Walzenbach, Sandra; Hinz, Thomas

Working Paper

Puzzling Answers to Crosswise Questions - Examining Overall Prevalence Rates, Primacy Effects and Learning Effects

Puzzling Answers to Crosswise Questions

Examining Overall Prevalence Rates, Primacy Effects and Learning Effects

Sandra Walzenbach
University of Konstanz
sandra.walzenbach@uni-konstanz.de

Thomas Hinz
University of Konstanz
thomas.hinz@uni-konstanz.de

ABSTRACT

This validation study on the crosswise model (CM) examines five survey experiments that were implemented in a general population survey. Our first crucial result is that in none of these experiments was the crosswise model able to verifiably reduce social desirability bias.
In contrast to most previous CM applications, we use an experimental design that allows us to distinguish a reduction in social desirability bias from heuristic response behaviour, such as random ticking, leading to false positive or false negative answers. In addition, we provide insights on two potential explanatory mechanisms that have not yet received attention in empirical studies: primacy effects and panel conditioning. We do not find consistent primacy effects, nor does response quality improve due to learning when respondents have had experiences with crosswise models in past survey waves. We interpret our results as evidence that the crosswise model does not work in general population surveys and speculate that the question format causes mistrust in participants.

1) Introduction

The crosswise model (CM) has lately received a lot of attention in sensitive question research. As a method that was designed to ensure anonymity and does not require a random device, it was hoped to overcome some of the key flaws associated with Randomized Response Techniques (Yu, Tian, and Tang 2008). Many studies drew - and we believe that too many keep drawing - positive conclusions regarding the method. Despite the fact that concerns about its validity have been voiced in some recent studies (Höglinger and Jann 2018; Jerke et al. 2021; Kuhn and Vivyan 2018; Walzenbach and Hinz 2019), the crosswise model is still implemented by researchers who believe in its attenuating effects on social desirability bias (Banayejeddi et al. 2019; Johann and Thomas 2017; Mieth et al. 2021; Vakilian, Mousavi, and Keramat 2019). Even two recent meta-analyses paint a positive picture; one considers the crosswise model a "promising" method (Sagoe et al. 2021; Schnell and Thomas 2021).

The trouble is that these conclusions are based on the finding that a crosswise model on average yields higher estimates than a direct question. In contrast to these views, we argue that this is a bad indicator for data quality. The CM estimator is systematically biased towards 50% whenever a socially undesirable behaviour with low prevalence is assessed and respondents disobey the instructions (inadvertently or deliberately). This fact leads most studies, including the existing meta-analyses, to unjustifiably positive conclusions and, maybe worse, keeps most authors from looking at the underlying mechanisms that cause bias in CM estimates.

In light of these current gaps in CM research, this paper theoretically explains why the assumption that "more is better" is faulty in the overwhelming majority of all existing CM applications and provides insights on the applicability of the crosswise model in a heterogeneous population sample. We present empirical results from five experiments on the validity of the

crosswise model that were implemented in a general population panel survey over the time span of several years. Our approach is superior to most previous research insofar as we do not merely rely on the comparison to direct questions. Instead, for two of our experiments, we use an innovative design that allows disentangling a reduction of social desirability bias from heuristic response behaviours, such as random ticking (explained in detail in Section 2b). In addition, we partly draw on external validation criteria. In a further step, we examine some of the underlying mechanisms that might drive the observed patterns, namely primacy/recency effects and learning through repeated exposure.

Last but not least, the study uses panel data from a heterogeneous population sample to examine the applicability of the crosswise model. Although it has been argued that convenience and general population samples respond differently to crosswise questions, and the positive conclusions authors draw might be due to the convenience samples they use (Schnell and Thomas 2021), validation studies with heterogeneous samples are still extremely rare.

