BMC Public Health

RESEARCH

Open Access

Prevalence estimates for COVID-19-related health behaviors based on the cheating detection triangular model

Shu-Hui Hsieh[1*], Pier Francesco Perri[2] and Adrian Hoffmann[3]

Abstract

Background Survey studies in medical and health sciences predominantly apply a conventional direct questioning (DQ) format to gather private and highly personal information. If the topic under investigation is sensitive or even stigmatizing, such as COVID-19-related health behaviors and adherence to non-pharmaceutical interventions in general, DQ surveys can lead to nonresponse and untruthful answers due to the influence of social desirability bias (SDB). These effects seriously threaten the validity of the results obtained, potentially leading to distorted prevalence estimates for behaviors for which the prevalence in the population is unknown. While this issue cannot be completely avoided, indirect questioning techniques (IQTs) offer a means to mitigate the harmful influence of SDB by guaranteeing the confidentiality of individual responses. The present study aims at assessing the validity of a recently proposed IQT, the Cheating Detection Triangular Model (CDTRM), in estimating the prevalence of COVID-19-related health behaviors while accounting for cheaters who disregard the instructions.

Methods In an online survey of 1,714 participants in Taiwan, we obtained CDTRM prevalence estimates via an Expectation-Maximization algorithm for three COVID-19-related health behaviors with different levels of sensitivity. The CDTRM estimates were compared to DQ estimates and to available official statistics provided by the Taiwan Centers for Disease Control. Additionally, the CDTRM allowed us to estimate the share of cheaters who disregarded the instructions and adjust the prevalence estimates for the COVID-19-related health behaviors accordingly.

Results For a behavior with low sensitivity, CDTRM and DQ estimates were expectedly comparable and in line with official statistics. However, for behaviors with medium and high sensitivity, CDTRM estimates were higher and thus presumably more valid than DQ estimates. Analogously, the estimated cheating rate increased with higher sensitivity of the behavior under study.

Conclusions Our findings strongly support the assumption that the CDTRM successfully controlled for the validity-threatening influence of SDB in a survey on three COVID-19-related health behaviors. Consequently, the CDTRM appears to be a promising technique to increase estimation validity compared to conventional DQ for health-related behaviors, and sensitive attributes in general, for which a strong influence of SDB is to be expected.

Keywords Cheating detection, COVID-19, Indirect questioning, Nonresponse, Privacy protection, Social desirability bias, Misreporting

*Correspondence:
Shu-Hui Hsieh
shhsieh@gate.sinica.edu.tw
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

In recent years, the daily lives of large segments of the global population have been severely impacted by the effects of the COVID-19 pandemic. Seeking to mitigate the spread of the SARS-CoV-2 coronavirus, government, and public health institutions provided behavioral recommendations, restrictions, and protective measures, including physical distancing, wearing masks, following personal hygiene rules, getting vaccinated, and isolating in case of a positive test. To evaluate the effectiveness of these measures, obtaining valid prevalence estimates for people's compliance with such interventions is of the utmost importance. To this end, numerous online surveys based on conventional self-reporting in a direct questioning (DQ) format have been conducted [1–6].

When answering questions on COVID-19-related issues, and generally, on the adherence to non-pharmaceutical interventions, respondents should primarily refer to their actual behavior; yet, responses are also likely influenced by personal beliefs, attitudes and expectations. These influences can have dramatic consequences for the validity of the results obtained if the topic under investigation is perceived as highly sensitive. To this end, [7] assume that the perceived sensitivity of a question is mainly determined by its intrusiveness, threat of disclosure, and social desirability. For example, a question about an individual's COVID-19 vaccination status will likely be perceived as intrusive and therefore highly sensitive by participants who consider health-related information as something innately private. The perceived threat of disclosure reflects personal concerns about potential consequences associated with either answer option; for example, if revealing one's vaccination status likely results in negative consequences such as legal sanctions or social stigma, this will result in high perceived sensitivity. Finally, a topic will be perceived as sensitive if one of the answer options is regarded as more socially desirable than the other; for example, if a respondent decided against a vaccination while the social norm is to get vaccinated, the respondent may perceive a question about the individual vaccination status as highly sensitive. As a consequence, when questioned about sensitive behaviors, some participants might decide in favor of protecting their individual privacy by denying a response. Alternatively, as nonresponse may also be perceived as indicative of a specific sensitive behavior, some participants might provide an answer that is socially desirable rather than truthful. This response behavior is typically referred to as social desirability bias (SDB) and results in the underreporting of socially undesirable, as well as the overreporting of socially desirable behaviors, respectively. In DQ surveys on COVID-19-related behaviors as well as sensitive issues in general, nonresponses and untruthful responses represent nonsampling errors that are chronically difficult to address; these errors severely threaten data quality and, therefore, the validity of results obtained in subsequent analyses.

Anonymous online survey situations make direct disapproval highly unlikely because of the absence of social peers. However, cognitive dissonance due to participants' actions contradicting social norms can still lead to misreporting. For example, overreporting of desirable behaviors has been shown to inflate estimates of COVID-19 vaccination rates in specific populations [8]. In other populations, the individual vaccination status may be perceived as less (or non-) sensitive, resulting in an attenuated influence of SDB. Understanding and addressing these complex issues is essential for obtaining accurate prevalence estimates for COVID-19-related health behaviors in a population under study. In summary, researchers from the medical, health, and social sciences must carefully and thoroughly consider all aspects of the data collection process, and should specifically focus on a potentially harmful influence of nonsampling errors if the topic under investigation may be perceived as sensitive.

While nonresponse and untruthful responses can hardly be completely avoided, numerous approaches have been developed to mitigate their negative influence by guaranteeing the confidentiality of individual answers, and thereby increasing respondents' cooperation (for an overview, see [9, 10]). One promising approach is given by indirect questioning techniques (IQTs), a class of data-collection methods that rely on the randomization of individual answers. These techniques allow participants to truthfully respond to sensitive questions without revealing anything about their true status with respect to the behavior under study.

Questions in the randomized response technique (RRT) format, the first IQT introduced by [11], rely on randomization of individual answers to maximize confidentiality and cooperation, but at the cost of lower estimation efficiency. When answering an RRT question, participants are instructed to use a randomization device (e.g., a die or a spinner) to determine whether to respond to a positively or negatively formulated sensitive statement. Since the individual outcome of the randomization remains unknown to the experimenter, respondents can reply "true" or "false" without revealing their true status regarding the sensitive attribute, thereby guaranteeing confidentiality of individual answers and presumably increasing the willingness to respond truthfully. The distribution of randomization outcomes is however known to the experimenter, thereby allowing for estimating the prevalence of the sensitive attribute on sample level with potentially higher validity than with a conventional DQ. IQTs are expected to provide higher and thus presumably more valid prevalence estimates for socially undesirable attributes ("more-is-better" assumption) and lower estimates for socially desirable attributes ("less-is-better" assumption). However, the "more-is-better" and "less-is-better" assumptions may fail in cases where the direction of SDB is unclear. Recent evidence suggests that the "more-is-better" principle may be problematic, as the validity of estimates could be threatened by false positives and false negatives [12].

Since Warner's original model [11], many advanced IQTs have been proposed that aim at further improving prevalence estimation for sensitive attitudes and behaviors (comprehensive overviews are provided in the monographs by [13–20]). Relating to the purposes of the current work, several previous studies in the field of public health policies and services have investigated whether SDB influences self-reports of compliance with non-pharmaceutical interventions. Specifically relevant here are studies that used the indirect questioning approach to investigate potentially sensitive attributes in the context of the COVID-19 pandemic (e.g., [8, 21–24]).

In the present paper, we applied the recently proposed Cheating Detection Triangular Model (CDTRM) [25] to estimate the prevalence of several COVID-19-related health behaviors in an online survey in Taiwan, and experimentally compared its performance to a conventional DQ control condition. Extending on the results of the work cited above, the application of the CDTRM in our study not only allows to control for nonresponse and SDB, but also for a dual assessment of the influence of social desirability on self-report data. First, this influence is visible in the form of differences between CDTRM and DQ prevalence estimates for sensitive behaviors; second, the CDTRM allows for the direct estimation of the proportion of respondents who are non-compliant to the instructions and provide a self-protective response (referred to as "cheaters" in the CDTRM framework) by means of a dedicated model parameter.

The Cheating Detection Triangular Model

While IQTs generally aim at increasing respondents' cooperation in sensitive surveys and reducing nonresponse and untruthful responding, many of the available models do not account for potential problems during the answering process and introduce new issues to be adequately considered.

