtime* (the logarithm of the time to complete the survey) and *difficulty* (the perceived difficulty of the ECWM instructions) as predictors. To facilitate comparison of the effects sizes, both predictors were standardized. There are no significant effects of *logtime*, but the effects of *difficulty* show a clear correlation with the length of the number sequence randomizer. The parameter estimates show that the probability of a response error increases with the perceived difficulty of the instructions, and that this effect becomes stronger as the length of the number sequence randomizer decreases. This result strongly suggests that the response errors are due to a mixture of overlooking the memorized numbers and misunderstanding the instructions, and that probability of the former occurring decreases as the length of the number sequence decreases.

A reviewer pointed out that there may be a negative association between education levels of the respondents and the error rates on the control question. In fact, in studies [19, 22] lower-educated respondents were found to produce more errors in the (E)CWM condition. We therefore conducted standard logistic regression analyses with the error rate as dependent variable and education as a nominal covariate (with 5 categories, including the level “other”) and compared the fit of the model with that of the intercept-only model using the difference in G^2 statistics ΔG^2 . The results showed no significant effect of education in Study I ($\Delta G^2 = 5.51$, $df = 4$, $p = 0.239$), and Study III ($\Delta G^2 = 2.90$, $df = 4$, $p = 0.576$). In Study II the effect of education was significant ($\Delta G^2 = 10.03$, $df = 4$, $p = 0.040$), with undergraduate and postgraduate respondents producing significantly less errors than respondents with GCSE or below.

Discussion

This paper evaluated the performance of the ECWM in combination with a number sequence randomizer. The ECWM distinguishes itself from other randomized response techniques by the use of the two neutral response options “I have TWO ‘Yes’ or TWO ‘No’ answers” and “I have only ONE ‘Yes’ answer” of the CWM, and by the use of two sub-samples with complementary probabilities of answering “Yes” to the unrelated question. This latter property allows for a goodness-of-fit test. The number sequence randomizer is proposed as an alternative for

the birthday question randomizer, that is commonly used in other applications of the CWM and ECWM.

The (E)CWM with a birthday randomizer has been investigated in numerous validation studies. The (E)CWM has been shown to provide higher levels of comprehension and perceived privacy protection [19] than other randomized response techniques. “Weak” validation studies, i.e. studies comparing the prevalence estimates obtained through the (E)CWM and DQ for sensitive attributes with unknown prevalence, have provided evidence for both the “more-is-better” criterion for socially undesirable attributes [15, 17, 25, 27, 29, 40], and the “less-is-better” criterion for socially desirable attributes [23]. Moreover, in “strong” validation studies, i.e., studies of a sensitive attribute with a known prevalence, the CWM yielded estimates that were close to the known prevalence [41, 42]. However, some validation studies of the CWM yielded results that were not in agreement with the “more-is-better” criterion [18, 42, 43]. A potential explanation of these latter results is the experimental character of the sensitive attributes, which either had a zero prevalence [44] or were experimentally induced [20, 22].

Although the validation studies with “non-experimental” sensitive attributes seem to support the validity of the (E)CWM, random responding has been identified as an alternative explanation for confirmation of the “more/less-is-better” criteria [44–46]. Random responding may occur due to insufficient comprehension of the instructions or inattentiveness or carelessness, and bias the prevalence estimates toward 50% for both socially desirable and undesirable attributes. In studies by [22, 47, 48], 2 to 19% of the respondents indeed admitted to answered the questions randomly. A study by [28] however strongly suggests that, while random responding cannot be ruled out completely, it only plays a minor role in conformations of the “more/less-is-better” criteria.

