Online Appendix

Online Appendix to "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT". For the full survey documentation, including questionnaire screenshots and data as well as extended analyses and analysis scripts see Höglinger and Diekmann (2016).

Contents

A. Recent individual-level validation studies

Of the handful of RRT individual-level validations published since 2000 only Höglinger and Jann (2016) and John et al. (2016) actually considered false positives in their analysis. The others surveyed only "guilty" respondents, i.e. true positives, which inhibits testing for false positives (van der Heijden et al. 2000; Moshagen et al. 2014; Wolter and Preisendörfer 2013), or used designs that allowed for identifying false positives to be identified in principle, but did not make use of this opportunity (Hoffmann et al. 2015; Kirchner 2015). This, too, indicates a profound lack of awareness of the potential occurrence of false positives in sensitive question research.

B. Design, data, and analysis details

Sample and survey details

Respondents were members of the PsyWeb-Panel, a non-representative online access panel administered by three German universities (see `https://psyweb.uni-muenster.de`). Of 10,000

members invited by email, 1,722 accessed our online questionnaire on "Organ donation and health" consisting of various questions on organ donation attitudes and behavior and containing an experimental information treatment on beliefs related to organ donation willingness.[1] After excluding one respondent who assessed his language skills (in German) as "rather poor"[2], we were left with 1,685 respondents who completed the survey part containing the sensitive questions. The median response time was 10.4 minutes, with the questionnaire version using the crosswise model taking one minute longer than the one using direct questioning. Break-off rates were almost identical for both the DQ version with 4% and the crosswise model (CM) with 5%. The sample consisted of German residents, with a median age of 47 years, 64% females, 54% married or living together with a partner, and 96% with German citizenship. Further, 46% worked full-time, 20% part-time, 5% were occasionally employed, 7% in training, and 22% not employed or on leave, while 13% were university students. Their educational background was quite above-average with 76% having completed the general or subject-specific university entrance qualification (about equivalent to a High School diploma).

The sensitive question techniques implemented

To validate the sensitive question techniques, one-third of the respondents were randomly assigned to the direct questioning (DQ) version of the sensitive questions (Figure B.1), and two-thirds to the crosswise-model variant (CM). The unbalanced assignment partly counterbalances the lower statistical efficiency of the crosswise-model RRT. The sensitive questions were preceded by a screen announcing some sensitive questions, stating the importance of honest answers for the success of the study and providing some privacy assurance.

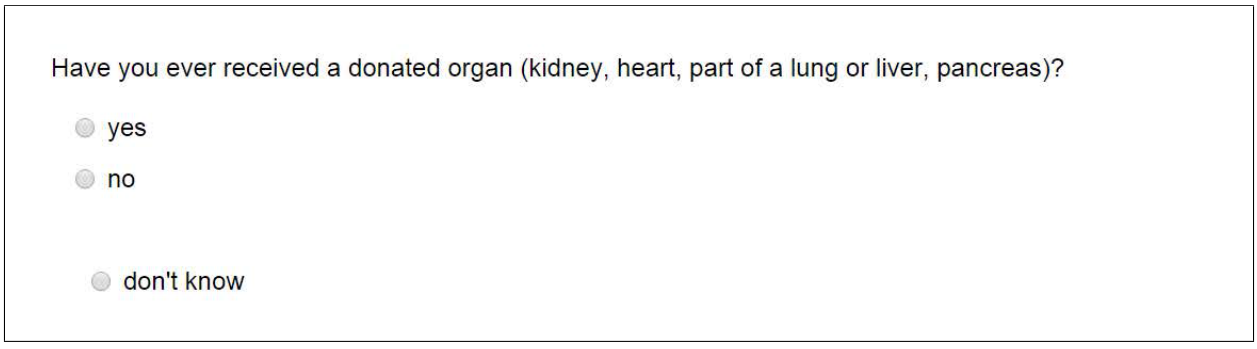

Figure B.1: Screen shot of the direct questioning implementation (translated from German)

[1] Because we used a fully-crossed experimental design, these treatments, which are not discussed here, have no impact on the sensitive question technique validation.

[2] We additionally performed most analyses excluding the 47 respondents who had assessed their language skills as only "medium" and not as "good" or "very good". The results are basically identical. See the online supplement for the corresponding analyses.

The crosswise-model RRT implemented was an unrelated question version as previously used in Jann, Jerke, and Krumpal (2012) and in most other studies using the crosswise model. Respondents were asked two questions at the same time: A sensitive question and an unrelated non-sensitive question (see Figure B.2). Respondents then had to indicate whether their answers to the two questions were identical (both "No" or both "Yes") or different (one "Yes", the other "No"). The CM procedure was carefully introduced to the respondents. On the first screen, we outlined the procedure and briefly explained how the technique protects individual answers. In addition, respondents were referred for further information about the RRT to a Wikipedia article which they could directly access by clicking on a button, with 18% of respondents making use of this possibility. On the second screen, respondents were shown a practice question on whether they had completed the "Abitur". Then, the five sensitive items followed.