We will proceed as follows: Section 2 contains a brief theoretical introduction to the crosswise model, gives a critical assessment of previous validation studies and looks at what we know from previous research. After discussing our research question, hypotheses, data, and concrete experiments in Section 3, we present our empirical results in Section 4: We evaluate the overall performance of the crosswise model throughout our series of experiments (Section 4a), discuss potential response order effects (Section 4b), and panel conditioning / learning effects (Section 4c).

2) The Crosswise Model and Common Flaws in Previous Research

a) Basic Logic of the Crosswise Model

In a nutshell, the crosswise model combines two dichotomous questions into one response task: the sensitive question of interest and a non-sensitive question with a known probability (see example in Figure 1). Respondents only provide information as to whether their answers to these two questions are equal or different. This means that there is no socially undesirable or revealing response option. Being structurally equivalent to Warner's Randomized Response Technique (Warner 1965)[1], the procedure adds additional random noise to the response process for the sake of greater privacy (for technical details, see Yu et al. 2008). Assuming that respondents answer more honestly to this question format, it should provide more accurate overall prevalence rates for the sensitive behaviour than a direct question.

Figure 1. Basic Logic of the Crosswise Model

1) non-sensitive question with known probability p

"Is your mother's birthday in January, February or March?"

2) sensitive question with unknown prevalence rate π

"Have you ever been arrested?"

Possible answers:

☐ YES to both questions or NO to both questions

☐ YES to one question and NO to the other question

b) Common Flaws: The Assumption That "More Is Better" Is Usually Wrong

[1] Knowing that the first answer category ($\lambda=1$) will be ticked if both items are answered with "yes" ($p\pi$) or if both items are answered with "no" $(1-p)(1-\pi)$, the prevalence rate π can be estimated for given λ and p by using the formula $\lambda=p\pi + (1-p)(1-\pi)$.

Most applications of the crosswise model have assessed socially undesirable behaviour with low prevalence rate such as plagiarism, xenophobia, tax evasion and drug consumption (Coutts et al. 2011; Hoffmann and Musch 2016; Höglinger, Jann, and Diekmann 2016; Jann, Jerke, and Krumpal 2012; Jerke et al. 2021; Korndörfer, Krumpal, and Schmukle 2014; Shamsipour et al. 2014). Typically, the crosswise estimate is compared to an experimental condition with a direct question. A higher CM than DQ prevalence is often interpreted as a successful reduction in social desirability bias. The fundamental problem with this approach is that also heuristic response behaviours such as random ticking would make the CM estimate tend towards 50% (also see Höglinger and Jann 2018). This means that respondents that are confused and/or do not comply with the procedure produce the same response pattern as a crosswise model that successfully reduces social desirability bias. This is true for socially undesirable behaviours with low prevalence rates of under 50% (such as having been arrested) and for socially desirable behaviours with high prevalence rates of above 50% (such as paying taxes).
Contrastingly, desirable but rare behaviours (such as blood donation) and undesirable but common behaviours (such as jaywalking), allow us to disentangle the two mechanisms (see highlighted cells in Table 1).

Table 1. Disentangling Random Ticking and Reduction of Social Desirability Bias

	Prevalence Rate <50%	Prevalence Rate >50%
Desirable Behaviour	Have you ever donated blood?	Have you always payed your taxes?
Undesirable Behaviour	Have you ever been arrested?	Have you ever passed a red traffic light?

In what follows, we will come back to this distinction and present some experimental designs that allow to distinguish random ticking from a successful reduction of social desirability bias. An additional strategy is the use of external validation criteria if available.

c) What Do We Know About Underlying Mechanisms?

For the reasons discussed above, most studies do not allow researchers to draw conclusions about bias in CM estimates at all. Reflecting the problem that CM estimates tend towards 50% in the overwhelming majority of implementations, a part of the recently published studies on the crosswise model have focused on examining false positive and/or false negatives - and have often drawn very skeptical conclusions (Höglinger and Jann 2018; Kuhn and Vivyan 2018; Walzenbach and Hinz 2019). This strand of studies suggests that the crosswise model might add to bias, more than reducing it. However, the underlying mechanisms why this is happening remain unclear. Very rarely do authors even report correlates of bias in CM estimates.

