Erschienen in: Public Opinion Quarterly ; 85 (2021), 3. - S. 836-863
https://dx.doi.org/10.1093/poq/nfab043

FALSE POSITIVES AND THE "MORE-IS-BETTER" ASSUMPTION IN SENSITIVE QUESTION RESEARCH
NEW EVIDENCE ON THE CROSSWISE MODEL AND THE ITEM COUNT TECHNIQUE

FELIX WOLTER*
ANDREAS DIEKMANN

Abstract Several special questioning techniques have been developed in order to counteract misreporting to sensitive survey questions, for example, on criminal behavior. However, doubts have been raised concerning their validity and practical value as well as the strategy of testing their validity using the "more-is-better" assumption in comparative survey experiments. This is because such techniques can be prone to generating false positive estimates, that is, counting "innocent" respondents as "guilty" ones. This article investigates the occurrence of false positive estimates by comparing direct questioning, the crosswise model (CM), and the item count technique (ICT). We analyze data from two online surveys ($N = 2{,}607$ and 3,203) carried out in Germany and Switzerland. Respondents answered three questions regarding traits for which it is known that their prevalence in reality is zero. The results show that CM suffers more from false positive estimates than ICT. CM estimates amount to up to 15 percent for a given true value of zero. The mean of the ICT estimates is not significantly different from zero. We further examine factors causing the biased estimates of CM and show that speeding through the questionnaire (random answering) and problems with the measurement procedure—namely regarding the unrelated questions—are responsible. Our findings suggest

FELIX WOLTER is a postdoc researcher in the Cluster of Excellence "The Politics of Inequality" and the Sociology Department at the University of Konstanz, Konstanz, Germany. ANDREAS DIEKMANN is a senior professor at the Eidgenössische Technische Hochschule (ETH) Zurich, Zurich, Switzerland, and at the University of Leipzig, Leipzig, Germany. The authors are very thankful to the reviewers for their helpful comments and to Marc Höglinger for his valuable advice in preparing the surveys that underlie this study. This work was supported by the German Research Foundation [DFG; project WO 2242/1-1 to F.W.], and by an ETH research grant to A.D. [Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder – EXC-2035/1 – 390681379]. *Address correspondence to Felix Wolter, Sociology; Cluster of Excellence "The Politics of Inequality," University of Konstanz, Box 32, D-78457 Konstanz, Germany; e-mail: felix.wolter@uni-konstanz.de.

that CM is problematic and should not be used or evaluated without the possibility of accounting for false positives. For ICT, the issue is less severe.

Introduction

Since its very beginning, modern survey methodology has been preoccupied with the problem of how to counteract misreporting to sensitive survey questions—for instance, questions on criminal behavior, substance abuse, or voter turnout. The most prominent starting points of research on special questioning techniques addressing this issue are the papers by Barton (1958, although written in a certain tongue-in-cheek style) and the paper by Warner (1965), which presented the randomized response technique (RRT). The principle of RRT consists of anonymizing respondents' answers to survey questions by introducing a random device administered by the interviewee; the outcome of this random mechanism is added to the data and conceals the "real" answer to the survey question. On an aggregate level, however, the distribution of the random device is known, which makes it possible to calculate estimates of the sensitive item that is of interest.

Since then, a vast amount of methodological literature on sensitive question techniques has emerged. One strand of literature has concentrated on further developing existing or proposing new anonymizing techniques. Important propositions in this regard are the crosswise model RRT (CM; Yu, Tian, and Tang 2008) and the item count technique (ICT; Droitcour et al. 1991). CM is a derivative of RRT (and statistically equal to it), making use of unrelated questions with known incidence rates as a random device (such as the month of birth of the respondent's mother). Respondents report whether their (yes-no) answers to the sensitive question *and* the unrelated question are equal or unequal. Because the individual value of the unrelated question is unknown to the interviewer and researcher, the answer to the sensitive question remains concealed. However, a prevalence estimate of the sensitive item can be calculated by taking account of the known aggregate distribution of the unrelated question.[1]

ICT employs item lists with combinations of binary non-key (filler) items and one sensitive item, with respondents only indicating a summary of their answers to the whole item list and not to individual items. Again, the individual answer to the sensitive item is concealed. The sample is randomly split into two subgroups. One group (the short-list group) receives a list containing only non-key items; the other group (the long-list group) answers an

1. In the Supplementary Material, we provide more detailed descriptions and formulae on CM and ICT.

item list containing the same non-key items plus the sensitive item. A prevalence estimate of the sensitive item is obtained by subtracting the mean of the short-list group from the one of the long-list group. A growing body of recent methodological literature has devoted itself to studying design issues, diagnostic tests, and different estimation strategies related to ICT (e.g., Blair and Imai 2012; Aronow et al. 2015; Ahlquist 2018; Hinsley et al. 2019; see the Supplementary Material for a discussion).

Another strand of literature has dedicated itself to conducting empirical studies investigating the performance of sensitive question techniques in mitigating response bias to sensitive questions. One approach to doing so are external validation studies, which examine differences between estimates obtained using one of these techniques and known true values from external records. These kinds of studies, however, are rare because external validation data are usually not available. For this reason, most empirical studies rely on experimental comparisons between conventional direct questioning (DQ) and techniques like RRT, CM, or ICT. If estimates are higher for items where under-reporting is expected (e.g., criminal behavior), they are considered to be more valid. This rationale is referred to as the "more-is-better" assumption. The opposite ("less-is-better") holds for questions where over-reporting is expected (e.g., voter turnout).

Although the findings of these studies are often mixed, meta-analyses of RRT and ICT studies generally arrive at the conclusion that, on average, these techniques indeed outperform DQ (Lensvelt-Mulders et al. 2005; Blair, Coppock, and Moor 2020; Ehler, Wolter, and Junkermann 2020). With respect to CM, no meta-analysis has been published as far as we are aware. However, several experimental studies have found remarkably positive results regarding its performance (e.g., Jann, Jerke, and Krumpal 2012; Hoffmann et al. 2015; Gingerich et al. 2016; Hoffmann and Musch 2016; Höglinger, Jann, and Diekmann 2016). Hoffmann et al. (2015, p. 409) conclude that CM "appears to be a very promising indirect questioning technique that can be used to successfully control for social desirability on surveys of sensitive behavior." Jann, Jerke, and Krumpal (2012, p. 183) state that CM "seems to be a promising alternative to conventional RRT variants." Hoffmann and Musch (2016, p. 1042) allege that CM "offers a valid and useful means for achieving the experimental control of social desirability."

However, recent studies have spoiled the party regarding CM and ICT. These have called into question the entire research field of more-is-better comparisons, and hence, the usefulness of these special questioning techniques generally. The core objection is that CM and ICT generate false positive estimates of sensitive behavior or traits: that is, respondents *not* having engaged in socially undesirable behavior are wrongly estimated as having done so. For example, Höglinger and Diekmann (2017) examine two items with practically zero prevalence in reality, namely having received an engrafted

organ (heart, lung, etc.) and having suffered from Chagas disease; they report CM prevalence estimates of 8 and 5 percent, respectively. Höglinger and Jann (2018), using individual validation data, find false positive rates of more than 10 percent for CM estimates. With respect to ICT, Riambau and Ostwald (2020) report a prevalence rate of 12 percent for a placebo item on having been invited to have dinner with the prime minister of Singapore. They further conduct a small-scale meta-analysis on previous literature (Holbrook and Krosnick 2010; Ahlquist, Mayer, and Jackman 2014; Kiewiet de Jonge and Nickerson 2014) and conclude, although these studies arrive at inconsistent conclusions, that "inflation is more likely than not" (Riambau and Ostwald 2020, p. 2). Kuhn and Vivyan (2020) present an individual-level validation study based on reported voter turnout and official turnout records in a New Zealand and a London (UK) sample. Their results are alarming with respect to the validity of ICT and the "more-is-better" (or in this case: "less-is-better") assumption: not only does ICT not alleviate strategic misreporting to the turnout question as compared to DQ (i.e., reduce false negative reports among the real nonvoters), it actually boosts false positive reports (i.e., real voters reporting they *did not* vote). Further, the authors show that according to a conventional "more-is-better" logic on the aggregate level, ICT estimates would falsely appear to be more valid (closer to the true aggregate value) than DQ.