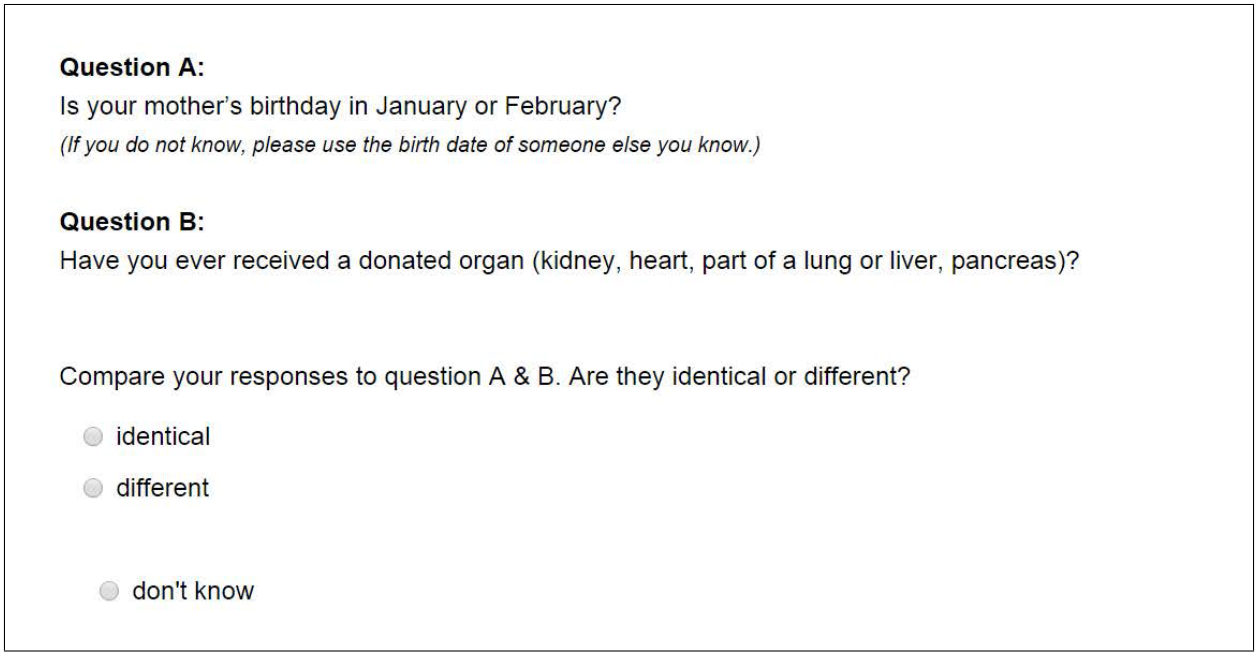

Figure B.2: Screen shot of the CM implementation (translated from German). Unrelated questions (Question A) were randomized across items and every question was only used once for each respondent.

Due to the mixing with the non-sensitive question, a respondent's answer to the sensitive question remains completely private. Nevertheless, at the aggregate level prevalence estimates for the sensitive question are possible because the probability distribution of the unrelated non-sensitive question is known. The unrelated questions used were about the birthdates of respondents' parents and of an arbitrarily chosen acquaintance such as "Is your mother's birthday in January or February?". Unrelated questions were randomly paired with the sensitive items for each respondent. Note that half the respondents received unrelated questions with a probability of a "yes" answer

of .15 to .20, the other half received inverted questions with a "yes" answer probability of .80 to .85 (see Table C.3 for a list of the unrelated questions used). Further, in both the DQ and the CM condition half the respondents were shown a "don't know" response option, whereas the other half were not.

The zero-prevalence items

As zero-prevalence items to test for systematic false positives served a question on having "ever received a donated organ" and on having "ever suffered from Chagas disease (Trypanosomiasis)". We deliberately chose zero-prevalence items that suited the survey topic and had near-zero prevalence in the surveyed population without being completely impossible so that they appeared meaningful to respondents. We did not find any statistics on living organ recipients in Germany. However, using the average number of transplanted organs in Germany from the last ten years ($4,400/year$) to extrapolate over the last 30 years and making the unrealistic but most conservative assumption that all patients who received an organ since 1985 are still alive and that each received only one organ, we can estimate an upper bound of organ recipients presently alive of 132,000, which corresponds to 0.16% of the population.