The only comparatively well-documented hypothesis is that the CM procedure is not well understood by respondents (Jerke et al. 2019; Khosravi et al. 2015; Meisters, Hoffmann, and Musch 2020). The cognitive burden is assumed to trigger satisficing (Krosnick 1991; Simon 1957) and heuristic responses, such as random ticking. This is why survey methodologists typically use self-reported comprehension or education as indicators for cognitive load and risk of satisficing. However, empirical studies usually have trouble linking comprehension to more honest responses: At least the typical indicators (self-reported comprehension and education background) tend to have inconsistent or no effects on bias in CM estimates (Jerke et al. 2019; Meisters et al. 2020; Walzenbach and Hinz 2019; Wolter and Diekmann 2021).

3) Our Study: Five Survey Experiments in a Heterogeneous General Population Sample

a) Research Question and Hypotheses

In light of previous research, the aim of our study is twofold: In addition to a general evaluation of the crosswise model‘s validity in a heterogeneous population sample, we are interested in two concrete mechanisms that can explain the patterns we observe: order and learning effects. Examining the role of order effects for bias is a logical follow-up arising from the findings we obtained in a previous experiment (see Section 3c for further details). The idea to analyse learning effects is very much based on theoretical arguments from previous research. As we are lucky enough to have panel data, we will go beyond what other studies have done by looking at comprehension from a different perspective. If the crosswise model is simply too complicated for respondents to understand, we expect that data quality improves when respondents are repeatedly confronted with a crosswise model, read the instructions for the second time, and have the chance to learn.

b) Data

Over the course of several years, we implemented a series of five survey experiments on the crosswise model in a general population panel survey (“Konstanz Citizen Survey”) including the registered citizens in a town in the south of Germany. Data is collected once a year, usually from October to before Christmas. Respondents are selected based on a random sample from the population register and invited by postal letter to join the online panel survey. Since wave 4, refreshment samples are drawn in regular intervals to mitigate higher rates of unit-nonresponse and panel attrition within young people and immigrants (see Appendix A1 for more details on sampling strategy and cooperation rates). This strategy leads to a general population sample that

reflects the target population in sex, age, migration background and area of residence within the city. With regard to content, the survey covers issues of general interest but has a focus on political participation and activities at the community level. As a consequence, higher educated and politically interested citizens are more likely to participate.

In most years, citizens could fill in a paper questionnaire upon request. Since these paper versions did usually not contain the experiments on the crosswise model, we limit our analyses to the online panel members.

c) Survey Experiments

The CM experiments were designed to elicit different desirable and undesirable behaviours: (1) voter turnout, (2) blood donation, (3) littering, (4) keeping too much change, and (5) jaywalking (see Appendix A2 for exact question wording). In all of them, respondents were randomly assigned to a crosswise model or a direct question.

In addition to this common but somewhat error-prone strategy to validate the crosswise model's performance, the design of experiments (2) and (5) allows us to directly disentangle a reduction in social desirability bias and the effects of random ticking. Moreover, external validation data were available for experiments (1) and (2). Experiments (4) and (5) varied the order of the presented response categories in the crosswise model to examine a potential primacy effect.

This research question was inspired by the results of the experiment on blood donation. In a previous paper, we tried to explain the method's failure to reduce bias by indicators that are traditionally related to satisficing. However, we could not find any significant correlations of respondent characteristics, such as age and education, and biased crosswise estimates (see

Walzenbach and Hinz 2019 for details). It seemed as if respondents had a general tendency to choose the first response category.

All experiments are listed in Table 2.