The consequence of this literature is clear-cut: If prevalence estimates obtained via CM or ICT generally suffer from false positives, then comparisons between question formats that rely on the more-is-better assumption are misleading, because the apparently more valid (higher) estimates are—at least partially—owed to false positives, that is, by counting "innocent" respondents as "guilty" ones. Therefore, CM and ICT (and maybe other questioning techniques as well) would not improve measurement validity as compared to conventional questioning techniques, but rather do the very reverse. Hence, the decades-old research area on sensitive questions based on the more-is-better approach would in fact be highly questionable and a dead-end street.

There are, however, some open issues in this debate. First, the findings with respect to false positives should be replicated and based on a sounder empirical basis before totally abandoning the techniques (and the more-is-better assumption). One also has to distinguish between the "genuine" model and the specific measurement technique, for example, the choice of the unrelated questions with CM. For example, in the study by Höglinger and Diekmann (2017), false positives for "received donated organ" and Chagas disease went down from 8 to 6 percent, and from 5 to 1 percent, respectively, when unrelated questions that deviated empirically from the theoretically expected values were dropped. Second, pertaining to CM, Schnapp (2019) has recently proposed procedures that aim to adjust CM estimates for false

positives. The idea is to empirically estimate the fraction of respondents answering CM questions randomly, which in turn yields false positives and causes estimates to be biased toward a prevalence rate of 50 percent. If this fraction is correctly estimated, the false positive part of the CM estimates can be identified and used to adjust the calculation of the prevalence estimates. This is implemented by directly asking respondents whether they have answered the CM items properly or not (i.e., rushed over them).[2] Although Schnapp (2019) carries out some empirical tests of his propositions (with the result that they are not very helpful), the sample size of only about $n = 100$ students does not permit inferring robust conclusions. Third, as far as we know, no study has ever undertaken a direct experimental comparison of CM, ICT, and DQ with respect to false positives. Höglinger and Jann (2018) compare DQ, CM, and two variants of the classic RRT (with the result that CM does and the classic RRT does not suffer from false positives), but no evidence exists regarding a relative assessment of ICT.

The present article aims to add empirical evidence to the discussion. Using data from two online surveys in Germany ($n = 2{,}607$) and Switzerland ($n = 3{,}203$), we experimentally compare DQ, CM, and ICT estimates of several zero-prevalence items with respect to false positives. First, we will replicate and extend the study of Höglinger and Diekmann (2017) with respect to CM (as compared to DQ). Our results confirm their findings and show that CM suffers considerably from false positives. Second, we investigate whether ICT has problems with false positives in this scenario as well. Our findings show that in five out of six test situations, ICT does not suffer from false positives—this contradicts the findings of the above-cited studies. Third, we provide results for Schnapp's (2019) proposal by empirically testing the possibility of adjusting for random answers in the CM. The results suggest that we cannot recommend the proposed procedures. Fourth, we examine design- and respondent-specific factors affecting the occurrence of false positives with CM. We empirically demonstrate that deviations in the unrelated questions from their theoretically expected prevalence rates, as well as careless answering (speeding) by respondents, are responsible.

The contribution of our study is relevant both for applied researchers who intend to use CM or ICT for measuring sensitive items, and for methodological researchers interested in investigating the validity of these techniques. Our main practical recommendations are that researchers should deal with CM only with the greatest care and caution. With respect to ICT, our results contradict those by Kuhn and Vivyan (2020) and Riambau and Ostwald (2020)—we do not find substantial rates of false positives and hence argue that the use of ICT is less problematic than the use of CM. This, however, should be further investigated in future studies. Also, methodological

2. In the Methods section and the Supplementary Material, we provide more detailed information.

researchers should not base their argumentation too naively on the "more-is-better" assumption and always bear in mind that differences between different questioning techniques, such as DQ, CW, RRT, and ICT, can be caused by false negative *and* by false positive responses.

Study Design, Data, and Methods

STUDY DESIGN

In order to validate and test CM and ICT for false positives, we adopt the strategy of Höglinger and Diekmann (2017) and use items for which the prevalence in the general population is practically zero. Hence, if survey estimates for these items are significantly different from zero (in a statistical and substantial sense), they can be considered as biased and as false positives. We conducted two online surveys (being almost identical as regards the questionnaires) in Germany and the German-speaking part of Switzerland. The German survey was carried out by the authors among a sample of all students of the University of Mainz. Out of 29,826 students invited by email, 2,607 completed the survey. The Swiss survey was conducted using a commercial online-access panel with $N = 3{,}203$ respondents of the German-speaking population.[3] The surveys were fielded in November and December 2019.

Both surveys were entitled "Environment, Health, and Organ Donation," and most of the questionnaire referred to these issues. Respondents were briefed, however, that the survey was also about testing new special questioning techniques. This "sensitive questions" part of the questionnaire was located approximately at the beginning of the second half of the questionnaire, after one question block on environmental issues and one on health issues and organ donation. At this point respondents were randomly allocated to one of four experimental groups: DQ format, CM format, and the short-list or long-list group of ICT, respectively. Due to the fact that the questioning techniques are different with respect to their statistical efficiency (with DQ producing the smallest standard errors and CM being more efficient than ICT), the probabilities of allocation to the experimental groups were chosen as $p = 0.13$ for DQ, $p = 0.3$ for CM, and $p = 0.283$ for the ICT short-list or long-list group, respectively. Within each group, five sensitive items were asked, of which three related to zero-prevalence items. The CM group further featured a trial question on whether respondents had attained the German or Swiss Abitur (university entrance diploma). For the student survey, we are

3. The only monetary incentive in the student survey was a lottery of five book vouchers. The Swiss online-access panel was provided by respondi and is a non-random quota (age, gender) sample of the adult population. Respondents receive a small monetary incentive of about 1 € for the completion of the survey.

able to validate this item as well, since all students (with some very few exceptions) at German universities have to pass the Abitur. The sensitive items and the Abitur question are depicted in the first part of table 1. The remainder of the article will, however, concentrate on the zero-prevalence items only; results for the items on blood donation and excessive drinking are not discussed.[4] Figure 1 presents the experimental survey design graphically.

In the DQ format, the five sensitive questions were asked directly. Afterward, respondents were asked to answer the six unrelated questions of the CM procedure directly, for which respondents were again randomly split, with equal chance, into two groups: one group received unrelated questions having low prevalence, and the other group received unrelated questions having high prevalence—see the second part of table 1. Looking at the direct answers to the unrelated questions and comparing them to the assumed distributions enables us to assess whether problems might have occurred with these questions when used in the CM procedure.

Respondents in the CM group first saw a screen introducing them to the CM procedure in detail (see Appendix for the exact wording), followed by a trial question on the Abitur diploma. Subsequently, they answered the five sensitive items, together with either low- or high-prevalence unrelated questions (which were assigned with equal probability at the respondent/questionnaire level). For each combination of the key item and the unrelated questions, respondents were asked to indicate whether their answers to the two questions were equal ("both yes or both no") or unequal ("one yes and one no"). The allocation of the unrelated questions to the sensitive items and the trial question was random. Further, the response order ("equal"—"unequal" versus "unequal"—"equal") was also varied randomly at the respondent/questionnaire level, as was the fact of whether a separate "don't know" answer option was displayed or not. The unrelated questions referring to the parents' birthdays were adopted from Höglinger and Diekmann (2017); the ones on house numbers were designed according to Benford's law (Diekmann 2012). After the key sensitive questions and following Schnapp's (2019) proposal, respondents were told that research has shown that the CM procedure confuses many respondents and that hence we would ask them to honestly indicate for each item whether they had answered correctly and carefully or whether they had clicked randomly.