In general, the above-mentioned validation studies provided evidence for the validity of the (E)CWM with a birthday randomizer. The central question of the current paper is how the ECWM with the number sequence randomizer relates to the findings of these validation studies. It has already been argued that number sequence randomizer has some theoretical advantages over the birthday randomizer, the main two being the elimination of potential biases due to informed self-protection and due to a non-uniform distribution of birthdays over the days or months of the year. Another advantage that has not been mentioned yet is that the number sequence randomizer is easier to implement when multiple sensitive questions are asked, as was the case in the three studies we presented in this paper. It is no problem to present the respondent with five different sets of number sequences, but it is much more difficult to come up with five different birthday questions. One option is to use the same person’s birthday but to select different periods, e.g., the months January and February for the first question, the months March and April for the second, etc. Problems with this approach are that it may decrease perceived privacy protection, especially if the person’s birthday could be known to the researcher, and it excludes a multivariate analysis of the questions because the probabilities to answer “Yes” to the birthday questions are no longer independent. This dependence issue is solved by using the birthdays of different persons, like the father’s and the mother’s birthday, but what persons to select for the remaining three questions? Finally, the number sequence randomizer is also better suited for “strong” validation studies because it allows the randomization probabilities to be set to either 0 or 1, thus eliminating uncertainty in the prevalence estimates due randomization. Obviously, setting the randomization probabilities to 1 and/or 0 is not possible with birthday questions.

Our study yielded some promising empirical results that are inline with the evaluation studies of the (E)CWM with birthday randomizer. Practically all prevalence estimates of the ECWM were in line with the “more-is-better” criterion; the ECWM yielded higher prevalence

estimates for the sensitive questions and similar prevalence estimates for non-sensitive questions. The latter result can also be taken as an indication that random responding has not played a major role, because in that case the ECWM would have yielded higher prevalence estimates than DQ. In general, the goodness-of-fit tests showed a satisfactory fit. Three of the six significant G^2 statistics only exceeded the critical value of 3.84 by a small margin, and the two of the three largest G^2 statistics occurred on the first question and in Study I with the 15 number sequences. These results suggest that respondents may not have fully understood the instructions on the first question, and that the 15-number sequences are more prone to response errors than the shorter number sequences. This last explanation seems to be confirmed by the analysis of the control question, which suggests that shorter number sequence results in less response errors. The non-zero error rates on the control question in the DQ conditions suggests that there is some degree of survey fatigue/rushed completion, which may have also contributed to the G^2 's of the significant goodness-of-fit tests in the ECWM conditions.

Somewhat to our surprise we found no differential effects for judgmental versus factual question formats. A potential explanation for this finding is that aside of the wording of the questions, other factors like the appropriateness and context of the question also play an important role in the willingness of the respondent to answer honestly. In a study on the effects of the wording of the sensitive questions on the willingness to give honest answer, [49] found mixed results.

The statistical contribution of this paper concerns the development of a maximum likelihood estimator for the prevalence of the sensitive characteristic, and its extension to a binary logistic regression model that can be used to investigate the effects of covariates on the probability of having the sensitive characteristic. We have illustrated its use by performing a logistic regression with sub-sample membership as covariate, and shown that the results are equivalent to conducting the goodness-of-fit test.

The current study has some limitations. Firstly, since the validation only involved comparisons of the ECWM and DQ conditions, we have no empirical evidence for the superiority of the number sequence randomizer over the birthday randomizer since our study lacked a condition with the birthday randomizer. A second limitation of our study is that as yet we do not have a satisfactory explanation for the misfit of some of the ECWM questions, except for the conjecture that the misfit on the first questions may be due to insufficient comprehension of the instructions. However, a misfit for ECWM questions is not specific to the number sequence randomizer. For example, [28] observed a misfit on three of the eight ECWM questions using the birthday randomizer. Finally, our observation that the error rates on the control question increased with the length of the number sequences led us to the conclusion that the longer sequences are more prone to errors than the shorter ones. Due to the confounding of randomization probabilities and the studies, with Study III having the lowest error rates and a randomization probability of 0 and Studies I and II having the highest error rates and randomization probabilities of 1, it cannot be ruled out that the randomization probabilities had an effect on the error rates. To avoid such confounding, we have set the randomization probability of the control question to 0 in one sub-sample, and to 1 in the other sub-sample in new surveys we are conducting.

Conclusion

In summary, the ECWM with the 5-number sequence randomizer exhibited a good performance in estimating the prevalence of sensitive issues. Its design offers a good combination of features that collectively ensure, as much as possible, that the data it generates are valid and

reliable. The ECWM with number sequence randomizer reduces the risk of response biases since it does not offer an obvious strategy for self-protective responses, while its use of two sub-samples stills allows testing for other forms of response bias. With its relatively simple instructions and its short length, the 5-number sequence randomizer also reduces the risk of response errors. Furthermore, it is also well suited for the anticipated large scale applications, for instance, international sport events. As a tool for WADA, a sequence of questions with incremental sensitivity is recommended, along with a control question to provide an indication of the degree to which the respondents understood the instructions.