Table 2. Summary of Crosswise Experiments

	(1) voter turnout (W4)	**(2) blood donation (W6)**	**(3) littering (W7)**	**(4) keeping too much change (W8)**	**(5) jaywalking (W8)**
elicited behaviour	desirable	desirable	undesirable	undesirable	undesirable
prevalence (DQ)	>50%	<50%	<50%	<50%	>50%
disentangling of mechanisms possible	no	yes	no	no	yes
external criterion for validity available	yes	yes	no	no	no
experiment on order of response categories	no	no	no	yes	yes
hypothesis if CM works	CM closer to real prevalence than DQ	CM < DQ	(CM>DQ)	(CM>DQ)	CM > DQ

d) Analytical Strategy

We evaluate the general performance of the crosswise model by comparing CM estimates to the respective direct questions and external validation criteria. This is done for all experiments (Section 4a). Note that due to random assignment to the experimental conditions, the general conclusions in this paragraph are unaffected by real differences in prevalence rates between groups of respondents (e.g. if younger people are more likely to donate blood). To examine order

effects in experiments (4) and (5), the CM estimates stemming from implementations with equal wording but different response orders are compared (Section 4b). Panel conditioning / learning effects are assessed in experiments (2) and (5), the ones that allow us to disentangle random ticking from a reduction in social desirability bias (Section 4c).

All reported significance tests are obtained from regression models using the stata ado rrreg (Jann 2008). It applies a least squares procedure to the transformed response variable $Y_i = \frac{\lambda_i + p_i - 1}{2p_i - 1}$, which indicates the answer "yes" to the sensitive question (for details see Jann et al. 2012).

4) Results

a) Prevalence Rates From Five Crosswise Experiments

For all five experiments, Figure 2 compares the estimated prevalence rates from a direct question and the crosswise model.

The estimates for experiments (3) and (4), littering and keeping too much change, follow the pattern that we would usually expect in most crosswise experiments, which typically assess socially undesirable behaviour with low prevalence rates: The crosswise estimator comes with a significantly higher share of respondents that admit the undesirable behaviour, but it is unclear if this is because the model reduces social desirability bias or because respondents did not follow the instructions correctly.

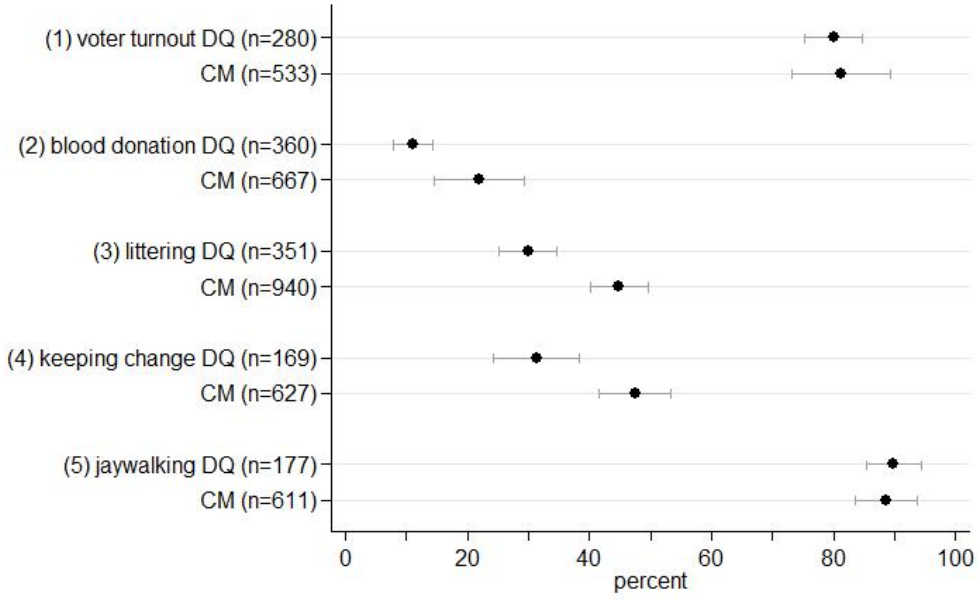
Although the difference between experimental conditions is smaller, the same pattern can be seen for voter turnout in experiment (1). Compared to the true value (46%), we vastly overestimate voter turnout in the survey data (DQ: 80.0% and CM: 81.3%; difference not significant with p=0.82). Without doubt, this is due to self-selection of politically interested

citizens into survey participation. However, although our sample is obviously not suitable to estimate voter turnout in the target population, the crosswise model should at least yield an estimate that is closer to the true value than the direct question – if it reduced social desirability. The fact that this is not the case casts first doubts on the crosswise model's performance.