The validity of IQTs such as the original RRT by [11], basically relies on two tacit assumptions: (i) all survey participants are completely honest when answering the statement chosen by the randomization procedure; (ii) all participants correctly execute the instructions prescribed by the procedure. However, some respondents may fail to understand the instructions and unintentionally provide an answer that does not actually apply to them. Others may distrust the method because they do not understand how it protects the confidentiality of their answer; some may even believe that there is a mathematical trick that somehow links their response to a specific status with respect to the sensitive attribute. Consequently, these respondents might deliberately disregard the instructions and resort to a presumably self-protective response (e.g., "no" in the sense of "I don't possess the sensitive attribute"), irrespective of the randomization outcome. In general, respondents who, for one reason or another, do not adhere to instructions are widely referred to as "cheaters". Outside the field of IQT research, the term "cheating" often carries negative connotations, implying that survey participants are intentionally behaving in an undesirable manner. However, it is important to recognize that participants have the right to choose responses that do not accurately represent their true behavior, and that such a choice is not inherently negative but may merely reflect a self-protective answering strategy. In the context of the CDTRM, the term "cheating" refers to non-compliance with the method's instructions, which could stem from either insufficient comprehension of the somewhat complex procedure, or mistrust towards the confidentiality protection of individual answers. Thus, in this study, we use the term "cheating" to denote non-compliant answering behavior.

Empirical studies have shown that participants' trust towards, and comprehension of, IQTs is often far from perfect [26], potentially resulting in questionable data quality. While compared to conventional DQ, IQTs are plausibly expected to attenuate the problem of cheating on the instructions, any substantial cheating rate may still negatively affect the prevalence estimates obtained by introducing a form of bias that needs to be controlled for. Hence, more sophisticated IQTs have been proposed that allow for detecting and estimating the share of cheaters in the sample. Among these are the Cheating Detection Model (CDM) [27] and the Stochastic Lie Detector [28], empirically applied by, for example, [23, 25, 26, 29–31]. As a current advancement of the IQT method, [25] recently proposed the CDTRM with the cheating detection mechanism of the CDM. Without loss of generality, let us consider here and ahead a survey that requires a binary response ("true" or "false") to a sensitive statement, with a "true" response implying carrying the sensitive attribute.

In the CDM format, respondents are presented with a sensitive statement with unknown population prevalence η (e.g., "I have at least once tested positive for COVID-19") as well as a nonsensitive statement used for randomization (e.g., "I was born between January and April"). The answer to the randomization statement determines whether they are instructed to respond honestly to the sensitive statement with probability p (e.g., if they were born between January and April, $p \cong 4/12$), or to simply answer "true" with probability $1 - p$ (if they were born in any other month, $1 - p \cong 8/12$) irrespective of their true status. Note that the randomization probability p denotes a design parameter with known value as it is controlled by the researcher. In addition, the CDM accounts for the possibility that some respondents disregard the instructions. Accordingly, the population is ideally classified into three nonoverlapping groups. The first two groups consist of respondents following the instructions, that is, honest carriers (in a proportion equal to π) and honest noncarriers of the sensitive attribute (in a proportion equal to β). The third group represents cheaters (in a proportion equal to $\gamma = 1 - \pi - \beta$), that is, respondents who disregard the instructions and choose the self-protective response ("false"), possibly due to insufficient trust or understanding. Notably, the true status of cheaters with respect to the sensitive attribute remains unknown; it is possible that none, some, or all of them carry the sensitive attribute. Therefore, it can be reasonably assumed that the true population prevalence of the sensitive attribute falls between the lower bound of π (if no cheater is a carrier) and the upper bound of $\pi + \gamma$ (if all cheaters were carriers), so that $\pi \leq \eta \leq \pi + \gamma$. As the CDM has two unknown parameters that need to be estimated (π and γ), two independent samples s_1 and s_2 with different randomization probabilities ($p_1 \neq p_2$) are required.

The Triangular Model (TRM) [32] was designed to provide comparatively simple instructions. Thereby, instead of a post-hoc detection of cheaters in the sample (e.g., via the cheating detection mechanism of the CDM), the TRM aims at reducing the cheating rate preventively by maximizing respondents' comprehension. Using the same example as for the CDM, the TRM presents respondents with two statements simultaneously: the sensitive statement A with unknown population prevalence η (e.g., "I have at least once tested positive for COVID-19") and a nonsensitive statement B with known prevalence p used for randomization (e.g., "I was born between January and April"). In a joint response to both statements A and B, participants are then simply required to indicate whether "at least one of the statements is true" or "none of the statements is true". While the second response option ("none of the statements is true") can of course be chosen truthfully, it is also a self-protective option that precludes being a carrier of the sensitive attribute.

The CDTRM now combines the favorable properties of the CDM and the TRM aiming at: (i) maximizing the proportion of respondents adhering to the instructions (i.e., minimizing the cheating rate); (ii) detecting the proportion of nonadherent respondents (i.e., the cheating rate); and (iii) providing a lower and an upper bound for the prevalence estimate of the sensitive attribute, thereby accounting for the estimated prevalence of cheaters. In analogy to the CDM, respondents in the CDTRM framework are reasonably classified into the three groups of honest carriers (proportion π), honest noncarriers (proportion β), and cheaters (proportion γ, with $\pi + \beta + \gamma = 1$); they receive the same instructions and answer options as in the TRM, while cheaters who deliberately try to conceal their status are expected to choose the self-protective option "none of the statements is true" irrespective of the instructions. As in the CDM, estimating two unknown parameters in the CDTRM requires two independent samples s_1 and s_2 with different randomization probabilities ($p_1 \neq p_2$). A graphical representation of the CDTRM is provided in Fig. 1. It is worth observing that the model assumptions of the CDM and the CDTRM are actually far more liberal than those of competing models incorporating different types of cheating such as, e.g., the Stochastic Lie Detector [28], which assumes that only carriers have a motivation to cheat, while all non-carriers respond with perfect honesty.

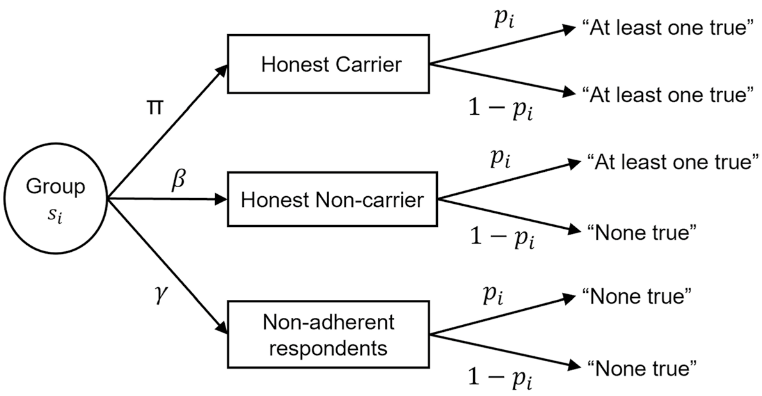

Fig. 1 Tree diagram of the CDTRM

In order to derive the estimates for π, β and γ in a closed-form framework, let λ_i be the probability of responding "at least one of the statements is true" in the sample s_i selected according to a simple random sampling design with replacement, $i = 1, 2$:

$$\lambda_i = \pi + \beta p_i.$$

Solving a system of two equations in the unknowns π and β yields to:

$$\pi = \frac{\lambda_1 p_2 - \lambda_2 p_1}{p_2 - p_1} \quad \text{and} \quad \beta = \frac{\lambda_1 - \lambda_2}{p_1 - p_2}.$$

From $\pi + \beta + \gamma = 1$, it readily follows that:

$$\gamma = 1 - \frac{\lambda_2(1 - p_1) - \lambda_1(1 - p_2)}{p_2 - p_1}.$$

Replacing λ_i with its sample counterpart $\widehat{\lambda}_i$ representing the observed proportion of respondents who answer "at least one of the statements is true" in sample s_i, the method-of-moments unbiased estimators of π, β and γ are straightforwardly obtained as:

$$\widehat{\pi} = \frac{\widehat{\lambda}_1 p_2 - \widehat{\lambda}_2 p_1}{p_2 - p_1}, \quad \widehat{\beta} = \frac{\widehat{\lambda}_1 - \widehat{\lambda}_2}{p_1 - p_2} \quad \text{and} \quad \widehat{\gamma} = 1 - \frac{\widehat{\lambda}_2(1 - p_1) - \widehat{\lambda}_1(1 - p_2)}{p_2 - p_1}.$$