Supporting information

S1 Fig. Power curves of the crosswise model.

(TIF)

S2 Fig. Allocation of respondents to experimental conditions.

(TIF)

S1 Appendix. Background characteristics of the participants in the three studies.

(PDF)

S1 Table. Summary of the project design in Studies I, II and III.

(PDF)

Author Contributions

Conceptualization: Khadiga H. A. Sayed, Maarten J. L. F. Cruyff, Peter G. M. van der Heijden, Andrea Petróczi.

Investigation: Khadiga H. A. Sayed, Maarten J. L. F. Cruyff, Andrea Petróczi.

Methodology: Khadiga H. A. Sayed, Maarten J. L. F. Cruyff, Peter G. M. van der Heijden, Andrea Petróczi.

Writing – original draft: Khadiga H. A. Sayed, Maarten J. L. F. Cruyff, Peter G. M. van der Heijden.

Writing – review & editing: Khadiga H. A. Sayed, Maarten J. L. F. Cruyff, Peter G. M. van der Heijden, Andrea Petróczi.

References

1. Pitsiladis YP, Abatzis-Papadopoulos M, Ali N, Aggeloudis S, Atkinson C, Constandache B, et al. The Olympia Declaration. *Current Sports Medicine Reports*. 2019; 18(12):448–451. <https://doi.org/10.1249/JSR.0000000000000660> PMID: 31834176
2. World Anti-Doping Agency. Prohibited list: January 2021. Available at: <https://www.wada-ama.org/en/resources/science-medicine/prohibited-list-documents>; 2021.
3. Gleaves J, Petróczi A, Folkerts D, De Hon O, Macedo E, Saugy M, et al. Doping prevalence in competitive sport: Evidence synthesis with “best practice” recommendations and reporting guidelines from the WADA Working Group on Doping Prevalence. *Sports Medicine*. 2021; 51:1909–1934. <https://doi.org/10.1007/s40279-021-01477-y> PMID: 33900578
4. De Hon O, Kuipers H, van Bottenburg M. Prevalence of doping use in elite sports: a review of numbers and methods. *Sports medicine*. 2015; 45(1):57–69. <https://doi.org/10.1007/s40279-014-0247-x> PMID: 25169441
5. Chaudhuri A, Mukerjee R. *Randomized response: Theory and techniques*. New York: Marcel Dekker; 1988.