For the second item, Chagas disease, some epidemiological findings were available. Chagas disease is a parasitic disease spread mostly by insects and potentially leading to heart and digestive disorders that is endemic in most countries in South and Middle America. In Western Europe, however, the disease is nearly non-existent, the exception being Latin American migrants for whom studies found prevalence rates of slightly above 10% for samples from Florence and Geneva. Strasen et al. (2014) estimate an incidence rate for Germany of between 0.0001% and 0.0004%.

Data analysis

To correct for the systematic error that is introduced by the randomization procedure of the crosswise model, the response variable must be transformed. Let Y be the observed response variable with $Y = 1$ if the response is "identical" and $Y = 0$ for "different". S is the actual answer to the sensitive item with $S = 1$ if the answer to the sensitive item is "yes", and $S = 0$ for "no". $p^{yes,u}$ is the known probability of a "yes" answer to the unrelated question. The probability of the response "identical" then is

$$\Pr(Y = 1) = \Pr(S = 1) \cdot p^{yes,u} + (1 - \Pr(S = 1)) \cdot (1 - p^{yes,u})$$

Solving for $\Pr(S = 1)$ results in the transformed response variable $\tilde{Y}$ for the CM:

$$\tilde{Y} = \Pr(S = 1) = \frac{\Pr(Y = 1) + p^{yes,u} - 1}{(2p^{yes,u} - 1)}$$

For the direct questioning data, we set $p^{yes,u}$ to 1 so that $\tilde{Y}$ equals the untransformed response variable with $Y = S = 1$ if the answer is "yes" and $Y = S = 0$ if the answer is "no". For the prevalence estimates, we used least-squares regressions on this transformed response variable with robust standard errors (i.e. Fox and Tracy 1986). Data analysis was carried out using the Stata program rrreg (Jann 2008) which readily accommodates the outlined procedure. In addition, we performed all analyses using a logistic regression as well as a non-linear least-squares estimation. The results are essentially identical (see the online supplement for the corresponding analyses and Höglinger, Jann, and Diekmann 2016 for a more thorough discussion of RRT estimation strategies). Figures and tables of the estimated parameters were generated using the Stata programs coefplot (Jann 2014) and esttab (Jann 2007).

C. Additional results

Sensitivity of the items

To assess the sensitivity of the five surveyed items, towards the end of the survey we asked participants to rate how touchy answering them might be. Most items were not assessed as particularly sensitive by the majority of respondents (see Table C.1). The question on blood donation was assessed as "quite touchy" or "very touchy" by only 2% of respondents, the question on organ donation willingness by 23%, and the one on excessive drinking by 43%, apparently being the most sensitive item. The zero-prevalence item on whether one had received a donated organ was assessed as sensitive by 11%, the one on having suffered from Chagas disease by 15%. The five items covered quite a range of sensitivity, but in general appeared not too sensitive to most respondents.

Individual-level validation

As a complementary individual-level validation of the sensitive question techniques, we used a barely sensitive question on whether respondents had (not) completed the "Abitur", the general university entrance qualification. The question was presented as a practice question in the CM condition and appeared as a normal question in the DQ condition. Answers were validated using previously collected information on respondents' basic characteristics when they registered for the online panel. Some limitations apply to this validation. First, the question was presented as a practice question in the CM but not in the DQ condition. It is therefore possible that respondents

Table C.1: Sensitivity assessment of surveyed items

Sensitive item	Respondents assessing an item as "quite touchy" or "very touchy"
Never donated blood	2%
Unwilling to donate organs after death	23%
Excessive drinking last two weeks	43%
Received a donated organ	11%
Suffered from Chagas disease	15%

Notes: Question wording: "Please indicate for the following questions, how touchy answering them might be for some respondents". Answer categories were "not touchy at all", "relatively not touchy", "partly", "quite touchy", and "very touchy". *N* from 1,630 to 1,634

exercised relatively less care in answering it in the CM compared to DQ. To minimize this as far as possible, we asked respondents in the CM condition to "nevertheless, carefully follow the procedure" and to "answer the question truthfully", regardless of the fact that it is not sensitive and for practice. Second, the format differed between the question posed in our survey and the elicitation in the panel's registration form. In the survey, the question read "Have you completed the 'Abitur?"' with the response options "yes" and "no". In the registration form, respondents had to select their educational achievement from among several categories.[3] Third, respondents had registered for the panel up to five years prior to our survey and so it is possible that a few had completed the "Abitur" in the meantime and had not updated the corresponding panel information. However, this would only decrease the false-positive rate. Moreover, the latter two sources of error are constant in both the DQ and the CM condition, hence by comparing the validation results between DQ and CM they are controlled for.