The corresponding variances of the estimators are:

$$\mathrm{Var}(\widehat{\pi}) = \frac{1}{(p_2 - p_1)^2}\left[p_2^2 \mathrm{Var}(\widehat{\lambda}_1) + p_1^2 \mathrm{Var}(\widehat{\lambda}_2)\right],$$

$$\mathrm{Var}(\widehat{\beta}) = \frac{1}{(p_1 - p_2)^2}\left[\mathrm{Var}(\widehat{\lambda}_1) + \mathrm{Var}(\widehat{\lambda}_2)\right],$$

and

$$\mathrm{Var}(\widehat{\gamma}) = \frac{1}{(p_1 - p_2)^2}\left[(1 - p_1)^2 \mathrm{Var}(\widehat{\lambda}_2) + (1 - p_2)^2 \mathrm{Var}(\widehat{\lambda}_1)\right],$$

with $\mathrm{Var}(\widehat{\lambda}_i) = \dfrac{\lambda_i(1 - \lambda_i)}{n_i}$ while n_i denotes the sample size of sample s_i, $i = 1, 2$. The above variance can be estimated unbiasedly by replacing $\mathrm{Var}(\widehat{\lambda}_i)$ with $\widehat{\mathrm{var}}(\widehat{\lambda}_i) = \dfrac{\widehat{\lambda}_i(1 - \widehat{\lambda}_i)}{n_i - 1}$. Finally, for hypothesis testing purposes in subsequent analyses, it may also be useful to obtain the covariance between $\widehat{\pi}$ and $\widehat{\gamma}$:

$$\mathrm{Cov}(\widehat{\pi}, \widehat{\gamma}) = \frac{1}{(p_1 - p_2)^2}\left[(p_1^2 - p_1)\mathrm{Var}(\widehat{\lambda}_2) + (p_2^2 - p_2)\mathrm{Var}(\widehat{\lambda}_1)\right].$$

It is worth observing that, while the method-of-moments provides closed-form solutions and is often convenient due to its simplicity, the resulting estimates may not always be feasible. In fact, the estimation procedure may lead to point and interval estimates for θ ($\theta = \pi, \beta, \gamma$) outside the interval [0,1], which is meaningless in practice. The Maximum Likelihood (ML) method improves on this issue by truncating point estimates to [0,1] which leads to the estimator:

$$\widehat{\theta}^* = \min\left(\max(0, \widehat{\theta}), 1\right) \tag{1}$$

However, despite the favorable statistical properties of the ML method, interval estimates may still fall outside the interval [0,1]. To overcome this problem, the Expectation-Maximization (EM) algorithm [33, 34] is a broadly applicable and widely accepted approach to the iterative computation of ML estimates, which has the advantage of producing both reliable point and interval estimates and is often associated with a smaller standard deviation. In the current work, we derived and applied an EM algorithm to obtain prevalence estimates, standard errors, and confidence intervals based on CDTRM response data. Algebraic details about the implementation of the expectation and maximization steps are provided in Appendix 1, which is in the supplementary material available only online.

In the only empirical application of the CDTRM reported to date [25], the model was shown to produce higher, and thus potentially more valid, prevalence estimates for experimentally induced cheating behavior than conventional DQ; moreover, the CDTRM estimate was substantially closer to the known true value than the DQ estimate. Finally, in a direct comparison, the CDTRM was evaluated more favorably than its predecessor models, the TRM and the CDM. In the CDTRM, a substantial cheating rate was determined; in contrast, a consideration of cheaters is generally not possible in the TRM, potentially biasing the prevalence estimates obtained. Estimation validity was comparable between the CDTRM and the CDM; however, the CDTRM increased both objective comprehensibility and subjective evaluation by participants, presumably due to its simplified instructions. Against this background, for the current study, the CDTRM appeared to be a promising means to reduce the influence of SDB with easy-to-understand instructions, a dedicated cheating detection mechanism, and a high level of acceptance by respondents.

Aims of the current study

Our study aimed at obtaining accurate prevalence estimates for several COVID-19-related health behaviors using self-reports. By obtaining potentially unbiased and comprehensive data on these behaviors, we intended to contribute to a better understanding of the COVID-19 situation in Taiwan and, most importantly, to validate a methodology designed to control for the influence of socially desirable responding that could be extended to other similar research questions, for instance on the adherence to non-pharmaceutical interventions, and inform public health strategies and measures. To meet our research objectives, we conducted an online survey on a sample drawn from the general public in Taiwan and collected responses to questions about three COVID-19-related health behaviors differing in their assumed level of sensitivity:

- Topic 1: Having received at least one dose of a COVID-19 vaccine (low sensitivity);
- Topic 2: Having at least once tested positive for COVID-19 (medium sensitivity);
- Topic 3: Having at least once intentionally concealed a positive COVID-19 test result from others (high sensitivity).

While the terms "low", "medium", and "high sensitivity" might suggest a quantitative metric, it should be explicitly noted here that we chose these terms arbitrarily without any pretests, for the sole purpose of plausibly ordering them from least sensitive to most sensitive. This choice allowed a test of hypotheses pertaining to relative topic sensitivity; an absolute interpretation of the assumed sensitivity was not intended, nor it is recommended.

When investigating sensitive issues such as COVID-19-related health behaviors, it is important to take potential consequences of the sensitivity of these behaviors on participants' response patterns into account. To this end, concerns regarding an invasion of privacy, stigmatization, and negative legal or social consequences can result in nonresponse, untruthful, or biased responses, threatening the validity of the results obtained. Online surveys such as the current one are expected to result in comparatively high levels of perceived anonymity, confidentiality, and trust among participants. In contrast to nonanonymous personal interviews, participants generally do not have to fear direct disapproval from interviewers or any third parties due to their absence during the online survey situation. However, even in anonymous online surveys, some participants choose to respond dishonestly to sensitive questions in DQ format, or not respond at all, in an effort to present themselves in a socially desirable light. Against this background, the conventional DQ survey mode appears inappropriate for assessing the prevalence of sensitive behaviors because of its susceptibility to nonresponse and response biases.

To address this issue, we applied the CDTRM, an IQT designed to increase response rates and control for the harmful influence of SDB in sensitive surveys. In an experimental design, the prevalence estimates obtained in a CDTRM condition were compared to estimates from a DQ control condition, as well as to available official data provided by the Taiwan Centers for Disease Control. To evaluate the validity of the estimates obtained and the effectiveness of the CDTRM to control for SDB, we resorted to a comparative validation approach (i.e., the "more-is- better" criterion for socially undesirable behaviors). Additionally, we estimated the cheating rate associated with each of the topics under study, that is, the share of participants who disregarded the instructions and provided a self-protective response to CDTRM questions. Finally, we conducted exploratory analyses to examine the influence of potential moderators such as gender and age.

In summary, our study was designed for the empirical evaluation of two research questions. Specifically, we intended to assess whether:

1. For socially undesirable COVID-19-related health behaviors, prevalence estimates obtained in the CDTRM condition are higher and thus presumably more valid than estimates obtained in the DQ control condition.
2. The difference between CDTRM and DQ prevalence estimates, as well as the CDTRM cheating rate, increase with higher levels of topic sensitivity. Specifically, both the difference between CDTRM and DQ estimates and the CDTRM cheating rate are expected to be negligible or small for Topic 1 with low sensitivity, larger for Topic 2 with medium sensitivity, and largest for Topic 3 with high sensitivity.

The following section provides a detailed description of the methodology applied in the current study including all relevant details of the survey design and the data collection process. Subsequently, we will present the results obtained and discuss their implications.

Methods

Participants and procedure

Our study was part of the probability-based web panel conducted by the Center for Survey Research (CSR) at the Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan. This panel includes respondents who have previously been selected for

Table 1 Distribution of demographic variables by experimental condition

	General sample		DQ		CDTRM		
	($n = 1,714$)		($n_{DQ} = 573$)		($n_1 + n_2 = 1,141$)		χ^2 **Test**
	Frequency	**Percentage**	**Frequency**	**Percentage**	**Frequency**	**Percentage**	(*p*-value)
Gender							
Male	821	47.90	272	47.47	549	48.12	0.041
Female	893	52.10	301	52.53	592	51.88	(.804)
Age (years)							
Age 18 − 34	465	27.13	152	26.53	313	27.43	0.432
Age 35 − 49	720	42.01	247	43.11	473	41.46	(.806)
Age 50 or older	529	30.86	174	30.36	355	31.11	
Education							
College degree or lower	449	26.20	140	22.43	309	27.08	1.564
Bachelor's degree	825	48.13	279	48.69	546	47.85	(.457)
Master's degree or higher	440	25.67	154	26.88	286	25.07	

probability-based samples and participated in surveys conducted by CSR using various methods such as face-to-face interviews, telephone, and online surveys via short message services. An initial sample of 3,296 adults at least 18 years of age (1,102 [33%] assigned to the DQ and 2,194 [67%] to the CDTRM condition) received an invitation email from CSR to participate in the survey; data were collected online from October 11 to 25, 2022. At the beginning of the survey, participants were explicitly informed about the confidentiality of their responses, and were assured that all information they provided would only be used in a strictly anonymized format, and exclusively by the research group that conducted the survey. Ethical approval for the survey was obtained from the Institutional Review Board on Humanities and Social Science Research at the Academia Sinica, Taiwan (No. AS-IRB-HS07-111165). All participants who completed the survey received financial compensation in the form of a NT$50 e-voucher for a convenience store, right after submitting their responses. The mean time respondents needed for completing the survey was 14 minutes.