6. Lensvelt-Mulders GJ, Hox JJ, van der Heijden PGM, Maas CJ. Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research*. 2005; 33(3):319–348. <https://doi.org/10.1177/0049124104268664>
7. Ulrich R, Pope HG, Cléret L, Petróczi A, Nepusz T, Schaffer J, et al. Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine*. 2018; 48(1):211–219. <https://doi.org/10.1007/s40279-017-0765-4> PMID: 28849386
8. Greenberg BG, Abul-Ela ALA, Simmons WR, Horvitz DG. The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*. 1969; 64(326):520–539. <https://doi.org/10.1080/01621459.1969.10500991>
9. Peeters CF, Lensvelt-Mulders GJ, Lasthuizen K. A note on a simple and practical randomized response framework for eliciting sensitive dichotomous and quantitative information. *Sociological Methods & Research*. 2010; 39(2):283–296. <https://doi.org/10.1177/0049124110378099>
10. Droitcour J, Caspar RA, Hubbard ML, Parsley TL, Visscher W, Ezzati TM. 11. In: *The item count technique as a method of indirect questioning: A review of its development and a case study application*. Wiley Online Library; 2004. p. 185–210.
11. Petróczi A, Nepusz T, Cross P, Taft H, Shah S, Deshmukh N, et al. New non-randomised model to assess the prevalence of discriminating behaviour: a pilot study on mephedrone. *Substance abuse treatment, prevention, and policy*. 2011; 6(1):1–18.
12. Nepusz T, Petróczi A, Naughton DP, Epton T, Norman P. Estimating the prevalence of socially sensitive behaviors: Attributing guilty and innocent noncompliance with the single sample count method. *Psychological methods*. 2014; 19(3):334–355. <https://doi.org/10.1037/a0034961> PMID: 24295152
13. Yu JW, Tian GL, Tang ML. Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika*. 2008; 67(3):251–263. <https://doi.org/10.1007/s00184-007-0131-x>
14. Coutts E, Jann B, Krumpal I, Näher AF. Plagiarism in student papers: prevalence estimates using special techniques for sensitive questions. *Jahrbücher für Nationalökonomie und Statistik*. 2011; 231(5-6):749–760. <https://doi.org/10.1515/jbnst-2011-5-612>
15. Jann B, Jerke J, Krumpal I. Asking sensitive questions using the crosswise model: an experimental survey measuring plagiarism. *Public opinion quarterly*. 2012; 76(1):32–49. <https://doi.org/10.1093/poq/nfr036>
16. Korndörfer M, Krumpal I, Schmukle SC. Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology*. 2014; 45:18–32. <https://doi.org/10.1016/j.joep.2014.08.001>
17. Hoffmann A, Musch J. Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods*. 2016; 48(3):1032–1046. <https://doi.org/10.3758/s13428-015-0628-6> PMID: 26182857
18. Höglinger M, Jann B, Diekmann A. Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods*. 2016; 10(3):171–187. <https://doi.org/10.18148/srm/2016.v10i3.6703>
19. Hoffmann A, De Puseau BW, Schmidt AF, Musch J. On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior research methods*. 2017; 49(4):1470–1483. <https://doi.org/10.3758/s13428-016-0804-3> PMID: 27631988
20. Höglinger M, Jann B. More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PloS one*. 2018; 13(8):e0201770. <https://doi.org/10.1371/journal.pone.0201770> PMID: 30106973
21. Hopp C, Speil A. Estimating the extent of deceitful behaviour using crosswise elicitation models. *Applied Economics Letters*. 2019; 26(5):396–400. <https://doi.org/10.1080/13504851.2018.1486007>
22. Meisters J, Hoffmann A, Musch J. Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? *PloS one*. 2020; 15(6):e0235403. <https://doi.org/10.1371/journal.pone.0235403> PMID: 32603352
23. Mieth L, Mayer MM, Hoffmann A, Buchner A, Bell R. Do they really wash their hands? Prevalence estimates for personal hygiene behaviour during the COVID-19 pandemic based on indirect questions. *BMC public health*. 2021; 21(1):1–8. <https://doi.org/10.1186/s12889-020-10109-5> PMID: 33397344
24. Hoffmann A, Meisters J, Musch J. Nothing but the truth? Effects of faking on the validity of the crosswise model. *PloS one*. 2021; 16(10):e0258603. <https://doi.org/10.1371/journal.pone.0258603> PMID: 34714838
25. Schnell R, Thomas K. A Meta-analysis of Studies on the Performance of the Crosswise Model. *Sociological Methods & Research*. 2021; p. 1–26. <https://doi.org/10.1177/0049124121995520>