Note that as for the items of the comparative validation the "Abitur" item was reverse-coded, such that the potentially socially undesirable response is the "yes" response, i.e. which corresponds to admitting not having completed the "Abitur". Results of the aggregate-level validation (upper panel of Figure C.3, also see Table C.2) show that the prevalence estimates of respondents not having completed the "Abitur" are nearly identical for DQ and the CM. Both are a negligible two percentage points above the corresponding validation values denoted by the diamond symbol (difference not significant). According to this, one would conclude that both techniques produce valid estimates equally well. This result does not seem surprising given that the question on whether one has completed the "Abitur" is neither barely sensitive nor ambiguous. Yet looking at results

[3]Because there is some disagreement in general understanding on whether one of the offered categories, the subject-specific university entrance qualification ("Fachhochschulreife"), is considered as "Abitur" or not, we excluded the 14% of respondents who selected it, restricting the validation to respondents who unequivocally indicated having completed the "Abitur" or not.

of the individual-level validation (middle and lower panel) tells a very different story. Note that the sensitive outcome is "having not completed the Abitur". Hence, the false negative rate is the share of respondents misclassified as having completed the "Abitur" even though they have not. It amounts to 9% in DQ and up to 29% for the CM. The false positive rate is the percentage of respondents incorrectly classified as not having completed the "Abitur" even though they have. It is not significantly different from zero in the DQ condition but a considerable 7% in the CM. Hence, the CM shows more missclassification than DQ in both directions. Note that the CM's high false negative and high false positive rates level each other out, resulting in an accurate aggregate prevalence estimate.

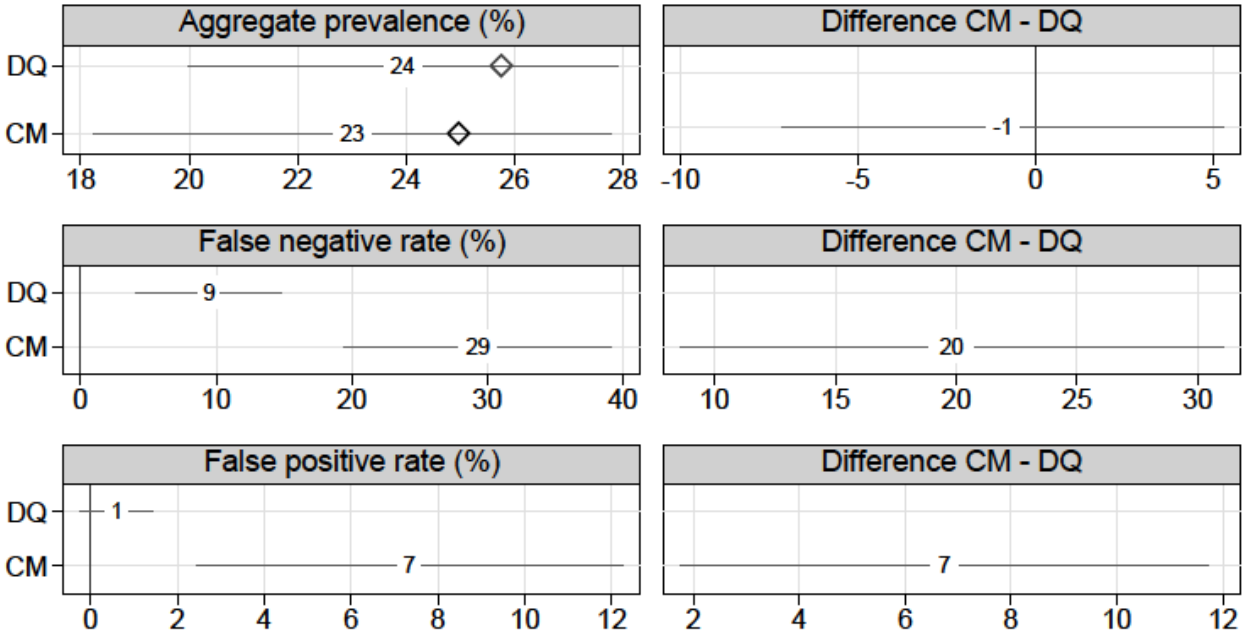

Figure C.3: Aggregate-level validation (upper panel) and individual-level validation (middle and lower panels). Diamond symbols denote the aggregate validation values of "no Abitur" (lines indicate a 95% confidence interval). $N = 458$ for DQ and $N = 953$ for CM

In sum, these results corroborate the findings from the zero-prevalence comparative validation. As mentioned, our individual-level validation had some limitations, mainly that we cannot rule out that the higher misclassification in the CM is caused to some extent by the fact the question was presented as a practice question in the CM condition. But what is most remarkable is not so much the finding that there was again misclassification in the CM, but that the substantial misclassification was not revealed in the aggregate-level validation. This demonstrates the serious weakness of such a validation strategy.