4. Regarding the prevalence and incidence rates of Chagas disease and living with an engrafted organ in the general population, Höglinger and Diekmann (2017) give some figures. The incidence rate of Chagas disease in Germany is less than 0.001 percent; the upper bound of people living with a donated organ amounts to 0.16 percent in 2016 if all patients having received an organ since 1985 were still alive. The incidence rate of dengue fever in Germany was 613 cases in 2018 (Robert Koch-Institut 2019); if this is extrapolated most conservatively over the last 80 years, the prevalence rate would be 0.06 percent of the general population.

Table 1. Sensitive/zero-prevalence items and unrelated questions for CM procedure

Item	Wording
Dengue fever	Have you ever suffered from dengue fever or have you ever been infected with the dengue virus?
Received donated organ	Have you received a donated organ (kidney, heart, part of a lung or liver, pancreas)?
Chagas disease	Have you ever suffered from Chagas disease (trypanosomiasis)?
Blood donation	Have you ever donated blood?
Excessive drinking	In the last two weeks, have you consumed five or more alcoholic drinks in a row (e.g., glasses of wine, bottles of beer etc.)?
Abitur diploma (CM only)	Have you obtained the Abitur? Swiss survey: Have you obtained the Maturitätsprüfung / Matura / Abitur?
URQ 1	Is the birthday of your mother in January or February [March to December]?
URQ 2	Is the birthday of your mother between the 1st and up to and including the 6th [the 7th and up to and including the 31st] of a month?
URQ 3	Is the birthday of your father in January or February [March to December]?
URQ 4	Is the birthday of your father between the 1st and up to and including the 6th [the 7th and up to and including the 31st] of a month?
URQ 5	Is the first digit of your physical address's house number either 7, 8, or 9 [1, 2, 3, 4, 5, or 6]?
URQ 6	Is the first digit of your mother's physical address's house number either 7, 8, or 9 [1, 2, 3, 4, 5, or 6]?

NOTE.—The order of the items was random in each questionnaire, except, for technical reasons, for the sensitive items in CM format. The Abitur item is reverse-coded for analysis (1 = no Abitur). URQ = unrelated question.

The ICT procedures also first introduced respondents to the technique and featured a trial question (see Appendix). Afterward, respondents answered five item lists, either without (short-list group) or including (long-list group)

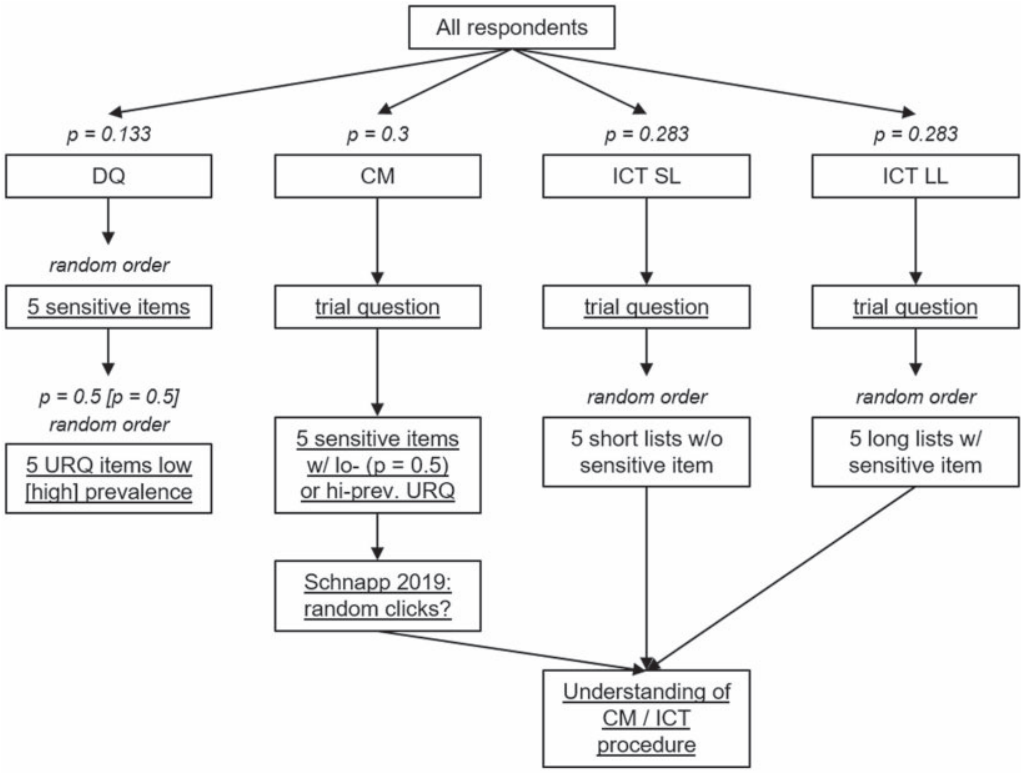

Figure 1. Experimental survey design. DQ = direct questioning; CM = crosswise model; ICT = item count technique; SL = short list; LL = long list; URQ = unrelated question.

the sensitive item. The list order and the order of the items within each list was random. Finally, in the CM and ICT format, respondents were asked how well they personally had "generally understood the special questioning technique for the previous five questions." The questionnaire also contained a short version of the Crowne-Marlowe social desirability scale for measuring the need for social approval (Crowne and Marlowe 1960; Stocké 2007).

DATA

Table 2 summarizes descriptive information about the two surveys. The response time for the whole survey shows outliers; the median time is about 15 minutes for the student survey and 13 minutes for the Swiss survey.[5] What is critical, however, is that a non-negligible fraction of respondents in the Swiss survey apparently clicked randomly through the questionnaire: the minimum response time is less than 2 minutes, and more than 5 percent of respondents took less than 7 minutes to answer the whole survey. Further,

5. This variable shows some missing values because the survey software only tracks response time for respondents who answered the questionnaire without interruptions.

Table 2. Descriptive statistics of the surveys

	German student survey				Swiss access panel survey			
	Mean	Min	Max	*n*	Mean	Min	Max	*n*
Question format								
DQ	0.121	0	1	316	0.146	0	1	466
CM	0.293	0	1	765	0.299	0	1	956
ICT SL	0.296	0	1	772	0.273	0	1	873
ICT LL	0.289	0	1	754	0.284	0	1	908
Response time (median)	14.667	6.3	3,099.4	2,498	13.417	1.9	1,166.2	3,046
Speeding in whole survey (1=yes)	0.002	0	1	2,498	0.054	0	1	3,046
Speeding on CM intro screen (1=yes)	0.272	0	1	765	0.579	0	1	956
Speeding on CM items (1=yes)	0.2	0	1	765	0.671	0	1	956
Answered CM randomly (self-report)								
Dengue fever	0.008	0	1	762	0.048	0	1	950
Donated organ	0.005	0	1	764	0.038	0	1	950
Chagas disease	0.013	0	1	764	0.038	0	1	947
Understanding of technique								
CM	4.68	1	5	763	4.538	1	5	956
ICT	4.707	1	5	1,524	4.643	1	5	1,774
Need for approval	4.102	0	7	2,601	4.709	0	7	3,201
Gender (1=female, other)	0.725	0	1	2,600	0.523	0	1	3,203
Age	24.937	18	78	2,580	47.646	15	90	3,177
Abitur diploma	0.989	0	1	2,602	0.26	0	1	3,203

NOTE.—Response time in minutes; the median difference was tested using quantile regression. Speeding in whole survey = response time less than 7 minutes. Speeding on CM intro screen = spent less than 30 seconds reading the instructions. Speeding on CM items = response time less than 90 seconds for CM items (excluding intro screen).