After removing invalid email addresses and incomplete data sets, a final sample of 1,714 participants (response rate: 52%) was obtained and used for subsequent analyses. Of these participants, 52% identified as female; with respect to age group, 27% reported to be 18-34 years old, 42% 35-49 years old, and 31% 50 years or older. These respondents provided complete information for all questions of the online survey described below, including the three questions assessing sensitive COVID-19-related health behaviors. An analysis of the allocation of the final sample to experimental conditions revealed that $n_{DQ} = 573$ (33%) respondents were assigned to the DQ control condition, and 1,141 (67%) respondents to the CDTRM condition; within the CDTRM condition, $n_1 = 545$ (32%) respondents were assigned to sample s_1 with randomization probability $p_1 = 5/6$, and $n_2 = 596$ (35%) to sample s_2 with randomization probability $p_2 = 1/6$.

The final sample size and the allocation of respondents to experimental conditions ensured sufficient statistical power for estimating the prevalence of the behaviors under study. This assumption is supported by the power analyses for the CDM – which is mathematically equivalent to the CDTRM used in our study – discussed in [35]. In the final sample, there were no significant differences between experimental conditions with respect to the distribution of the demographic variables gender, age group, and educational achievement (see Table 1).

Materials and design

The online questionnaire began with a brief introduction explaining the purpose and content of the study, as well as questions on demographic variables, and continued with questions on participants' personal satisfaction with respect to general well-being (quality of life and happiness) and environmental factors (air quality, behaviors, and attitudes towards air pollution). Participants were then surveyed about the three COVID-19-related health behaviors in either CDTRM or DQ format, depending on the experimental condition they had been assigned to. In the CDTRM condition (samples s_1 and s_2, respectively), a training phase preceded the actual questions on COVID-19-related behaviors. In this phase, participants first received comprehensive instructions and an explanation of how the novel CDTRM format protected the confidentiality of their answers. Subsequently, participants were presented with two fictitious examples and a total of three questions pertaining to these examples, which were designed to measure objective and subjective comprehension of the instructions. Details on the comprehension questions and a discussion of the potential influence of instruction comprehension are provided in Appendix 2.

After the training phase was over, participants in the CDTRM condition with randomization probability p_1 (sample s_1) were assured that their status with respect to the statements used for randomization (their own month of birth, and that of their parents) were not known and would not be asked for. Subsequently, they were required to answer the three questions on COVID-19-related health behaviors adhering to the following instructions:

CDTRM question 1: (for Topic 1 with low sensitivity)

Please think about your previous behavior with respect to a COVID-19 vaccination. You are now presented with two statements labeled Statement A and Statement B:

- Statement A: I have received at least one dose of a COVID-19 vaccine.
- Statement B: I was born between January and October.

Instead of answering each statement individually, provide a joint response to both statements simultaneously. Please choose one of the following response options:

1. Both statements are false.
2. At least one statement is true, irrespective of which one.

CDTRM question 2: (for Topic 2 with medium sensitivity)

Please think about your previous COVID-19 test results. You are now presented with two statements labeled Statement A and Statement B:

- Statement A: I have at least once tested positive for COVID-19.
- Statement B: My father was born between January and October.

Instead of answering each statement individually, provide a joint response to both statements simultaneously. Please choose one of the following response options:

1. Both statements are false.
2. At least one statement is true, irrespective of which one.

CDTRM question 3: (for Topic 3 with high sensitivity)

Please think about your previous behavior with respect to the nondisclosure of COVID-19 test results. You are now presented with two statements labeled Statement A and Statement B:

- Statement A: I have at least once intentionally concealed a positive COVID-19 test result from others.
- Statement B: My mother was born between January and October.

Instead of answering each statement individually, provide a joint response to both statements simultaneously. Please choose one of the following response options:

1. Both statements are false.
2. At least one statement is true, irrespective of which one.

For participants in the CDTRM condition with randomization probability p_2 (sample s_2), instructions and questions were identical to the participants in sample condition s_1 (with randomization probability p_1) with the only exception that the nonsensitive Statement B used for randomization referred to the birth month of respondents and parents falling into November or December ($p_2 = 1 - p_1$).

Participants in the DQ control condition were presented with the three sensitive statements only (identical to Statement A in each of the examples above) and required to respond either "True" or "False". The wording of the statements was as follows:

DQ question 1: I have received at least one dose of a COVID-19 vaccine.
DQ question 2: I have at least once tested positive for COVID-19.
DQ question 3: I have at least once intentionally concealed a positive COVID-19 test result from others.

The three questions on COVID-19-related health behaviors were presented in consistent and identical order in all experimental conditions.

Statistical analyses

In the current study, we provided a plausible range for the prevalence estimate of three COVID-19-related health behaviors among Taiwanese individuals in a CDTRM and a DQ control condition as described in the previous sections. To ensure transparency and reproducibility of our results, we report observed response frequencies by experimental questioning technique conditions and demographic subgroups in Appendix 3.

Table 2 Prevalence estimates for COVID-19-related health behaviors by questioning technique

	DQ			CDTRM					
	$\widehat{\eta}_{DQ}$	**SE**	95% **CI**	$\widehat{\pi}_L$	**SE**	95% **CI**	$\widehat{\pi}_U$	**SE**	95% **CI**
Topic 1	0.925	0.011	[0.903, 0.947]	0.874	0.017	[0.839, 0.906]	0.936	0.026	[0.886, 0.987]
Topic 2	0.384	0.020	[0.344, 0.424]	0.417	0.026	[0.368, 0.471]	0.554	0.041	[0.477, 0.640]
Topic 3	0.045	0.009	[0.028, 0.062]	0.184	0.023	[0.140, 0.229]	0.362	0.039	[0.280, 0.438]

$\widehat{\pi}_L$ denotes the lower bound estimate in the CDTRM condition (if none of the cheaters were carriers of the sensitive attribute: $\widehat{\pi}_L = \widehat{\pi}$); $\widehat{\pi}_U$ denotes the upper bound estimate (if all cheaters were carriers: $\widehat{\pi}_U = \widehat{\pi} + \widehat{\gamma}$); SE: the standard error; 95% CI: the 95% confidence interval

In the CDTRM framework, in addition to estimates for the share of honest carriers of the sensitive attribute and honest noncarriers, an estimate for the share of participants disregarding the instructions and choosing a self-protective response is provided (cheating rate). Notably, the status of cheaters with respect to the sensitive attribute remains unknown; cheaters could be carriers of the sensitive attribute, noncarriers, or a mixture of both. Consequently, the true prevalence of the sensitive attribute η is expected to fall within the range of the lower bound π (if none of the cheaters were carriers of the sensitive attribute) and the upper bound $\pi + \gamma$ (if all cheaters were carriers).

To obtain estimates, standard errors, and confidence intervals for π, β and γ based on observed response data, we derived and applied an EM algorithm detailed in Appendix 1. In contrast to other procedures such as, for example, closed-form estimation or the application of the function RRuni from the R package `RRreg` [36], our EM-based procedure has the advantage that parameter estimates will always fall into the range of [0,1].

To test for significant differences in the population between the prevalence estimates obtained ($\widehat{\eta}_{DQ}$ in the DQ condition; $\widehat{\pi}_L = \widehat{\pi}$ and $\widehat{\pi}_U = \widehat{\pi} + \widehat{\gamma}$ in the CDTRM condition), we considered the Z-test statistics

$$Z = \frac{\widehat{\eta}_{DQ} - \widehat{\pi}_L}{\sqrt{\widehat{\text{var}}(\widehat{\eta}_{DQ}) + \widehat{\text{var}}(\widehat{\pi}_L)}} \quad \text{and} \quad Z = \frac{\widehat{\eta}_{DQ} - \widehat{\pi}_U}{\sqrt{\widehat{\text{var}}(\widehat{\eta}_{DQ}) + \widehat{\text{var}}(\widehat{\pi}_U)}},$$

where $\widehat{\text{var}}(\widehat{\eta}_{DQ}) = \widehat{\eta}_{DQ}(1 - \widehat{\eta}_{DQ})/n_{DQ}$, while $\widehat{\text{var}}(\widehat{\pi}_L)$ and $\widehat{\pi}_U$ obtained in a bootstrap setting as described in Appendix 1.