Table C.2: Aggregate and individual-level validation as displayed in Figure C.3 (standard errors in parentheses)

	Aggregate prevalence	False negative rate	False positive rate
Direct questioning (DQ)	23.94	9.48	0.60
	(2.02)	(2.73)	(0.43)
Crosswise model (CM)	23.01	29.29	7.34
	(2.43)	(5.03)	(2.51)
Difference CM - DQ	-0.93	19.81	6.74
	(3.16)	(5.72)	(2.54)

Notes: $N = 1,361$. Aggregated validation values are 25.76 for DQ, and 24.97 for CM

Exploring the causes and correlates of false positives in the CM

Having shown that false positives occurred in the CM with a non-ignorable frequency, we now look at some potential causes and mechanisms underlying this type of misclassification. We can think of two main causes: Careless answering and a bias in the unrelated question outcome that served as a randomizing device. Socially desirable responding can be excluded because the less incriminating answer to the zero-prevalence items is "no", i.e. denying having received a donated organ or having suffered from Chagas disease. Hence, it is hard to imagine why respondents would deliberately give a false "yes" answer to these questions.

The first, careless answering, might be the result of respondents not complying with the CM procedure to evade the effort involved or because they simply were unable to cope with the special procedure's complexity. Due to the privacy-protecting nature of the CM, false answers can never be revealed and so respondents might be more inclined to careless answering in the CM than in the direct questioning mode where answers are potentially verifiable (for this argument, also see Wolter and Preisendörfer 2013). Assuming that careless answering results in random responses, i.e. ticking the response options "different" and "identical" with equal probability[4], the share of respondents randomly answering needed to produce the bias found in our data would be twice the actual false positive rate: 15% for the "received organ" item and 10% for "Chagas disease" (see the left panel of Figure C.4).[5] Randomly answering always produces more false positives than

[4]Because the order of the response options was randomized across respondents and also because half the respondents received inverted unrelated questions, hence the correct response ("identical" or "different") was exactly the inverse, this assumption is quite plausible.

[5] The function for the false positive bias is derived from the transformed response variable $\tilde{Y}$ for the CM:

$$\tilde{Y} = \Pr(S = 1) = \frac{\Pr(Y = 1) + p^{yes,u} - 1}{(2p^{yes,u} - 1)}$$

We introduce the probability r of answering randomly, hence of giving the response "identical" with a probability of

negatives for a prevalence that in reality is below 0.5, which is typical for sensitive items.[6] Hence, in principle it could explain the overestimation bias found in our study as well as the consistently higher estimates from previous validations.

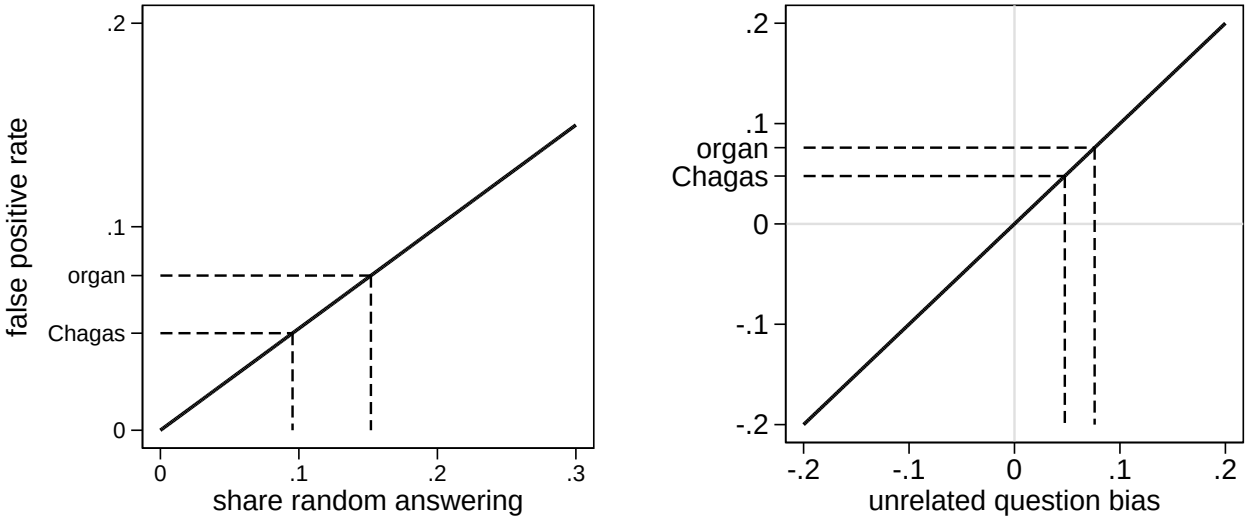

Figure C.4: Effect of random answering and unrelated question bias on the false positive rate for zero-prevalence items. (Dashed lines indicate false positive rates found and the corresponding extent of error necessary to generate them.)