We will now turn to experiments (2) and (5), assessing blood donation and the prevalence of jaywalking. Both experiments allow disentangling a valid CM estimate from random ticking. Jaywalking is undesirable but has a prevalence rate of above 50% in both question formats. More honest answers should thus result in higher CM estimates. Empirically, however, this is not what we find. If anything, the CM share is slightly lower (DQ: 89.8% and CM: 88.7%; difference not significant with $p=0.82$).

In the experiment on blood donation, a desirable low-prevalence behaviour was assessed and we would expect lower CM than DQ prevalence rates if the crosswise model worked properly. However, we again fail to observe such a pattern. The share of blood donors even is eleven percentage points higher in the crosswise model (22.0%) than in the direct question (11.1%), a statistically significant difference ($p=0.37$). External validation data from the Red Cross suggests a true prevalence rate of below 5% (for a detailed discussion including online and paper respondents, see Walzenbach and Hinz 2019).

Figure 2. Estimated Prevalence Rates in DQ and CM



Data: Konstanz Citizen Survey (online-panel respondents from waves 4, 6, 7, 8)

All in all, there was no empirical evidence for a successful reduction of social desirability bias in any of the survey experiments under study. In some cases, the crosswise model even produced worse estimates than the direct question.

b) Results on Response Order Effects

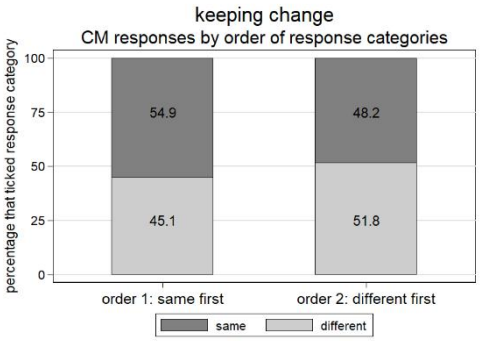
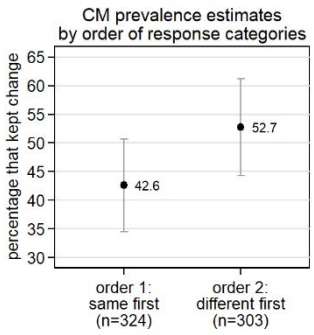
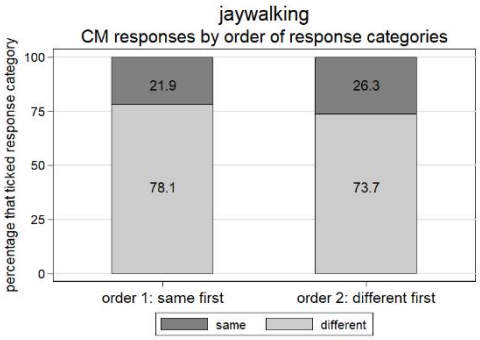
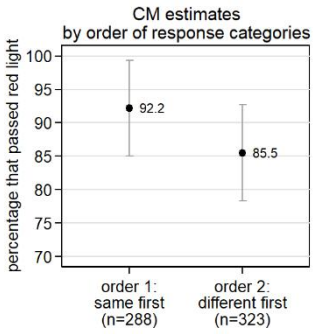
Experiments (4) and (5) on keeping too much change and jaywalking were designed to test if answers to the crosswise model depended on the order in which the response categories were presented. Considering our two follow-up experiments, however, we only found weak empirical evidence for a primacy effect in experiment (4) (see row 1 in Figure 3). In this case, the response category that was displayed first was picked slightly more often irrespective of content (e.g. 54.9% of respondents ticked ‘same’ if this answer was displayed first, but only 48.2% chose it when it came second). This tendency is suggesting a primacy effect, meaning that respondents