58 percent of respondents in the Swiss survey speeded over the CM introduction screen (took less than 30 seconds to read the instructions), and 67 percent took less than 90 seconds to answer the six CM items. These fractions are distinctively lower in the German student survey. Hence, the data of the access panel survey is probably of inferior quality and will possibly affect the results regarding false positives and the validity of CM and ICT estimates in a negative direction. We decided, however, not to attempt to exclude fast respondents or "random clickers" because this would be arbitrary in the end and would counteract the intention to test CM and ICT in a realistic scenario and not by investigating "good" survey respondents only. However, we will examine whether speeding affects the results.

Table 2 also reports on our efforts to apply Schnapp's suggestion that respondents should be asked directly whether they answered randomly. The results show that only a negligible fraction of respondents reported having clicked randomly in the student survey. In the Swiss survey, the figures are higher but still less than 5 percent. Taken together, these suggest that this approach is not useful. The results also show that subjective understanding of CM and ICT is good, showing estimates of 4.5 to 4.7 on a scale from 1 to 5. The distribution of the remaining variables in table 2 underlines that our two samples are distinct and not comparable.[6]

METHODS

The analyses first present results on false positives for zero-prevalence items by question format—separately for both surveys and then for the joint data. We also conduct overall tests for all three items simultaneously by stacking the data into long format and accounting for repeated measurements (clustered data). CM estimates are calculated based on the assumed theoretical values of the unrelated questions. Next, if false positives occur, we investigate to what extent they are associated with design-specific and personal factors. This concerns the CM data only. Design factors are the content of the unrelated question, their prevalence (low versus high), whether a "don't know" response option was displayed, and the order of the response options (equal–unequal versus the inverse). Personal factors are speeding (response time), the subjective understanding of the questioning technique, need for social approval, and basic socio-demographic variables. The formulae for calculating prevalence estimates and standard errors for the CM and ICT data are given in the Supplementary Material.

Formulae for Schnapp's (2019) proposal also appear in the Supplementary Material. The idea is to identify respondents answering randomly to CM

6. Within each survey, however, there are no differences in the distribution of socio-demographic variables by experimental groups.

questions, causing estimates to be biased. If random answers are identified, they can be used to adjust the calculation of prevalence estimates. We empirically explore the Schnapp correction and estimates excluding all "random clickers" by using the above-described direct question on whether respondents answered randomly. Because a noteworthy number of respondents said they did so only in the Swiss survey, these analyses are not carried out for the German survey.

Results

Figure 2 reports the main results of our study: namely, estimates of zero-prevalence items by question formats DQ, CM, and ICT. The first result is that CM consistently, in both surveys and for all items, produces false positive estimates of zero-prevalence items. The estimates lie between 5.2 and 14.7 percent and are all statistically different from zero and from the DQ estimates (here with the exception of the Abitur item). The overall test of all three CM items (excluding the trial question) taken together confirms the result (not displayed in figure 2; see the Supplementary Material): the global prevalence estimate for CM amounts to 10.7 percent over both surveys (standard error = 1.02), which is a highly significant difference from zero. In sum, our findings corroborate the evidence from previous studies regarding the severe flaws in CM estimates, and the use of CM in conjunction with the more-is-better assumption. This, however, is only the first part of the story, as more detailed analyses will show.

For ICT, we do not observe significant false positives for any of the items in the student survey and for two items in the Swiss survey. In the latter one, the ICT estimate for "having received a donated organ" is 9.6 percent, and we can only speculate about the reasons why this item especially yields false positives. It might be that the item content was misunderstood (as, for instance, relating to being willing to donate an organ), but in the student survey, no problems occur. Nevertheless, our overall inference regarding ICT is that, as compared to CM, it does not suffer so much from false positives. This conclusion is also supported by the global test of all items at once, which does not show any significant difference from zero in either survey and for the combined data (see the Supplementary Material). We also performed several diagnostic tests and robustness analyses with different ICT estimation strategies, which are all documented in the Supplementary Material.

Above we argued that taking into account the response times, the Swiss access panel survey is probably of inferior quality. This is also confirmed by the fact that the DQ estimates of 3, 2.4, and 1.8 percent are also too high, and significantly different from zero.[7]

7. If we exclude, for the Swiss data, respondents taking less than 7 minutes for the whole survey, the DQ estimates reduce to 1.6, 1.4, and 0.7 percent, respectively.

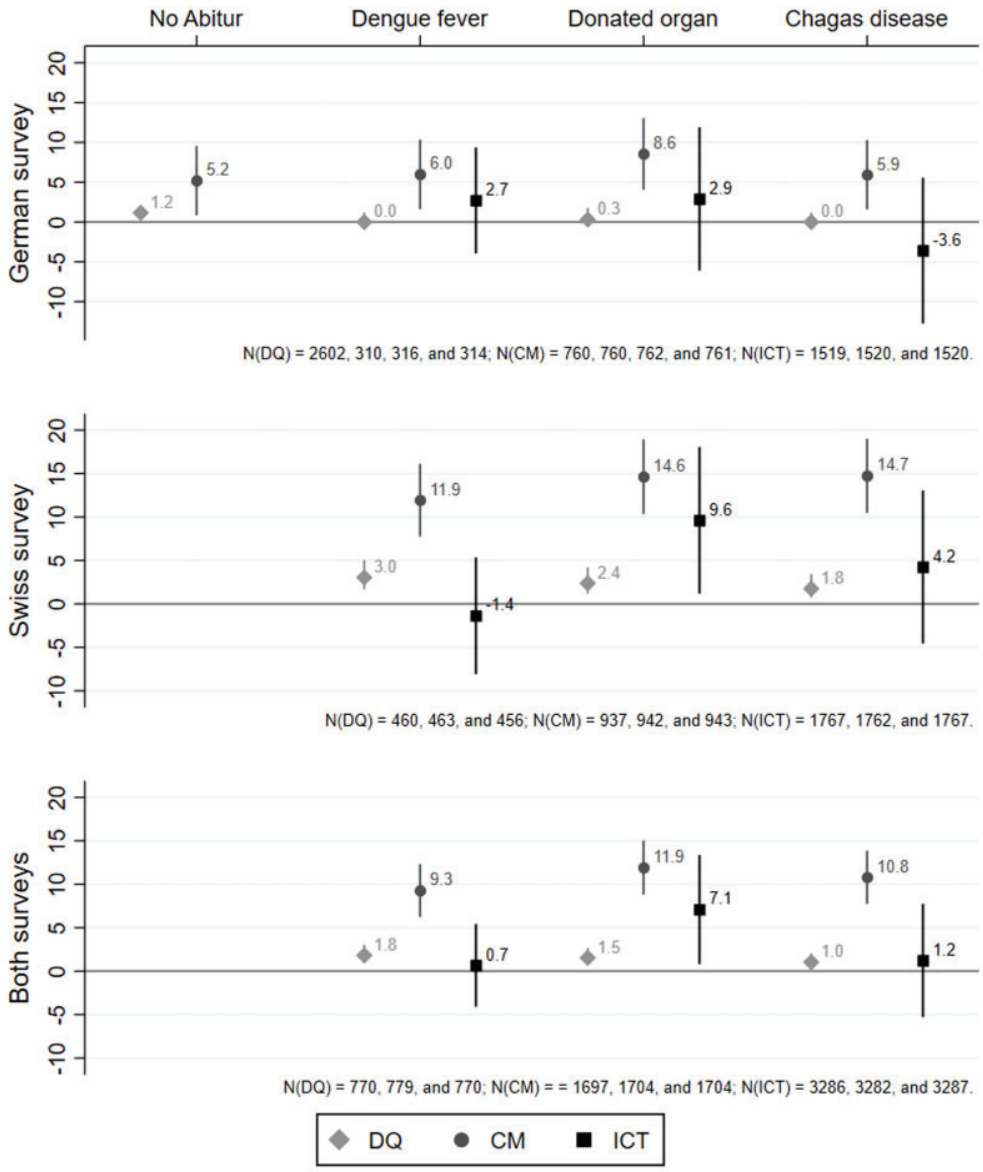

Figure 2. False positive estimates by question format. DQ = direct questioning; CM = crosswise model; ICT = item count technique. 95 percent confidence intervals (binomial exact for DQ estimates). Abitur is the German university entrance diploma.