Notes: With an expected "yes" probability for the unrelated questions of 0.18 as in the CM implemented. If the "yes" probability is inverted to 0.82 random answering has the same effect, but the effect of the unrelated question bias goes in the opposite direction.

The second potential cause, a bias in the unrelated question outcome, occurs if the unrelated questions do not produce the theoretically expected "yes" answer prevalence. We used unrelated questions about the birth dates of respondents' mother and father, and of arbitrarily chosen acquaintances. A bias in the "yes" probability could occur if there is actually a different prevalence of the underlying attribute in the study sample, which is quite unlikely for birthdate questions, or if respondents do not know the status of the attribute, i.e. the date of their parents' birth. In addition, for the question on an acquaintance's birthday which in one version read "Think of an acquaintance of yours whose birthday you know: Is this person's birthday in January or February?" respondents might be more inclined to choose an acquaintance whose actual birthdate falls within the specified time frame (January or February) or whose birthday falls about the time the survey was carried

.5, and a bias b:

$$\Pr(S = 1) + b = \frac{.5r + \Pr(Y = 1)(1 - r) + p^{yes,u} - 1}{(2p^{yes,u} - 1)}$$

After setting the "yes" prevalence $\Pr(S = 1)$ to zero and $\Pr(Y = 1)$ to $1 - p^{yes,u}$ as assumed, we arrive after some transformations at $b = r/2$.

[6]For estimates with a true prevalence above 0.5 the inverse holds: random answering leads to more false negatives and an underestimation in the aggregate. Complete random answering would lead in both cases to an estimate of 0.5.

out. To minimize such effects (and test them, see below), we randomized the unrelated questions across items and also used an inverted form for every unrelated question (instead of "in January or February", "in March to December, including December").

To generate the false positive rates found in our data, the "yes" answer bias must be of the same size, namely 8 and 5 percentage points (see the right panel of Figure C.4). We subjected the unrelated questions to a test by asking respondents of the DQ condition to explicitly answer the unrelated questions used in the CM.[7] A comparison of the so elicited "yes" prevalence with the theoretically expected prevalence showed a good match in general (see Table C.3). With the exception of three out of twelve questions, the differences were in the range of -5 to +3 percentage points and not significant. In part, very sizeable differences were found for the questions on "acquaintance's birthday in January or February" (36% instead of 16%, +20 percentage points bias), "acquaintance's birthday from the 1st to the 6th of the month" (31% instead of 20%, +11 percentage points), and for "father's birthday in March to December, including December" (77% instead of 84%, -7 percentage points). Interestingly, these prevalence estimates were all biased towards 50%, suggesting that choosing an answer at random might be the cause. Excluding responses based on these three potentially problematic unrelated questions indeed reduced false positive rates from 8% to 6% (received donated organ) and from 5% to 1% (Chagas disease, see the online supplement for the corresponding analysis). Apparently, some of the unrelated questions used might have been problematic. Most likely that is because they leave too much wiggle-space to respondents (the question on an acquaintance's birthday), or some respondents simply do not know the answer (the question on the father's birthday). A less unequivocal non-sensitive question or another randomizing device might therefore be preferable.

Note that, in contrast to random answering, a bias in the unrelated question outcome can lead to more false positives as well as more false negatives depending on the direction of the "yes" answer bias. This would not quite fit the pattern whereby the CM consistently produced more false positives. Still, the problematic questions identified with our test all showed a bias towards 50%, which would result in relatively more false positives. Therefore, the unrelated questions are likely responsible for some false positives, although they do not explain the whole bias.

Irrespective of the actual cause of the false positives (it might well be a mix of various mechanisms), we expected to find systematic patterns regarding implementation details of the CM as well as respondents' behavior and characteristics. In the following, we first present the effects of experimentally manipulated details of the CM implementation on false positives. Our analytical

[7] The questions were introduced as a "seemingly strange" task without detailing the purpose. To increase the certainly limited comparability, we employed a procedure as similar as possible and also randomized the question order. Of course, because the context of the questions when they were tested was very different to when they were used in the CM, we cannot directly infer that the same bias occurred in the CM. Still, the test provides some insights into the direction and possible size of the potential bias, and indicates potentially problematic questions.