partly apply a heuristic response strategy. As a consequence, the estimated prevalence rates stemming from the two different orders of response categories differ by roughly 10 percentage points (p=0.09 according to a regression of CM prevalence on experimental condition).

In experiment (5), however, the order in which the response categories are presented hardly influenced response behaviour (see row 2 in Figure 3). The estimated prevalence rates do not differ significantly by response order (p=0.20). If we wanted to interpret the direction of the effect, it would rather suggest a recency than a primacy effect.

Although the differences between the DQ and CM conditions in Experiment (4) and (5) fail to reach traditional levels of significance (yielding p-values of 0.09 and 0.2), we think that these two findings are somewhat contradictory. Keeping in mind that the crosswise model is a procedure that inflates standard errors and considering that the estimates point towards opposite directions, differences of 7 to 10 percentage points do not seem trivial. We conclude that, instead of clear evidence for a primacy effect, our findings rather suggest that response behaviour is inconsistent and susceptible to minor differences, e.g. in question wording or survey setting.

Figure 3. Results on Order Effects for Experiments (4) and (5)

	ticked response categories	prevalence rates by response order
Experiment 4: keeping too much change		
Experiment 5: jaywalking		

Data: Konstanz Citizen Survey (wave 8)

c) Results on Panel Conditioning / Learning Effects

We argued that data quality should improve when respondents are repeatedly confronted with a crosswise model and have the chance to learn. Table 3 compares experiments (2) and (5), for which we can clearly disentangle a reduction in social desirability bias and random ticking. For the socially desirable behaviour with low prevalence (experiment 2), the CM estimate should be smaller for experienced than for unexperienced respondents. For the socially undesirable behaviour with high prevalence (experiment 5), the CM estimate should be higher when respondents had the chance to learn.

The columns represent the prevalence rates from the direct question format and the crosswise estimates for different levels of familiarity with the crosswise model (without / with previous

experience[2]). Significance tests of a potential learning effect were obtained by running regression models of the CM prevalence rate on the experience level and are shown in the penultimate column.

Due to non-random panel attrition, respondents with different experience levels differ in sample composition. We provide two types of additional analyses to account for this. First, the last column of Table 3 shows significance tests from regression models controlling for respondent sex, age (18-30 / 31-59 / 60 and older) and highest educational degree (below high school / high school diploma / university degree). Secondly, we ran separate regression models for respondents with comparable sociodemographic characteristics, whenever case numbers allowed this (see Appendix A3).

Table 3. Results on Learning and Panel Conditioning / Robustness Checks

	CM without experience	CM with experience	significance tests from regression (without controls)	(controlling for sex, age, education)
(2) Blood donation	20.6% (N=459)	25% (N=208)	p=0.58	p=0.23
(5) Jaywalking: order 1	94.5% (N=128)	90.3% (N=160)	p=0.57	p=0.94
(5) Jaywalking: order 2	94.4% (N=162)	76.6% (N=161)	p=0.015	p=0.06

[2] For experiment 2, previous experience means that respondents have already answered the crosswise model on voting behaviour in the previous wave. Experiment 5 was implemented later and respondents can have answered more than one crosswise experiment in previous survey waves. For this reason, respondents with previous experience entail people that have previously answered one, two or three crosswise models, dependent on when they joined the panel and how often they were randomly assigned to the crosswise condition instead of the direct question.