The next question is whether Schnapp's (2019) proposals are able to remedy CM's flaws by adjusting for respondents speeding through CM items. Figure 3 depicts the respective results and shows the answer is no. Adjusting estimates for random clickers or excluding them from the estimation slightly reduces the false positive rates, but not in a statistically and substantially significant amount. Hence, at least in its actual form and implementation, asking respondents directly about their response behavior on CM items is not a proper

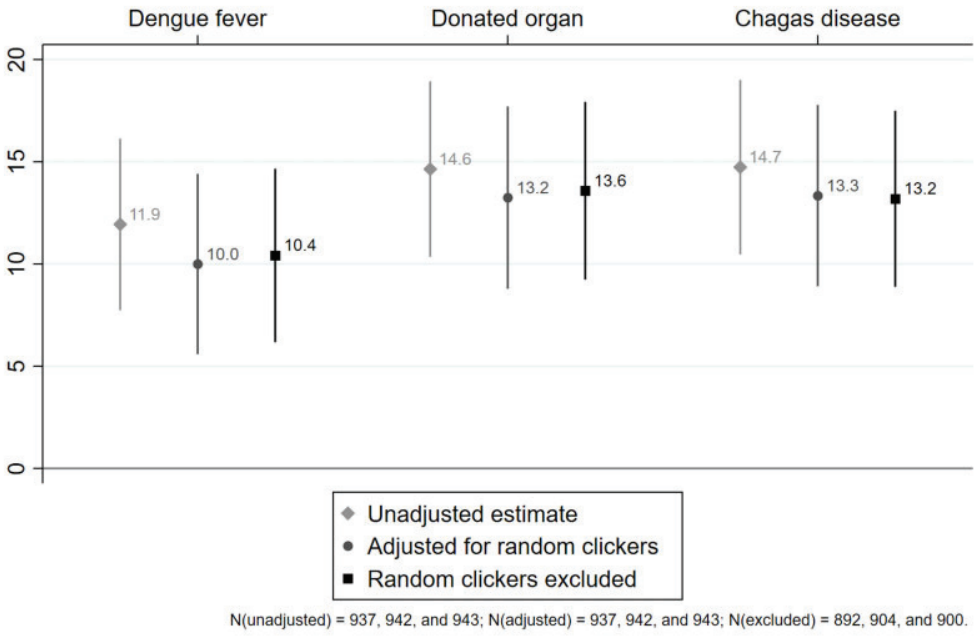

Figure 3. Adjusting CM estimates for random answers (Swiss survey only).

tool to account for CM's deficits. This, however, does not rule out that the adjustment method might work well provided a valid measurement device of the proportion of random clickers is available (see the Supplementary Material).

We now turn to the question of the causes of CM's biased estimates. As already pointed out by Höglinger and Diekmann (2017), two mechanisms can be hypothesized in this regard: namely, random answers (non-compliance with the CM instructions) and problems regarding the unrelated questions.[8] Random clicking could be caused by carelessness (e.g., speeding over the survey questions) or by confusion—that is, respondents not understanding the procedure. Further, all factors are interrelated and various mechanisms can be imagined. For example, a badly formulated unrelated question could lead some respondents to give biased answers to it, resulting in a deviation in the empirical prevalence of the unrelated question from the theoretically expected one. Alternatively, other respondents might "give up" because of their failure to understand the unrelated question, and randomly click on answers. These mechanisms also make sense in the light of a more general theoretical explanation of response behavior in surveys. The satisficing perspective (Krosnick 1991, 2000) predicts that if survey procedures become cognitively too demanding, respondents tend to avoid the costs involved with carefully running through all steps of

8. A third mechanism would be deliberate over-reporting. However, it is not realistic that up to 15 percent of respondents would deliberately report false "yes" answers to our health items in the CM format (and not in the DQ or ICT format).

Table 3. Factors affecting the occurrence of false positives in the crosswise model

	Student survey		Swiss survey		Both surveys	
	AME	(*p*-value)	AME	(*p*-value)	AME	(*p*-value)
Design variables						
URQ on father (1=yes, 0=other)	−0.009	(0.741)	0.003	(0.914)	−0.002	(0.923)
URQ on house number (1=yes, 0=other)	0.036	(0.204)	0.020	(0.443)	0.027	(0.148)
URQ on birthday (1=yes, 0=other)	0.003	(0.933)	−0.014	(0.591)	−0.006	(0.754)
URQ on month of birth (1=yes, 0=other)	−0.039	(0.147)	−0.008	(0.754)	−0.023	(0.211)
With "don't know" answer option (1=yes)	0.037	(0.185)	−0.050	(0.080)	−0.012	(0.553)
Response option order (1=equal/unequal first)	−0.040	(0.151)	0.016	(0.578)	−0.009	(0.653)
High prevalence of URQ (1=>0.8, 0=<0.2)	0.068	(0.016)	0.159	(0.000)	0.119	(0.000)
Personal variables						
Speeding in whole survey (1=yes)	n.a.	(n.a.)	0.115	(0.073)	0.160	(0.011)
Speeding on CM intro screen (1=yes)	−0.003	(0.913)	0.131	(0.000)	0.092	(0.000)
Speeding on CM items (1=yes)	−0.003	(0.940)	0.123	(0.000)	0.092	(0.000)
Understanding of CM procedure	0.006	(0.664)	−0.051	(0.000)	−0.032	(0.000)
Need for approval	0.008	(0.415)	−0.011	(0.285)	0.002	(0.824)
Gender (1=female, other)	0.052	(0.080)	−0.038	(0.189)	−0.016	(0.451)
Age	0.002	(0.084)	−0.001	(0.229)	0.001	(0.097)
No Abitur diploma (1=no Abitur)	n.a.	(n.a.)	0.046	(0.165)	n.a.	(n.a.)
N	from 2,182 to 2,283		from 2,676 to 2,822		from 4,858 to 5,105	

NOTE.—Average marginal effects (AME; reporting binary change from 0 to 1 for dummy variables) from bivariate binary logistic regressions. Significance tests (two-tailed) based on robust standard errors adjusted for the clustering of items in respondents (see the Supplementary Material for a table reporting standard errors). No "speeding in whole survey" and Abitur effect for the student survey because of too little variance in this variable. CM = crosswise model. URQ = unrelated question. The models include "URQ content variables" on father, house number, birthday, and month of birth derived from the original six URQs.

the cognitive answer process and to give cursory or even arbitrary answers that satisfice their overall aim to finish the survey.

In what follows, we investigate several potential factors that could explain the occurrence of false positives produced by CM. Table 3 depicts results from bivariate regression models[9] regarding the effects of design-specific and personal variables on the false positive rate. For the models, the data were stacked into long format; the outcome is the mean false positive rate over three zero-prevalence items for each respondent; standard errors are adjusted for the clustering in respondents.

The content of the unrelated question—that is, whether it asked about the respondent's father, about house numbers, birthdays, or months of birth—all have no effect on the emergence of false positives. Also, whether or not a "don't know" response option was displayed has no effect at conventional significance levels. The negative effect in the Swiss survey, however, is in line with the expectation that a "don't know" option reduces random answering—it is of borderline statistical significance ($p = 0.08$). In contrast, the theoretical prevalence of the unrelated question has a distinct effect in both surveys, with high-prevalence unrelated questions producing more false positives than low-prevalence ones. Although we cannot identify the definite reasons for this, there are two plausible explanations. First, the correct response for most of the respondents receiving a high-prevalence unrelated question is "unequal," because their correct answer to the zero-prevalence item is "no" and their correct answer to the unrelated question is "yes" (with probability greater than 0.8). Hence, a bias occurs if these respondents falsely report "equal" instead of "unequal." Because assessing something as unequal is cognitively more demanding than judging something as equal, respondents who are speeding over the items might be more inclined to choose the "equal" response option. Second, our discussion below concerning the empirical prevalence estimates of the unrelated questions will show that, especially in the Swiss survey, many of the high-prevalence unrelated questions are underestimated as compared to the theoretically expected values. If this is valid, then again there are too many "equal" answers in comparison to the theoretically assumed ones and the CM estimate will be biased upward.