Table C.3: Comparison of the elicited and theoretical "yes"-prevalence to unrelated questions used in the CM (standard errors in parentheses)

	"Yes" prevalence in test	Theoretical "yes" prevalence	Difference
Mother's birthday Jan-Feb	15.30	15.95	-0.65
	(2.20)		(2.20)
Mother's birthday 1st-6th	18.35	19.71	-1.36
	(2.37)		(2.37)
Father's birthday Jan-Feb	17.16	15.95	1.22
	(2.31)		(2.31)
Father's birthday 1st-6th	18.87	19.71	-0.85
	(2.41)		(2.41)
Acquaintance's birthday Jan-Feb	35.82	15.95	19.87*
	(2.93)		(2.93)
Acquaintance's birthday 1st-6th	30.57	19.71	10.85*
	(2.84)		(2.84)
Mother's birthday Mar-Dec	81.01	84.05	-3.05
	(2.45)		(2.45)
Mother's birthday 7th-31st	83.01	80.29	2.72
	(2.34)		(2.34)
Father's birthday Mar-Dec	77.38	84.05	-6.67*
	(2.64)		(2.64)
Father's birthday 7th-31st	75.60	80.29	-4.69
	(2.72)		(2.72)
Acquaintance's birthday Mar-Dec	82.75	84.05	-1.31
	(2.37)		(2.37)
Acquaintance's birthday 7th-31st	76.77	80.29	-3.52
	(2.65)		(2.65)

Notes: N from 250 to 268 per question. $^{*}p < 0.05$

strategy consisted of running bivariate regressions on the pooled response variables of the two zero-prevalence items, where answering "yes" is equivalent to giving a false positive. The results show that none of the experimental manipulations had a significant effect on false positives (Table C.4). The largest, albeit not significant effect (-4 percentage points, $p = 0.108$) was found for the introduction of a "don't know" response option.[8] All other manipulations such as reversing the

[8]Because only 0.7% (organ recipient) and 0.5% (Chagas) of the respondents provided with a "don't know" response option actually ticked it, the effect of the "don't know" option on false positives was not caused by respondents actually

Table C.4: Effects of CM implementation details on false positive rate (bivariate regression coefficients, standard errors in parentheses)

	Percentage points change
With "don't know" response option	-4.48
	(2.79)
Response order different - identical (vs. inverse)	-1.18
	(2.79)
Unrelated question on father (vs. mother)	-2.82
	(2.87)
Unrelated question on acquaintance (vs. mother)	2.69
	(2.91)
Unrelated question on birthday (vs. birth month)	2.04
	(2.73)
Yes-probability unrelated question .82 (vs. .18)	-2.10
	(2.79)
Item position (linear)	0.09
	(0.96)
Item position 1st or 2nd (vs. 4th or 5th)	-1.54
	(3.77)

Notes: Bivariate regressions on pooled responses to zero-prevalence items. Standard errors corrected for clustering in respondents. $N = 2,243$. $^{*}p < 0.05$

order of the response options from identical–different to different–identical, the type of the unrelated question (birthday of mother, father, or acquaintance; birthday vs. birth month), or inverting the "yes" probability of the unrelated question from on average $p = .18$ to $p = .82$ clearly had no effect. Moreover, no effects were found for the placement of the sensitive item, i.e. whether they were displayed as the first, second, third, fourth, or fifth item.

In the final step, we explored bivariate associations between giving a false positive and respondents' behavior and personal characteristics. Again, the results are far from conclusive (Table C.5). Being among the 10% of respondents who passed the CM introduction page with the explanations on the special technique the fastest was positively related to giving a false positive (+9 percentage points, albeit not significant at a conventional level, $p = 0.063$). This suggests that speeding respondents did not carefully read the instructions and thus did not fully understand the CM procedure, and consequently gave more false positive responses. But, somehow in contrast to this finding, being among the 10% fastest respondents in answering the five sensitive items was

making use of this option. It was the response behavior of those who ticked the "different" or "identical" response that was altered by simply having this option offered.

Table C.5: Bivariate associations between respondents' behavior and personal characteristics and false positive rate (bivariate regression coefficients)

	Percentage points change
Among the fastest 10% on CM introduction screen	9.05
	(4.87)
Among the fastest 10% answering sensitive items (without intro)	-4.33
	(4.46)
Clicked button referring to the RRT Wikipedia link	6.05
	(3.90)
Social desirability (Crown-Marlowe scale)	1.62*
	(0.80)
Completed the university entrance qualification	-5.17
	(3.53)
Age	-0.03
	(0.10)
Female	-1.73
	(2.95)

Notes: Bivariate regression on pooled zero-prevalence items. Standard errors corrected for clustering in respondents. N from 2,208 to 2,243. $^{*}p < 0.05$

clearly not positively associated with false positives. Clicking on the button provided to access the Wikipedia page with further RRT information on the introduction screen also showed no significant association. Scoring high on the Crowne-Marlowe social desirability scale (Crowne and Marlowe 1960) was positively related to giving a false positive (+1.6, $p = 0.042$, $scaleSD = 1.7$), meaning that respondents more prone to socially desirable responding were also more likely to give a false positive. We have no explanation for this finding because, if any social desirability bias existed, it would instead work against falsely admitting having suffered from Chagas disease or having received a donated organ. Finally, having completed the university entrance qualification is not systematically related to false positives, nor are age or gender.