As the majority of comparisons referred to directed research questions, we mostly applied one-sided tests representing the expected direction of the differences between estimates on population level. Specifically, for the sensitive Topics 2 and 3 for which a substantial influence of social desirability was to be expected, we tested the null hypothesis that the lower and upper bound CDTRM prevalence estimates were identical to the DQ estimate in the population ($\eta_{DQ} = \pi_L, \eta_{DQ} = \pi_U$) against the alternative hypothesis that the CDTRM prevalence estimates were higher than the DQ estimate ($\eta_{DQ} < \pi_L$, $\eta_{DQ} < \pi_U$). In contrast, for Topic 1, the low sensitivity of the topic did not suggest any substantial influence of social desirability bias. Therefore, we had no clear expectations with respect to the direction of potential effects relating to the first two research questions, and consequently based all respective comparisons on two-sided Z-test. Similarly, we considered tests for the significance of the cheating rate ($\widehat{\gamma}$) comparing the null hypothesis of no cheaters in the population ($\gamma = 0$) against the alternative hypothesis of any cheaters ($\gamma > 0$).

Moreover, pairwise comparisons of cheating rates (see Tables A3 and A4 in Appendix 2) allowed for testing the null hypothesis that in the population, participants with low comprehension would show a comparable or lower cheating rate than participants with high comprehension ($\gamma_{lo} \leq \gamma_{hi}$) against the alternative hypothesis of a higher cheating rate in case of low comprehension ($\gamma_{lo} > \gamma_{hi}$).

Results

The main findings of our analyses are summarized in Tables 2 and 3, and in Fig. 2. Table 2 provides prevalence estimates, with the corresponding standard error (SE) and 95% confidence interval (95% CI), for the three surveyed topics in the general sample by questioning technique as well as by participant gender and age group. In Table 3, we report the observed value of the Z-test statistic, say z, and the respective p-value for pairwise comparisons between prevalence estimates, and for the significance of the cheating rate. Figure 2 shows that prevalence estimates differed significantly, and mostly in the expected direction, between DQ and CDTRM experimental conditions, specifically for the topics with medium (Topic 2) and high sensitivity (Topic 3). In the general sample, we evaluated whether for socially undesirable attributes, the CDTRM would result in higher and thus presumably more valid prevalence estimates than DQ as well as substantial estimated cheating rates, and whether these effects would be more pronounced with increasing topic sensitivity (the two research questions). In the following, we present results obtained for Topics 1 to 3, differing in sensitivity from low to high.

Table 3 Results of tests for the significance of differences between prevalence estimates in DQ versus CDTRM conditions, and of the cheating rate in the CDTRM condition

	$\widehat{\eta}_{DQ} - \widehat{\pi}_L$	*z*	*p*-value	$\widehat{\eta}_{DQ} - \widehat{\pi}_U$	*z*	*p*-value	$\widehat{\gamma}$	*z*	*p*-value
Topic 1	0.051	2.563	.010	-0.011	-0.406	.685	0.063	4.328	$<$ .001
Topic 2	-0.033	-1.013	.156	-0.170	-3.741	$<$ .001	0.137	5.907	$<$ .001
Topic 3	-0.139	-5.612	$<$ .001	-0.317	-7.845	$<$ .001	0.178	7.280	$<$ .001

Two-sided test for Topic 1 and (left) one-sided test for Topics 2 and 3

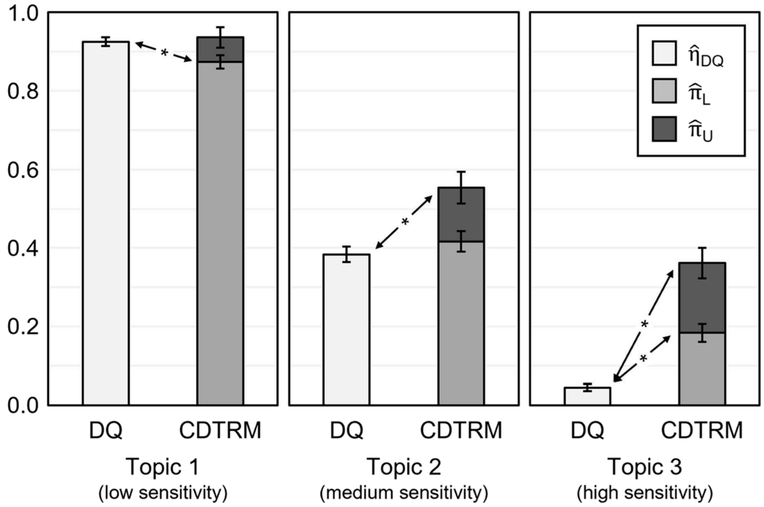

Fig. 2 Comparison between CDTRM and DQ

For having received at least one dose of the COVID-19 vaccine (Topic 1), prevalence estimates were high, and widely comparable between DQ and CDTRM conditions. The DQ estimate (92.5%) and the CDTRM upper bound estimate (93.6%) were not significantly different. Therefore, the CDTRM estimate did not suffice the "more-is-better" assumption; this finding was actually in line with our expectations due to the presumably low sensitivity of Topic 1 in Taiwan. Notably, both the DQ and the CDTRM upper bound estimates closely aligned with the official vaccination rate for Taiwanese adults as of October 24, 2022, which was approximately 95.6%[1]. Unexpectedly, however, the lower bound estimate in the CDTRM condition (87.4%) was significantly lower than the DQ estimate (92.5%), and an estimated cheating rate in the CDTRM condition significantly higher than zero was observed ($\widehat{\gamma}$ = 6.3%). A potential explanation for these findings is that in the collectivistic society of Taiwan, unvaccinated individuals surrounded by vaccinated people may choose to conceal their vaccination status in fear of social stigma. Consequently, participants in the DQ condition may have actually overreported their vaccination status, as being vaccinated was presumably perceived as socially desirable. The significant share of non-compliant respondents in the CDTRM condition is, however, hardly attributable to the influence of SDB, but rather to insufficient comprehension of the instructions among some participants. Previous applications of the CDTRM have similarly reported that insufficient comprehension of the method may result in slightly biased estimates, and increased cheating rates [25]. Considering the cheating rate in the CDTRM, both the upper bound CDTRM estimate and the DQ estimate closely reproduced the true COVID-19 vaccination rate in the Taiwanese population, as indicated by official statistics. Previous studies on public compliance with COVID-19 measures have similarly reported that no substantial influence of SDB was observed when using IQTs [37–39] or face-saving strategies [40]. Most importantly, results for Topic 1 are widely in line with the two research questions of the current work, stating that differences between DQ and CDTRM estimates are expected to be small (or even negligible) in the case of low topic sensitivity.

[1] The vaccination rate was calculated as the ratio between the cumulative number of first doses administered to individuals aged 18 and above (18,894,367) as of October 24, 2022, and the total population of individuals in the same age group (19,752,983). Official data (in Chinese) are available from the Taiwan Centers for Disease Control via the following link: https://www.cdc.gov.tw/En/File/Get/Ct02SqB4TU8xCKkRJfnkUQ.

For having at least once tested positive for COVID-19 (Topic 2), an attribute with medium sensitivity, Tables 2 and 3 show that prevalence estimates were in the medium range, and differed somewhat between DQ and CDTRM conditions. In the CDTRM condition, the prevalence was estimated to fall between 41.7% and 55.4%, and the cheating rate was substantially and significantly above zero ($\widehat{\gamma} = 13.7\%$). Expectedly sufficing the "more-is-better" assumption, the CDTRM upper bound estimate (55.4%) significantly exceeded the DQ estimate (38.4%); the CDTRM lower bound estimate (41.7%) did however not differ significantly from the DQ estimate. Notably, all prevalence estimates obtained in the DQ and CDTRM conditions were substantially higher than the rate of confirmed COVID-19 cases for individuals aged 20 or older in Taiwan as of October 25, 2022, which was approximately 30.7%[2]. One potential explanation for these discrepancies could be that participants may have been more honest when answering our survey due to the strictly anonymous survey situation, and the methodology applied. Moreover, survey participants in our study may have included positive results of home rapid tests that had not been included in official data. Finally, a potential explanation for the specific observation that CDTRM estimates exceeded DQ estimates and official statistics is that the influence of SDB on participants from collectivistic cultures may be particularly strong. This influence may result in a high pressure to underreport socially undesirable behaviors under non-confidential conditions, thus biasing both DQ estimates and official data. Importantly, results for Topic 2 were widely in line with the two research questions, showing that for a topic with medium sensitivity, the upper bound CDTRM estimate was higher and thus presumably more valid than a DQ estimate and that a significant rate of participants cheated by choosing a self-protective response.