As concerns the effects of the personal variables, there are no effects for the student survey. However, this is an interesting result as well, because it shows that speeding or a subjectively reported non-understanding of the CM procedure are not responsible for the occurrence of false positives in this survey. The effects for the Swiss survey, in contrast, corroborate our conjecture that a non-ignorable portion of the respondents speeded through the

9. The results do not change substantially in a full model containing all independent variables, with the exception that the three speeding/response time variables are intercorrelated and lose some of their effects.

Table 4. Theoretical and estimated prevalence of unrelated questions

	Theoretical value	Student survey		Swiss survey	
		Estimate	(*p*-value)	Estimate	(*p*-value)
Mother's birthday Jan–Feb	0.162	0.179	(0.577)	0.157	(0.843)
Mother's birthday 1st–6th	0.197	0.266	(0.039)	0.270	(0.005)
Father's birthday Jan–Feb	0.162	0.159	(0.907)	0.197	(0.145)
Father's birthday 1st–6th	0.197	0.269	(0.030)	0.296	(0.000)
Own house number 7–9	0.155	0.159	(0.904)	0.103	(0.027)
Mother's house number 7–9	0.155	0.153	(0.941)	0.098	(0.017)
Mother's birthday Mar–Dec	0.838	0.818	(0.476)	0.782	(0.024)
Mother's birthday 7th–31st	0.803	0.794	(0.774)	0.775	(0.293)
Father's birthday Mar–Dec	0.838	0.839	(0.958)	0.798	(0.109)
Father's birthday 7th–31st	0.803	0.821	(0.546)	0.741	(0.021)
Own house number 1–6	0.845	0.792	(0.056)	0.752	(0.000)
Mother's house number 1–6	0.845	0.725	(0.000)	0.650	(0.000)
N		from 143 to 170		from 217 to 235	

NOTE.—The Supplementary Material provides this table containing standard errors. Two-tailed z-tests of the empirical value against the theoretically expected one.

questionnaire and clicked randomly on answers. All speeding indicators have pronounced effects (the first one being significant at a 10 percent level only). Also, one point on the 5-point scale regarding understanding of the CM procedure reduces the false positive rate by 5 percentage points. One should be aware that random answers generally have large effects. Given a zero-prevalence item, the upward bias b is $^1/_2$ of r, the proportion of respondents answering randomly (Höglinger and Diekmann 2017). Thus, a proportion of, for example, 10 percent of random clickers yields a false positive rate of 5 percent.

Finally, a further possible source of CM bias are deviations in the empirical prevalence rates of the unrelated questions from the theoretically expected ones. If this is the case, calculations of CM estimates that are based on the (wrong) theoretical values are biased. Table 4 reports theoretical and empirical prevalence estimates for all unrelated questions as measured in the DQ experimental groups. Despite the low number of cases, the results show several significant differences in this regard. With respect to the high-prevalence unrelated questions, all significant differences show an underestimation of the theoretically assumed values. This supports our conjecture pointed out above regarding the positive effect of high-prevalence unrelated questions on the occurrence of false positives. Further, and most importantly, if we calculate the CM estimates using the empirical prevalence rates of the unrelated questions, false positive rates reduce dramatically. This is depicted in

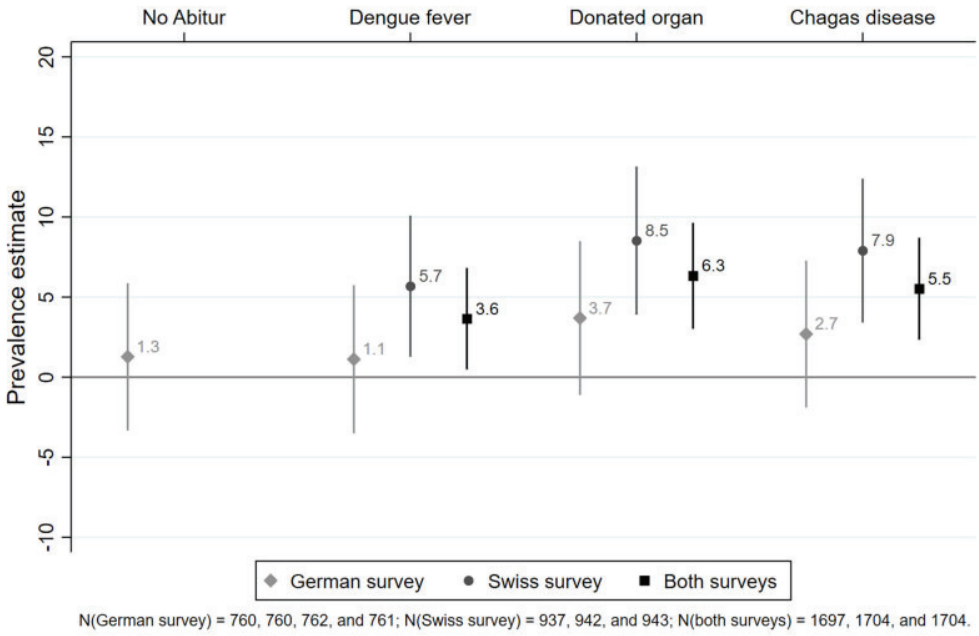

Figure 4. Prevalence estimates using empirical prevalence rates of unrelated questions. DQ = direct questioning; CM = crosswise model; ICT = item count technique. 95 percent confidence intervals (binomial exact for DQ estimates).

figure 4. CM estimates for the student survey are no longer significantly different from zero; those for the Swiss survey and the ones for the combined data still are but show considerably lower—though still non-ignorable—rates.

To sum up, the evidence in our view strongly suggests that there are two main factors causing false positives in the CM procedure: deviations in the empirical prevalence rates of the unrelated questions, and careless responses due to speeding and random clicking through the questionnaire.

Discussion

This article has investigated the occurrence of false positive estimates of zero-prevalence survey items in a sensitive question context. The four main questions were whether recent findings from the literature regarding the serious problems of CM with respect to false positives can be replicated and confirmed; whether ICT also suffers from these issues; whether a recent proposal to adjust for random answers in the CM is practicable; and to what extent design- and respondent-specific factors affect the appearance of false positives. The background of these issues is that applied researchers planning to measure sensitive traits in their surveys may choose one of the special questioning techniques in order to avoid social desirability bias and enhance

data validity. However, even if these techniques improve measurement validity with respect to under-reporting of negatively connoted traits, but also generate false positives in the sense that "innocent" respondents are flagged as "guilty" ones, their use is highly questionable. Moreover, the validity of CM, ICT, and other techniques has almost exclusively been assessed by relying on the "more-is-better" assumption. However, if questioning techniques produce false positive estimates, comparing survey estimates by question format and judging higher estimates as more valid is entirely misleading.

The main results are that CM is prone to a false positive bias. Second, ICT suffers less from this problem. Third, the recent proposal in the literature of adjusting CM estimates for random answers did not turn out to be a practical tool. Fourth, the findings concerning possible causes of CM indicate relatively clearly that deviations in the unrelated questions from their theoretically expected prevalence rates, as well as careless answering (speeding) by respondents, are responsible for the failure of CM.