Note that the statistical power of the previous analyses was relatively weak due to the low prevalence of the false positives. In addition, we tested several potential causes and covariates without having a clear theory about how they are related to false positives in the CM. Hence, the risk of both alpha and beta errors increased considerably and the findings presented in this section must be interpreted as exploratory. However, in light of the novelty of the finding that the CM produced false positives and a unique possibility to analyze the potential causes these results are, in our view, nevertheless valuable for informing future studies dealing with improving the crosswise model or related techniques. In sum, the analysis of the causes and correlates of false positives

did not reveal any pattern that would clearly point to a particular explanation. We could, however, identify some candidate causes of false positives whose effect should be investigated more systematically in future studies: The unrelated questions used and their respective bias, not offering a "don't know" response option (albeit the reason is unclear), and respondents speeding over the CM instructions. Still, each of these factors accounts for only a share of the false positives that occurred and, very likely, false positives might have been caused by a mix of different mechanisms.

D. **Table underlying the figure in the main text**

Table D.6: Comparative validation of sensitive question techniques as displayed in Figure 1 (standard errors in parentheses)

	Never donated blood	Unwilling to donate organs	Excessive drinking	Received a donated organ	Suffered from Chagas disease
Levels					
Direct questioning (DQ)	48.82	22.01	20.58	0.00	0.37
	(2.14)	(1.82)	(1.73)	(.)	(0.26)
Crosswise model (CM)	51.58	27.30	32.71	7.60	4.77
	(2.33)	(2.23)	(2.28)	(1.95)	(1.91)
Difference					
CM – DQ	2.76	5.29	12.13	7.60	4.40
	(3.16)	(2.88)	(2.86)	(1.95)	(1.92)
N	1669	1641	1672	1669	1669

References

Crowne, Douglas P., and David Marlowe. 1960. "A New Scale of Social Desirability Independent of Psychopathology". *Journal of Consulting Psychology* 24:349–354.

Fox, James Alan, and Paul E. Tracy. 1986. *Randomized response: A method for sensitive surveys*. Newbury Park, CA: Sage.

Hoffmann, Adrian, Birk Diedenhofen, Bruno Verschuere, and Jochen Musch. 2015. "A Strong Validation of the Crosswise Model Using Experimentally-Induced Cheating Behavior". *Experimental Psychology* 62:403–414.

Höglinger, Marc, and Andreas Diekmann. 2016. *Replication Data for: Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT*. Harvard Dataverse. doi:10.7910/DVN/SJ2RP1.

Höglinger, Marc, and Ben Jann. 2016. *More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model*. University of Bern Social Sciences Working Paper No. 18. University of Bern. https://ideas.repec.org/p/bss/wpaper/18.html.

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 (3): 171–87. doi:10.18148/srm/2016.v10i3.6703.

Jann, Ben. 2007. "Making regression tables simplified". *Stata Journal* 7:227–44.

— . 2014. "Plotting regression coefficients and other estimates". *Stata Journal* 14:708–37.

— . 2008. *rrreg: Stata module to estimate linear probability model for randomized response data*. S456962. Boston College Department of Economics.

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.

John, Leslie K., George Loewenstein, Alessandro Acquisti, and Joachim Vosgerau. 2016. *When and Why Randomized Response Techniques (Fail to) Elicit the Truth*. Harvard Business School Working Paper No. 16-125. Harvard Business School. http://www.hbs.edu/faculty/Pages/item.aspx?num=51059.

Kirchner, Antje. 2015. "Validating Sensitive Questions: A Comparison of Survey and Register Data". *Journal of Official Statistics* 31:31–59.

Moshagen, Morten, Benjamin E. Hilbig, Edgar Erdfelder, and Annie Moritz. 2014. “An Experimental Validation Method for Questioning Techniques That Assess Sensitive Issues”. *Experimental Psychology* 61:48–54.

Strasen, Jörn, Tatjana Williams, Georg Ertl, Thomas Zoller, August Stich, and Oliver Ritter. 2014. “Epidemiology of Chagas Disease in Europe: Many Calculations, Little Knowledge”. *Clinical Research in Cardiology* 103:1–10.

van der Heijden, Peter G. M., Ger van Gils, Jan Bouts, and Joop J. Hox. 2000. “A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning. Eliciting Sensitive Information in the Context of Welfare and Unemployment Benefit”. *Sociological Methods & Research* 28:505–537.

Wolter, Felix, and Peter Preisendörfer. 2013. “Asking Sensitive Questions: An Evaluation of the Randomized Response Technique vs. Direct Questioning Using Individual Validation Data”. *Sociological Methods & Research* 42:321–353.