26. Heck DW, Hoffmann A, Moshagen M. Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior research methods*. 2018; 50(5):1895–1905. <https://doi.org/10.3758/s13428-017-0957-8> PMID: 28916924
27. Meisters J, Hoffmann A, Musch J. Controlling social desirability bias: An experimental investigation of the extended crosswise model. *PloS one*. 2020; 15(12):e0243384. <https://doi.org/10.1371/journal.pone.0243384> PMID: 33284820
28. Meisters J, Hoffmann A, Musch J. More than random responding: Empirical evidence for the validity of the (Extended) Crosswise Model. *Behavior Research Methods*. 2022; p. 1–14. <https://doi.org/10.3758/s13428-022-01819-2> PMID: 35449499
29. Sagoe D, Cruyff M, Spendiff O, Chegeni R, De Hon O, Saugy M, et al. Functionality of the Crosswise Model for assessing sensitive or transgressive behavior: A systematic review and meta-analysis. *Frontiers in Psychology*. 2021; 12:1–19. <https://doi.org/10.3389/fpsyg.2021.655592> PMID: 34248750
30. Sieff K. In Afghanistan, Jan. 1 is everybody's birthday; 2013 Dec 31. Available from: https://www.washingtonpost.com/world/in-afghanistan-its-everyones-birthday/2013/12/31/81c18700-7224-11e3-bc6b-712d770c3715_story.html.
31. Clark SJ, Desharnais RA. Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods*. 1998; 3(2):160–168. <https://doi.org/10.1037/1082-989X.3.2.160>
32. Böckenholt U, van der Heijden PGM. Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika*. 2007; 72(2):245–262. <https://doi.org/10.1007/s11336-005-1495-y>
33. Pitsch W, Emrich E. The frequency of doping in elite sport: Results of a replication study. *International Review for the Sociology of Sport*. 2011; 47(5):559–580. <https://doi.org/10.1177/1012690211413969>
34. Elbe AM, Pitsch W. Doping prevalence among Danish elite athletes. *Performance enhancement & health*. 2018; 6(1):28–32. <https://doi.org/10.1016/j.peh.2018.01.001>
35. Umesh UN, Peterson RA. A critical evaluation of the randomized response method: Applications, validation, and research agenda. *Sociological Methods & Research*. 1991; 20(1):104–138. <https://doi.org/10.1177/0049124191020001004>
36. Scheers N, Dayton CM. Covariate randomized response models. *Journal of the American Statistical Association*. 1988; 83(404):969–974. <https://doi.org/10.1080/01621459.1988.10478686>
37. Van der Heijden PGM, Van Gils G. Some logistic regression models for randomized response data. In: Forcina, A., Marchetti, G.M., Hatzinger, R., Falmacci, G. (Eds.), *Proceedings of the 11th International Workshop on Statistical Modelling*; 1996. p. 341–348.
38. Van den Hout A, van der Heijden PGM. Randomized response, statistical disclosure control and misclassification: a review. *International Statistical Review*. 2002; 70(2):269–288. <https://doi.org/10.1111/j.1751-5823.2002.tb00363.x>
39. Peer E, Brandimarte L, Samat S, Acquisti A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*. 2017; 70:153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>
40. Hoffmann A, Musch J. Prejudice against women leaders: Insights from an indirect questioning approach. *Sex Roles*. 2019; 80(11):681–692. <https://doi.org/10.1007/s11199-018-0969-6>
41. Hoffmann A, Diedenhofen B, Verschuere B, Musch J. A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology*. 2015; 62(6):403–414. <https://doi.org/10.1027/1618-3169/a000304> PMID: 27120562
42. Hoffmann A, Meisters J, Musch J. On the validity of non-randomized response techniques: an experimental comparison of the crosswise model and the triangular model. *Behavior research methods*. 2020; 52(4):1768–1782. <https://doi.org/10.3758/s13428-020-01349-9> PMID: 32185638
43. Jerke J, Rauhut D, Thomas K, Velicu A. Handle withCare: Implementation of the list experiment and crosswise model in a large-scale survey on academic misconduct. *Field Methods*. 2022; 34(1):69–81. <https://doi.org/10.1177/1525822X20985629>
44. Höglinger M, Diekmann A. Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis*. 2017; 25:131–137. <https://doi.org/10.1017/pan.2016.5>
45. Walzenbach S, Hinz T. Pouring water into wine: Revisiting the advantages of the crosswise model for asking sensitive questions. *Survey Methods: Insights from the Field*. 2019; p. 1–16. <https://doi.org/10.13094/SMIF-2019-00002>
46. Atsusaka Y, Stevenson RT. A bias-corrected estimator for the crosswise model with inattentive respondents. *Political Analysis*. 2021; p. 1–15. <https://doi.org/10.1017/pan.2021.43>

47. Enzmann D. Die Anwendbarkeit des Crosswise-Modells zur Prüfung kultureller Unterschiede sozial erwünschten Antwortverhaltens. In: Eifler, S., Faulbaum, F. (eds) *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung*. Springer Fachmedien Wiesbaden; 2017. p. 239–277.
48. Schnapp P. Sensitive question techniques and careless responding: Adjusting the crosswise model for random answers. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*. 2019; 13(2):307–320. <https://doi.org/10.12758/mda.2019.03>
49. Näher A, Krumpal I. Asking sensitive questions: the impact of forgiving wording and question context on social desirability bias. *Quality & Quantity: International Journal of Methodology*. 2012; 46(5):1606–1616. <https://doi.org/10.1007/s11135-011-9469-2>