There are several consequences and practical recommendations that result from our findings. These concern both researchers who intend to use one of the special questioning techniques in actual field applications, and survey methodologists who conduct studies that empirically investigate their validity. First, any researcher dealing with one of these techniques should bear in mind that false positive estimates could arise. Hence, if possible, we should build survey designs that explicitly allow checking for the occurrence of false positives. Second, the "more-is-better" assumption should not be adopted uncritically, and researchers should always take into account that differences between questioning techniques can be caused both by false negatives and by false positives. Furthermore, even external validation studies could be erroneous if the design does not allow us to account for false positives. This concerns aggregate/sample-level validations, but also individual validation studies in which the true value of the sensitive item has no variance (e.g., "guilty only" samples).

Third, the validity of CM is questionable if assumptions concerning unrelated questions and respondents' compliance with the procedures are not met. CM may be a splendid model in theory, but in practice it is very sensitive to the two sources of randomization bias: the randomization mechanism of the unrelated questions and random responses. The latter problem may be enhanced by misunderstanding and low motivation to answer the questions in an online survey. The validity of CM seems to depend very much on the population sampled. Comparing the Swiss and German sample, the method worked better with more highly educated and more motivated respondents than with a poor-quality access panel from the general population.

If used without care, CM does not improve measurement validity, but instead accomplishes the very reverse. All three studies investigating false positives in the CM (Höglinger and Diekmann 2017; Höglinger and Jann 2018; and our study), using different samples, survey settings, and CM

implementations, unanimously led to this conclusion, so the evidence in this regard is rather convincing. This also indicates that all studies exploring the validity of CM on the basis of the more-is-better assumption are prone to being misleading and should be abandoned.

Another result from our findings is that we were able to identify two main causes of the CM bias, namely random answering and biased prevalence rates of the unrelated questions. This in principle gives hope, because if these issues and the emergence of false positives are remedied, CM may be able to work as intended. Hence, our practical recommendations are that researchers should always compare theoretical prevalence rates of the unrelated questions to empirical ones, which in turn means that CM users should measure these prevalence rates for their sample empirically. Additionally, it is critical to ensure that respondents do not answer the CM items carelessly. Our findings regarding the Swiss access panel data clearly indicate that poor survey quality makes the resulting CM data more or less unusable. Hence, another recommendation is to use screener or filter questions and other means (such as tracking response latencies) that allow for identifying random answering. According to our results, however, we cannot recommend the proposition of Schnapp (2019) to ask respondents directly whether they have answered carefully or not. While the adjustment technique is promising in theory, the diagnostic of the random answers remains the Achilles's heel of the proposed procedure.

As concerns ICT, our results are more positive because we do not find substantial rates of false positives in five out of six test situations. Therefore, our findings are more favorable than the results of recent studies of this issue (see above; Kuhn and Vivyan 2020; Riambau and Ostwald 2020), which found clear evidence for false positives. Thus, it seems like this issue "sometimes" is a problem with ICT procedures and "sometimes not." Future research should dedicate itself to clarify which ICT design issues (including sample/population characteristics) stimulate the generation of false positives and how we can avoid it. One strategy in this regard would be to accumulate further empirical studies and then carry out meta-analytic methods of the empirical material.

Of course, our study has some limitations that should be accounted for in future studies. One is that we used a self-administered online survey, which generally fosters careless responding. Hence, our findings should be replicated using in-person and telephone surveys. Normally, one would expect that problems alleviate if guidance by a human interviewer is available. Also, we did not undertake any efforts to mitigate the poor quality of the Swiss access panel survey—for example, by excluding speeding respondents. One should remember, however, that excluding respondents speeding through the CM introduction screen or the CM items would have meant deleting 58 and 67 percent of respondents, respectively, which presents no solution in practice.

Further, we did not carry out cognitive interviews to check the understanding of the CM and ICT procedures. This could be a promising tool to gain further knowledge about the genesis of response bias and means of avoiding it. Another issue is that our design as a whole only allows detecting false positive estimates but does not allow us to check for the main task of sensitive question techniques, namely, reducing rates of false negative responses. If the latter dominate the former to large extents, then researchers might accept small rates of false positives. Also, by design, we are working in our study at the extreme low end of the response distribution (zero prevalence), so the question arises of how our results relate to real-world applications with more realistic and higher prevalence rates of the sensitive items. In order to account for both points, individual-level validation studies should be conducted in which both false negative and false positive rates can be compared by different question formats and without relying on zero-prevalence scenarios (see Kuhn and Vivyan 2020 for such a design).

Appendix: Questionnaire Wordings[10]

DQ GROUP

Introduction screen:

In the following pages we will ask you several rather personal questions, for which truthful answers are central to our research. We would like to point out again that your data will be treated confidentially and the data will be made anonymous, so that no conclusions about individual persons are possible. [For the wording of the sensitive items, see table 1 in the main text.]

Introduction screen to the unrelated questions asked of the DQ group:

The following six questions may seem a little strange to you. But for our research it is important that you answer them carefully and correctly. They relate to statistical methods, and concern how we try out new questioning techniques. Thank you very much!

[For the wording of the unrelated question, see table 1 in the main text.] The items asking about the respondent's mother or father additionally gave the following information:

If you don't know, please choose another person whose birthday [house number] you know.

10. Translated from German.

Appendix Table A1. Item lists for the ICT procedure

List for item	Non-key items
Blood donation	Did you have a doctor's appointment this week: yes or no? Have you ever been abroad: yes or no? Do you have a private health insurance: yes or no? [Switzerland: Do you have a Spitalzusatzversicherung (private/half private): yes or no?] Have you ever been to the cinema: yes or no?
Excessive drinking	Are you vegan: yes or no? Have you ever eaten at McDonald's: yes or no? Do you have blond hair: yes or no? Do you have life insurance: yes or no?
Dengue fever	Did you ever have a traffic accident: yes or no? Have you ever moved: yes or no? Does your house number start with the figure "8": yes or no? Did you have a dentist's appointment in the last five years: yes or no?
Received donated organ	Do you have a domestic animal now or have you had one before: yes or no? Do you own a car: yes or no? Are you a spectacles wearer: yes or no? Are you a member of a football [i.e., soccer] club: yes or no?
Chagas disease	Do you use aspirin regularly: yes or no? [Switzerland: Do you use pills against headaches regularly: yes or no?] Do you use an electric toothbrush: yes or no? Do you ride your bike regularly: yes or no? Have you ever been hospitalized: yes or no?

CM GROUP

Introduction screen (accentuation as in the German original):

In the following pages we will ask you some rather personal questions, the truthful answers to which are central to our research.

In order to ensure your privacy is unconditionally protected, we use a special survey method in which your details remain 100% secret at all times.

The whole thing works like this: You answer two questions A and B silently for yourself and then only indicate whether your answers to the two questions were. . .

- equal (both "no" or both "yes") or
- different (once "no" and once "yes").

Your answer to one of the single questions is therefore not apparent to us researchers. But for all participants of this survey taken together, we can calculate the percentage of yes and no answers in a statistically correct way.

In order for the procedure to work, we depend on you to follow the procedure exactly. Thank you for your cooperation!

On the next page we will ask you a practice question so that you can get to know the special survey method. Afterwards we will ask you five questions, which you will answer following the special method.

More detailed information about the method can be found in this Wikipedia article: [link to the Wikipedia article on CM].

Trial question screen:

The following question is a practice question. Nevertheless, please follow the procedure carefully and answer the questions truthfully.

Question A:

[Randomly allocated unrelated question]

Question B:

Have you obtained the Abitur: yes or no?

Please compare your answers to questions A and B. Are they equal or unequal?

- Equal (both "no" or both "yes")
- Unequal (once "no" and once "yes")
- [Don't know, if randomly allocated at the questionnaire level]

No matter what you ticked, your answer to a single question is not visible to anyone. Only for all participants of this survey taken together can we calculate the percentage of yes and no answers in a statistically correct way.

On the following pages, we will now ask five "real" questions on the topic of health/organ donation, which we present using this special method.