For having at least once intentionally concealed a positive COVID-19 test result from others (Topic 3), a behavior with high sensitivity, the results reported in Tables 2 and 3 paint a very clear picture. While the DQ estimate (4.5%) suggests a relatively low rate of individuals having concealed a positive test result, this estimate was significantly and substantially exceeded by both the CDTRM lower bound (18.4%) and the CDTRM upper bound estimates (36.2%). Moreover, the cheating rate in the CDTRM condition was estimated at 17.8% which was well above zero. Unfortunately, our results for Topic 3 could not be compared to official data as no official statistics on the prevalence of intentionally concealed positive test results were available. The current findings however suggest that for the highly sensitive Topic 3, socially desirable responding has likely resulted in an underestimation in the DQ condition; this was to be expected, as the intentional concealment of a positive COVID-19 test is very plausibly considered to be socially undesirable in Taiwan. The CDTRM, on the other hand, has provided higher and thus presumably more valid prevalence estimates through a successful control of SDB, and additionally provided an estimate for a substantial rate of cheaters choosing a self-protective response. To this end, results for Topic 3 were in full support of the two research questions of this work.

To further validate the empirical results observed in our study, and to explore the usefulness of the CDTRM in other health-related contexts such as the general adherence to non-pharmaceutical interventions, we realized a simulation study considering five different hypothetical situations. Across these situations, we systematically varied the population prevalence (η), and the proportion of cheating (γ). The simulation results reported in Table A6 (see Appendix 4) clearly show that all estimate intervals $[\widehat{\pi}_L, \widehat{\pi}_U]$ include the respective true value of η; in some cases the true value is closer to the upper bound, in others it is closer to the lower bound. Moreover, the simulation shows that the CDTRM does not accidentally overestimate the cheating rate, but rather provides very precise estimates of γ. Therefore, the apparently rather high estimates for γ in our empirical study on COVID-19-related health behaviors (for Topic 2, and especially for Topic 3) are very likely a precise representation of intentional or unintentional non-compliance (in case of Topic 1, probably unintentional), rather than a statistical artifact.

In summary, across Topics 1 to 3, the observed differences between prevalence estimates in the DQ and CDTRM conditions and the estimated cheating rates in the CDTRM condition strongly support the two research questions of this work. Differences between DQ and CDTRM estimates and the CDTRM cheating rate were least pronounced for Topic 1 with low sensitivity, more pronounced for Topic 2 with medium sensitivity, and most pronounced for Topic 3 with high sensitivity. These results, and the confirmation of a precise estimation of the cheating rate from our simulation study, suggest that the CDTRM is a promising tool for controlling for the influence of SDB on self-reports, and for providing additional estimates for the rate of non-compliant respondents also when attitudes or behaviors with low prevalence are surveyed.

[2] The rate of confirmed COVID-19 cases was calculated as the ratio of the cumulative number of confirmed cases among individuals aged 20 and above (5,927,140) as of October 25, 2022, and the total population of individuals in the same age group (19,306,244). Official open-access data (in Chinese) are available at: https://data.gov.tw/dataset/151770.

For a deeper analysis of the results in the general sample, we also examined prevalence estimates and comparisons across subgroups by gender and age. Results of these analyses along with a few comments on their implications are presented in Appendix 5. We observed a significant CDTRM cheating rate for COVID-19 vaccination (Topic 1), particularly among women. This aligns with findings of gender-related differences in health behaviors during pandemics [41]. Additionally, older participants (50+) showed higher cheating rates for testing positive for COVID-19 or concealing a positive test (Topics 2 and 3). These findings suggest increased SDB among older individuals. This contrasts with studies showing older people were generally more truthful about following health guidelines [42] and underscores the need for further research on factors affecting honesty in self-reports of sensitive health behaviors.

Overall, our findings provide comprehensive insights into the prevalence of the COVID-19-related health behaviors under study and strongly support the assumption that the CDTRM provides increased estimation validity. In general, our results indicate that the method can be used effectively to survey public health issues beyond pandemic-specific behaviors, particularly with regard to adherence to non-pharmaceutical interventions.

Discussion

In the current study, we assessed the prevalence of three COVID-19-related health behaviors differing in sensitivity in a large online survey in Taiwan. To control for the potentially validity-threatening influence of SDB, we applied the CDTRM, a current indirect questioning technique with particularly simple instructions and a dedicated cheating detection mechanism. This model guarantees the confidentiality of individual responses while maintaining the possibility of obtaining prevalence estimates at the sample level, presumably leading to a higher willingness of participants to provide truthful responses compared to conventional DQ. Additionally, the model provides an estimate for the share of participants disregarding the instructions and providing a self-protective response (i.e., the cheating rate). In an experimental design, we compared prevalence estimates for the three behaviors under study obtained in a CDTRM condition to those obtained in a DQ control condition. According to the “more-is-better” assumption, and as reflected in our a priori-formulated research questions, we expected prevalence estimates in the CDTRM condition to be higher and thus potentially more valid than DQ estimates if the CDTRM would indeed be capable of controlling for the influence of SDB. This difference between conditions as well as the cheating rate in the CDTRM condition were expected to be positively associated with topic sensitivity, that is, to increase from small or negligible effects for Topic 1 (low sensitivity) over medium effects for Topic 2 (medium sensitivity) to large effects for Topic 3 (high sensitivity). Additionally, we explored potentially moderating influences of participant gender and age group.

For having received at least one dose of the COVID-19 vaccine (Topic 1), we expected rather high prevalence estimates and a low (if any) effect of SDB. Official statistics indicated that the vast majority of the Taiwanese population had been vaccinated, and the sociopolitical climate emphasized being vaccinated as the social norm; consequently, this behavior was expected to be of low sensitivity in Taiwan although, in other societies (e.g., Germany), this topic may indeed be considered sensitive [8]. Consistent with these expectations and with the two a priori-formulated research questions of the current work, prevalence estimates for Topic 1 in our study were high, largely comparable between the DQ and the CDTRM condition, and close to official statistics; furthermore, the share of participants choosing a self-protective response in the CDTRM condition (i.e., the cheating rate) was small, although unexpectedly significant. For having at least once been tested positive for COVID-19 (Topic 2), we expected a somewhat stronger influence of SDB. This topic was presumably of medium sensitivity, as testing positive for COVID-19 is not necessarily a result of disregarding social norms or official behavioral guidelines, but may nevertheless be perceived as indicative of less-than-optimal behavior during the pandemic. As expected, the CDTRM (higher bound) prevalence estimate exceeded the estimate in the DQ condition, presumably due to a successful control of SDB; the CDTRM cheating rate was also substantial, and higher than that for Topic 1. Having at least once intentionally concealed a positive COVID-19 test result from others (Topic 3) was expected to be highly sensitive, as it clearly violated social norms – especially within the socio-political climate of Taiwan. Consistent with these expectations, a very low DQ estimate was obtained; both CDTRM lower and upper bound estimates were found to be substantially higher, thus presumably less distorted by SDB, and ultimately more valid. For Topic 3, the highest cheating rate was observed, indicating a strong tendency of participants towards choosing a self-protective response. Taken together, the observed results strongly support both of our research questions.

Unexpectedly, prevalence estimates for having at least once been tested positive for COVID-19 (Topic 2) in both the DQ and CDTRM conditions exceeded respective numbers from official statistics. Similar findings have been reported by [43] who found that certain aspects of the testing and reporting process may have led to an underestimation of COVID-19 cases in official statistics for various countries. In particular, the increasing availability of home antigen tests (rapid tests) over the course of the pandemic may have introduced specific challenges in accurately capturing COVID-19 cases via official statistics. From May 12, 2022, the Taiwanese government had implemented a new policy stating that individuals with a positive rapid test result subsequently confirmed by a medical unit would officially be recognized as confirmed cases of COVID-19, and would have to undergo home isolation. However, some individuals with a positive rapid test result, especially if they were asymptomatic or had only mild symptoms, may have decided against reporting their status to the authorities, for example, due to aversion to home isolation, fear of being stigmatized, or concerns with respect to a necessary disclosure of detailed information about their personal contacts. In contrast, participants in our study may have been more honest in reporting previous positive test results due to the strictly anonymous survey situation, and the methodology applied to guarantee the confidentiality of individual answers, especially in the CDTRM condition. These factors could contribute to discrepancies between survey estimates and official statistics, with survey data potentially resulting in higher prevalence estimates such as those observed in the current study.

Our results clearly indicate that the CDTRM is effective in controlling for the validity-threatening influence of SDB in surveys on sensitive topics, and that its superiority in validity over conventional DQ as well as the utility of the CDTRM cheating detection mechanism increase with higher sensitivity of the topic under study. Notably, estimating the CDTRM cheating rate also allows for identifying subgroups in which the pressure of SDB towards choosing a self-protective response is particularly high.