For the wordings of the sensitive items and the unrelated questions, see table 1 in the main text.

Schnapp (2019) procedure for adjusting for random clicking:
Thank you very much.

Previous research has found that the special technology being used confuses many respondents and the questions are often difficult to answer.

We would therefore ask you to honestly indicate below whether you have answered the five questions correctly and conscientiously, or whether you have just clicked on one or more questions at random.

It doesn't matter if you have clicked at random, we just want an honest answer. Your anonymity will not be affected.

Respondents then indicated for each of the five items separately whether they gave a "correct answer" or whether they "clicked at random."

ICT GROUPS

Introduction screen:

In the following pages we will ask you some rather personal questions, the truthful answers to which are central to our research.

In order to ensure your personal privacy unconditionally, we use a special survey method in which your details remain 100% secret at all times.

The whole thing works like this: We will present you with lists of five [four] yes-no questions. However, you will not answer the questions individually, but will simply state the number of questions that you have answered with "yes," i.e., a number between 0 and 5 [4]. In so doing, no one gets to know which questions you have answered with "yes." It is very unlikely that you will answer all the questions with "yes" or "no."

Your answer to one of the single questions is therefore not apparent to us researchers. But for all participants of this survey taken together, we can use special statistical methods to correctly calculate the percentage of yes and no answers statistically.

In order for the procedure to work, we depend on you to follow the procedure exactly. Thank you for your cooperation!

In the next page we will ask you a practice question so that you can become familiar with the special survey method. Afterwards we will ask you five questions, which you will answer following the special method.

More detailed information about the method can be found in this Wikipedia article: [link to the Wikipedia article on ICT].

Trial question screen:

First we will use a simple and funny example to practice this process.

How many of the following not entirely serious situations apply to you?

- Were you born on the North Pole: yes or no?
- Is your name Donald Duck: yes or no?
- Are younger than 100 years old: yes or no?
- Have you ever been on planet Mars: yes or no? [long-list group only]
- Do you live on planet Earth: yes or no?

Sum of your "yes" answers in total:

If respondents ticked a response other than "2," the following screen appeared:

Actually you should have answered with "2" because we asked you to count the "yes" answers together.

- The first question was whether you were born at the North Pole. This should be a "no."
- The second question was whether your name was Donald Duck, which would be a "no."
- The third question was whether you were less than 100 years old. That would most likely be a "yes."
- The fourth question was whether you have ever been to Mars. That would be a "no."
- And the last question was whether your place of residence is on planet Earth. Here you should have answered "yes."

There would therefore be two "yes" answers in total.

Screen for all respondents:

Wonderful. In the next pages you will find the actual question lists.

Please indicate for each list how many of the questions you have answered with "yes."

The items lists are given in Appendix table A1. In some cases, slight adaptions were necessary for the Swiss survey as compared to the German one due to country-specific peculiarities, which are indicated in square brackets. In the short-list group, the sensitive item was missing.

CM AND ICT GROUPS: QUESTION ON UNDERSTANDING OF THE QUESTIONING TECHNIQUE

What would you say, how well did you understand the principle of the special questioning technique for the last five questions overall?" with response options "understood completely," "understood rather well," "partly," "did not understand very well," "not understood at all."

CROWNE-MARLOWE SCALE FOR NEED FOR SOCIAL APPROVAL[11]

And now finally a few statements describing personal attitudes and behavior. Please indicate for each statement whether it applies to you personally ("correct") or does not apply ("incorrect").

- I sometimes feel resentful when I don't get my way.
- No matter who I'm talking to, I'm always a good listener.
- There have been occasions when I took advantage of someone.
- I'm always willing to admit it when I make a mistake.
- I always practice what I preach.
- I am always courteous, even to people who are disagreeable.
- I am sometimes irritated by people who ask favors of me.

References

Ahlquist, John S. 2018. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators." *Political Analysis* 26:34–53.

Ahlquist, John S., Kenneth R. Mayer, and Simon Jackman. 2014. "Alien Abduction and Voter Impersonation in the 2012 U.S. General Election: Evidence from a Survey List Experiment." *Election Law Journal* 13:460–75.

11. As this variable only plays a minor role for the analysis and in order to save survey time, the short Crowne-Marlowe scale (Stocké 2007) was preferred over more elaborated multidimensional scales such as the Balanced Inventory of Desirable Responding.

Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. "Combining List Experiments and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3:43–66.

Barton, Allen H. 1958. "Asking the Embarrassing Question." *Public Opinion Quarterly* 22:67–68.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to Worry About Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114:1297–315.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20:47–77.

Crowne, Douglas P., and David Marlowe. 1960. "A New Scale of Social Desirability Independent of Psychopathology." *Journal of Consulting Psychology* 24:349–54.

Diekmann, Andreas. 2012. "Making Use of 'Benford's Law' for the Randomized Response Technique." *Sociological Methods and Research* 41:325–34.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena M. Ezzati. 1991. "The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application." In *Measurement Errors in Surveys*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, 185–210. New York: Wiley.

Ehler, Ingmar, Felix Wolter, and Justus Junkermann. 2020. "Sensitive Questions in Surveys: A Comprehensive Meta-Analysis of Experimental Survey Studies on the Performance of the Item Count Technique." *Public Opinion Quarterly* 85:6–27.

Gingerich, Daniel W., Virginia Oliveros, Ana Corbacho, and Mauricio Ruiz-Vega. 2016. "When to Protect? Using the Crosswise Model to Integrate Protected and Direct Responses in Surveys of Sensitive Behavior." *Political Analysis* 24:132–56.

Hinsley, Amy, Aidan Keane, Freya A. V. St. John, and Harriet Ibbet. 2019. "Asking Sensitive Questions Using the Unmatched Count Technique: Applications and Guidelines for Conservation." *Methods in Ecology and Evolution* 10:308–19.

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

Hoffmann, Adrian, and Jochen Musch. 2016. "Assessing the Validity of Two Indirect Questioning Techniques: A Stochastic Lie Detector Versus the Crosswise Model." *Behavior Research Methods* 48:1032–46.

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

Höglinger, Marc, and Ben Jann. 2018. "More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model." *PLoS One* 13:1–22.

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

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74:37–67.

Jann, Ben, Julia Jerke, and Ivar Krumpal. 2012. "Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism." *Public Opinion Quarterly* 76:32–49.

Kiewiet de Jonge, Chad P., and David W. Nickerson. 2014. "Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys." *Political Behavior* 36: 659–82.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.

———. 2000. "The Threat of Satisficing in Surveys: The Shortcuts Respondents Take in Answering Questions." *Survey Methods Newsletter* 20:4–8.

Kuhn, Patrick, and Nick Vivyan. 2020. "The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries." *Political Analysis* (forthcoming). https://doi.org/10.1017/pan.2021.10.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, and Cora J. M. Maas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods and Research* 33:319–48.

Riambau, Guillem, and Kai Ostwald. 2020. "Placebo Statements in List Experiments: Evidence from a Face-to-Face Survey in Singapore." *Political Science Research and Methods* 9: 172–79.

Robert Koch-Institut. 2019. *Infektionsepidemiologisches Jahrbuch meldepflichtiger Krankheiten für 2018*. Berlin: Robert Koch-Institut.

Schnapp, Patrick. 2019. "Sensitive Question Techniques and Careless Responding: Adjusting the Crosswise Model for Random Answers." *Methods, Data, Analyses (MDA)* 13:307–20.

Stocké, Volker. 2007. "Deutsche Kurzskala zur Erfassung des Bedürfnisses nach sozialer Anerkennung." *ZUMA-Informationssystem. Elektronisches Handbuch sozialwissenschaftlicher Erhebunginstrumente. ZIS Version 11.00*, edited by A. Glöckner-Rist. Bonn: GESIS.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60:63–69.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. "Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis." *Metrika* 67:251–63.