Limitations and future research directions

One important limitation of our study is that the different levels of sensitivity prescribed to the three topics under investigation was based on our a priori assumptions, rather than tested empirically (e.g., in a dedicated pre-study). To this end, we were unable to quantify the extent to which our survey participants actually regarded having received at least one dose of the COVID-19 vaccine (Topic 1) as low in sensitivity, having at least once been tested positive for COVID-19 (Topic 2) as medium sensitive, and having at least once intentionally concealed a positive COVID-19 test result from others (Topic 3) as highly sensitive. Specifically, with respect to Topic 1, we explicitly acknowledge that the low sensitivity assumed in the current study due to the socio-political climate in Taiwan cannot be generalized to any population, as the individual vaccination status has been shown to be (more) sensitive in other societies, such as in Germany [8]. However, notably, a specific quantification of the sensitivity of the three topics under investigation was neither an aim of the current study, nor necessary for testing our main research questions. Due to the experimental design, the differences observed between DQ and CDTRM conditions are most likely attributable to a successful control of SDB by the CDTRM; this assumption is further supported by the expected and observed positive association of these differences, and of the CDTRM cheating rate, with the assumed sensitivity of the three topics under study. Whether our findings can be replicated for other sensitive attributes and in different populations should be the subject of future research projects.

The current study shares a second limitation with any study employing an indirect questioning technique including a cheating detection extension such as the CDM or the CDTRM: the status of cheaters in the CDTRM condition with respect to the sensitive attribute under study remains explicitly unknown. To this end, the estimated share of cheaters could be entirely comprised of carriers, entirely comprised of noncarriers, or comprised of a mixture of both groups (which currently seems most likely given the findings of [25]). Consequently, in applications of the CDTRM such as the current one, an estimate for the true prevalence of the sensitive attribute is only provided in terms of an interval ranging from the lower bound (if none of the cheaters were carriers of the sensitive attribute: $\widehat{\pi}_L = \widehat{\pi}$) to the upper bound estimate (if all cheaters were carriers: $\widehat{\pi}_U = \widehat{\pi} + \widehat{\gamma}$). Compared to other IQTs providing a single estimate, the comparatively broad CDTRM interval could sometimes be at odds with the requirements for high precision in prevalence estimation, especially when the cheating rate is high. Additionally, estimating the cheating rate in the CDTRM does also not allow any inferences about participants' motivation for disregarding the instructions. Plausibly, some participants may have intentionally decided to provide a self-protective response; others may have failed to understand the instructions and provided a self-protective response without the specific intention to do so. Our data indeed suggest that both intentional and unintentional cheating may have occurred, while this assumption cannot be explicitly tested within the CDTRM framework. These apparent weaknesses of the CDTRM may lead researchers to consider IQTs that include more explicit assumptions about the status of cheaters. However, empirical studies have shown that such stronger assumptions in models such as, for example, the Stochastic Lie Detector [28] are often violated and may result in potentially invalid estimates [44]. The obvious alternative of not accounting for cheaters at all also seems rather unreasonable given that the estimated cheating rate in the current study, especially for the sensitive Topics 2 and 3, was well above zero. Therefore, despite its apparent drawbacks, the CDTRM currently appears to be a reasonable choice among the available models offering a cheating detection extension.

Related to this point, it should also be noted that other models with a cheating detection mechanism have recently been proposed. For example, a refined version of the Unrelated Question Randomized Response Model [45], the Unrelated Question Model with Cheating Extension (UQMC) [23], allows for the detection of cheaters in a way similar to the CDM, or CDTRM. The UQMC has been empirically shown to outperform its predecessor model without cheating detection in a survey on intimate partner violence victimization and perpetration during the COVID-19 pandemic [31]. Following a different approach, the Extended Crosswise Model (ECWM) [46], an extension of the Crosswise Model [32] with a mechanism to detect instruction nonadherence in the sample, has been researched intensively over the past few years. In contrast to the CDTRM or the UQMC, the ECWM does not allow for a quantitative assessment of the share of participants cheating on the instructions. Instead, the model aims at minimizing the proportion of nonadherent respondents by providing particularly simple instructions and symmetric response options. In the context of the COVID-19 pandemic, the ECWM has been successfully applied in a survey on the socially desirable attribute of personal hand hygiene, in which it obtained a presumably more valid prevalence estimate than conventional DQ [21]. In this light, both the UQMC and the ECWM appear to be promising alternatives to the CDTRM in sensitive surveys.

In our opinion, now that the CDTRM has repeatedly been shown to successfully control for the influence of SDB and provide valuable additional information in terms of the estimated cheating rate (see the current study and [25]), the model should be further explored with respect to its capability of obtaining valid estimates for other sensitive attributes. With regard to the COVID-19-related health behaviors surveyed in the current study, it is still likely that their sensitivity will decrease further over the course of time. The more time that has passed since the acute phase of the pandemic, the lower the actual participants' perceived probability of being subject to social or legal sanctions for past misbehavior should be. Accordingly, self-reports should also be less and less influenced by SDB with increasing temporal distance from the end of the pandemic.

A key limitation of the current study pertains to the type of cheating behavior that can be detected via the CDTRM (and the predecessor CDM). Due to the central model assumptions, honest respondents strictly follow the instructions, while cheaters always respond with "none true". However, other forms of cheating, such as partial (dis-)honesty or misunderstanding of the instructions, could lead to misclassifications of participants, affecting the overall accuracy of results. This issue is notably not an exclusive threat to the CDTRM, but to all questioning techniques including DQ and other IQTs. The results reported in the current study are valid only to the extent that the central model assumptions of the CDTRM hold. Future research should therefore critically evaluate these assumptions, and work towards models that are capable of detecting a wider range of different types of cheating behavior.

Finally, the results of the current study leave open whether cheating behavior is influenced not only by the sensitivity of the topic but also by the order of the questions, especially when similar, nonsensitive statements are used for randomization. In this respect, the repeated use of birthday randomizers could possibly promote distrust in the procedure. To address this issue, [47] suggested replacing birthday randomizers with number sequence randomizers to potentially increase trust among participants and prevent evasive responses. In future implementations of the CDTRM involving multiple questions, it would be interesting to explore the usefulness of such number sequence randomizers, and to compare their performance with birthday randomizers in terms of perceived trust, comprehension, and the validity of the prevalence estimates obtained.

Conclusions

The findings of the current study support the assumption that the CDTRM successfully controlled for the validity-threatening influence of social desirability bias in a survey on three COVID-19-related health behaviors. As expected, the advantage of the CDTRM in estimation validity over a conventional DQ survey, as well as the utility of the CDTRM's cheating detection mechanism, increased with higher topic sensitivity. In summary, our results suggest that IQTs in general, as well as the CDTRM in particular, are a promising means to increase the validity of prevalence estimates based on self-reports for health-related behaviors, and for all sensitive personal attributes for which a strong influence of social desirability bias can be expected. To this end, techniques such as the CDTRM can help to better inform political and societal decisions in the context of current and future pandemics, in cases in which the adherence to non-pharmaceutical interventions in general is important, as well as in any other situation in which an accurate knowledge of the prevalence of sensitive attributes is essential.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-024-19819-6.

Supplementary Material 1.

Authors' contributions

Shu-Hui Hsieh conceptualized and obtained funding for the study, coordinated data collection, and performed data analyses. Pier Francesco Perri and Adrian Hoffmann contributed to the formulation and the design of the questionnaire and the survey plan, and to the interpretation of the results obtained. Shu-Hui Hsieh, Pier Francesco Perri, and Adrian Hoffmann contributed to the written documentation of the study conducted in the form of the current manuscript and provided critical comments to earlier drafts. All authors read and approved the final manuscript.

Funding

The author(s) disclose receipt of the following financial support for the research, authorship, and/or publication of this article:
- The research of S.H. Hsieh was supported by the National Science and Technology Council of Taiwan, ROC (111- 2410-H-001-046-SS2);
- The research of P.F. Perri was supported by MUR-PRIN 2022, grant number 2022XRHT8R - The SMILE project (Statistical Modelling and Inference to Live the Environment).

Availability of data and materials

Data and materials will be made available from the authors upon request.

Declarations

Ethics approval and consent to participate

This study received approval from the Institutional Review Board of Humanities and Social Sciences Research (No. AS-IRB-HS07-111165) at Academia Sinica, Taiwan. In this study, informed consent was waived by the Institutional Review Board of Humanities and Social Sciences Research since it was retrospective, and the personal information in the data set was anonymized.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Center for Survey Research, Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, Taiwan. [2]Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Arcavacata di Rende, CS, Italy. [3]Department of Experimental Psychology, University of Duesseldorf, Duesseldorf, Germany.

Received: 11 November 2023 Accepted: 16 August 2024
Published online: 17 September 2024

References

1. Chen T, Lucock M. The mental health of university students during the COVID-19 pandemic: An online survey in the UK. PLoS ONE. 2022;17(1):e0262562. https://doi.org/10.1371/journal.pone.0262562.
2. Coley RL, Carey N, Baum CF, Hawkins SS. COVID-19 vaccinations and mental health among US adults: Individual and spillover effects. Soc Sci Med. 2023:116027. https://doi.org/10.1016/j.socscimed.2023.116027.
3. Gao H, Hu R, Yin L, Yuan X, Tang H, Luo L, et al. Knowledge, attitudes and practices of the Chinese public with respect to coronavirus disease (COVID-19): An online cross-sectional survey. BMC Public Health. 2020;20:1–8. https://doi.org/10.1186/s12889-020-09961-2.
4. Hlatshwako TG, Shah SJ, Kosana P, Adebayo E, Hendriks J, Larsson EC, et al. Online health survey research during COVID-19. Lancet Digit Health. 2021;3(2):e76–7. https://doi.org/10.1016/S2589-7500(21)00002-9.
5. Singh S, Sagar R. A critical look at online survey or questionnaire-based research studies during COVID-19. Asian J Psychiatr. 2021;65. https://doi.org/10.1016/j.ajp.2021.102850.
6. Ziauddeen N, Gurdasani D, O'Hara ME, Hastie C, Roderick P, Yao G, et al. Characteristics and impact of Long Covid: Findings from an online survey. PLoS ONE. 2022;17(3):e0264331. https://doi.org/10.1371/journal.pone.0264331.
7. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull. 2007;133(5):859–83. https://doi.org/10.1037/0033-2909.133.5.859.
8. Wolter F, Mayerl J, Andersen HK, Wieland T, Junkermann J. Overestimation of COVID-19 vaccination coverage in population surveys due to social desirability bias: Results of an experimental methods study in Germany. Socius. 2022;8. https://doi.org/10.1177/23780231221094749.
9. Tourangeau R, Smith TW. Asking sensitive questions: The impact of data collection mode, question format, and question context. Public Opin Q. 1996;60(2):275–304. https://www.jstor.org/stable/2749691.
10. Groves RM, Fowler Jr FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. Survey methodology. John Wiley & Sons; 2004.
11. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. J Am Stat Assoc. 1965;60(309):63–9. https://doi.org/10.2307/2283137.
12. 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.
13. Fox JA, Tracy PE. Randomized response: A method for sensitive surveys. Sage Publications; 1986.
14. Chaudhuri A, Mukerjee R. Randomized response: Theory and techniques. Sage Publications; 1987.
15. Chaudhuri A. Randomized response and indirect questioning techniques in surveys. CRC Press; 2016.
16. Chaudhuri A, Christofides TC. Indirect questioning in sample surveys. Springer Science & Business Media; 2013.
17. Tian GL, Tang ML. Incomplete categorical data design: Non-randomized response techniques for sensitive questions in surveys. CRC Press; 2013.
18. Chaudhuri A, Christofides T, Rao C. Handbook of Statistics 34, Data gathering, analysis and protection of privacy through randomized response techniques. Elsevier; 2016.
19. Fox JA. Randomized response and related methods: Surveying sensitive data. Sage Publications; 2015.
20. Chaudhuri A, Pal S, Patra D. Randomized response techniques. Certain thought-provoking aspects. Springer; 2024.
21. 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.
22. Kaufmann TH, Lilleholt L, Böhm R, Zettler I, Heck DW. Sensitive attitudes and adherence to recommendations during the COVID-19 pandemic: Comparing direct and indirect questioning techniques. Personal Individ Differ. 2022;190:111525. https://doi.org/10.1016/j.paid.2022.111525.
23. Reiber F, Pope H, Ulrich R. Cheater detection using the unrelated question model. Sociol Methods Res. 2023;52(1):389–411. https://doi.org/10.1177/0049124120914919.
24. Becher M, Stegmueller D, Brouard S, Kerrouche E. Ideology and compliance with health guidelines during the COVID-19 pandemic: A comparative perspective. Soc Sci Q. 2021;102(5):2106–23. https://doi.org/10.1111/ssqu.13035.
25. Meisters J, Hoffmann A, Musch J. A new approach to detecting cheating in sensitive surveys: The cheating detection triangular model. Sociol Methods Res. 2024;53(1):328–68.
26. Hoffmann A, Waubert de Puiseau B, Schmidt AF, Musch J. On the comprehensibility and perceived privacy protection of indirect questioning techniques. Behav Res Methods. 2017;49:1470–1483. https://doi.org/10.3758/s13428-016-0804-3.

27. Clark SJ, Desharnais RA. Honest answers to embarrassing questions: Detecting cheating in the randomized response model. Psychol Methods. 1998;3(2):160–80. https://doi.org/10.1037/1082-989X.3.2.160.
28. Moshagen M, Musch J, Erdfelder E. A stochastic lie detector. Behav Res Methods. 2012;44:222–31. https://doi.org/10.3758/s13428-011-0144-2.
29. Ostapczuk M, Moshagen M, Zhao Z, Musch J. Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. J Educ Behav Stat. 2009;34(2):267–87. https://doi.org/10.3102/1076998609332747.
30. Moshagen M, Musch J, Ostapczuk M, Zhao Z. Brief report: Reducing socially desirable responses in epidemiologic surveys: An extension of the randomized-response technique. Epidemiology. 2010;21(3):379–82. https://doi.org/10.1097/EDE.0b013e3181d61dbc.
31. Reiber F, Bryce D, Ulrich R, Self-protecting responses in randomized response designs: A survey on intimate partner violence during the coronavirus disease 2019 pandemic. Sociol Methods Res. 2019;2022. https://doi.org/10.1177/004912412110431.
32. Yu JW, Tian GL, Tang ML. Two new models for survey sampling with sensitive characteristic: Design and analysis. Metrika. 2008;67:251–63. https://doi.org/10.1007/s00184-007-0131-x.
33. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Methodol. 1977;39(1):1–38. https://www.jstor.org/stable/2984875.
34. Bourke PD, Moran MA. Estimating proportions from randomized response data using the EM algorithm. J Am Stat Assoc. 1988;83(404):964–8.
35. Ulrich R, Schröter H, Striegel H, Simon P. Asking sensitive questions: A statistical power analysis of randomized response models. Psychol Methods. 2012;17(4):623. https://doi.org/10.1037/a0029314.
36. Heck DW, Moshagen M. RRreg: An R package for correlation and regression analyses of randomized response data. J Stat Softw. 2018;85:1–29. https://doi.org/10.18637/jss.v085.i02.
37. Larsen M, Nyrup J, Petersen MB. Do survey estimates of the public's compliance with COVID-19 regulations suffer from social desirability bias? J Behav Public Adm. 2020;3(2):1–9. https://doi.org/10.30636/jbpa.32.164.
38. Munzert S, Selb P. Can we directly survey adherence to non-pharmaceutical interventions? Evidence from a list experiment conducted in Germany during the early Corona pandemic. Surv Res Methods. 2020;14(2):205–9. https://doi.org/10.18148/srm/2020.v14i2.7759.
39. Timmons S, McGinnity F, Belton C, Barjaková M, Lunn P. It depends on how you ask: Measuring bias in population surveys of compliance with COVID-19 public health guidance. J Epidemiol Commun Health. 2021;75(4):387–9. https://doi.org/10.1136/jech-2020-215256.
40. Daoust JF, Nadeau R, Dassonneville R, Lachapelle E, Bélanger É, Savoie J, et al. How to survey citizens' compliance with COVID-19 public health measures: Evidence from three survey experiments. J Exp Polit Sci. 2021;8(3):310–7. https://doi.org/10.1017/XPS.2020.25.
41. Tan J, Yoshida Y, Ma KSK, Mauvais-Jarvis F. Gender differences in health protective behaviors during the COVID-19 pandemic in Taiwan: An empirical study. MedRxiv. 2021. https://doi.org/10.1101/2021.04.14.21255448.
42. O'Connor AM, Evans AD. Dishonesty during a pandemic: The concealment of COVID-19 information. J Health Psychol. 2022;27(1):236–45. https://doi.org/10.1177/1359105320951603.
43. Alvarez E, Bielska IA, Hopkins S, Belal AA, Goldstein DM, Slick J, et al. Limitations of COVID-19 testing and case data for evidence-informed health policy and practice. Health Res Policy Syst. 2023;21(1):11. https://doi.org/10.1186/s12961-023-00963-1.
44. Hoffmann A, Musch J. Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model. Behav Res Methods. 2016;48:1032–46. https://doi.org/10.3758/s13428-015-0628-6.
45. Greenberg BG, Abul-Ela ALA, Simmons WR, Horvitz DG. The unrelated question randomized response model: Theoretical framework. J Am Stat Assoc. 1969;64(326):520–539. https://www.jstor.org/stable/2283636.
46. Heck DW, Hoffmann A, Moshagen M. Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. Behav Res Methods. 2018;50:1895–905. https://doi.org/10.3758/s13428-017-0957-8.
47. Sayed KH, Cruyff MJ, van der Heijden PG, Petróczi A. Refinement of the extended crosswise model with a number sequence randomizer: Evidence from three different studies in the UK. PLoS ONE. 2022;17(12):e0279741.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.