Empirically, there is no evidence for a learning effect, neither in experiment 2 nor in experiment 5. Contrary to our expectations, respondents with more experience in answering crosswise models, show more biased prevalence rates than those without any experience: Among the experienced respondents, a higher share pretends that they have donated blood and a lower share admits having jaywalked. For the experimental group with order 2, experienced respondents even produce a 17 percentage points lower prevalence rate than the unexperienced, a difference that remains significant on a 10%-level if sociodemographic characteristics are controlled ($p=0.06$). Put differently, we find more socially desirable answers among more experienced respondents, although two out of the three differences are far from reaching statistical significance.

This finding is corroborated by the additional robustness checks in the appendix: There are no significant learning effects for any respondent group, with most coefficients even pointing towards the opposite direction (see Appendix A3).

The findings related to learning effects and panel conditioning can be summed up as follows: While we expected that learning helps to deal with the rather complex CM format, the contrary was the case in our data. Repeated exposure to the crosswise model seems to have no or a detrimental effect on data quality. In line with these findings, it is possible that the unusual question format triggers mistrust or privacy concerns that respondents did not have in the first place. However, this is a mere theoretical hypothesis that would need empirical testing in a future project.

5) Discussion

Summing up, this validation study casts serious doubts about the applicability of the crosswise model in heterogeneous respondent samples. We presented empirical evidence from five experiments that were implemented in a general population sample and elicited different socially desirable and undesirable behaviours. Some of these survey experiments were specifically designed to distinguish a reduction in social desirability bias from random ticking, some of them could rely on external validation data. Our main finding is that the crosswise model consistently failed to verifiably reduce social desirability bias. In some cases, the crosswise model even produced more biased prevalence rates than a direct question.

Concerning the mechanisms underlying these findings, a cautious conclusion from our analysis on panel conditioning and learning is that it is not the complexity of the model that motivates respondents to use heuristic response strategies. This finding is in line with some previous studies that do not find any link between education or understanding of the procedure and honest responses (Jerke et al. 2019; Walzenbach and Hinz 2019) but contradicts others (Schnell and Thomas 2021). Our suspicion is that the question format might trigger privacy concerns irrespective of respondents' experience or cognitive skills. This argument has been made for traditional RRT implementations (John et al. 2018). However, we are not aware of any study that has explicitly examined this hypothesis for the crosswise model, which leaves room for future research.

Our empirical results are in line with the idea that respondents react to highlighted privacy concerns by randomly ticking an answer. At the same time, there does not seem to be one response category that respondents consistently prefer. At least in this study, we only found weak evidence for a primacy effect in one out of two CM implementations. However, the fact that

different response orders can trigger considerable differences in prevalence rates shows its susceptibility to small changes in the questionnaire and should in itself be interpreted as a warning sign.

Considering strengths and limitations, our study provides a neat experimental approach to evaluate the general applicability of the crosswise model in a probability-based general population sample. We believe this is a valuable contribution for two reasons: First, previous studies in the field very rarely use anything but convenience samples of students or academics, and access panels. Secondly, it seems to be a timely and necessary counterbalance to the two recently published meta-analyses on the crosswise model, which (for reasons discussed above) have come to dubiously positive conclusions.

Generally, our study leaves many open questions concerning the mechanisms that cause the response patterns we observe. We examined response order effects and learning through repeated exposure but found only inconsistent or null effects in our data. Nonetheless, we believe these results are a valuable step on the way to a fuller understanding of the crosswise model and the response behaviour it triggers.

All in all, our findings point towards the crosswise model's failure to reduce social desirability bias. Based on the current state of research, we cannot generally recommend implementing such question formats in general population surveys. This paper has shown that just having a direct question to compare CM estimates to is not enough to truly assess bias in the overwhelming majority of CM implementations with undesirable low prevalence items, as also random ticking leads to higher prevalences in the CM condition. If at all, we suggest using crosswise models to elicit desirable behaviours with low prevalence rates and undesirable behaviours with high

prevalence rates, in combination with a DQ condition, as this design allows researchers to identify potential problems.

REFERENCES

Banayejeddi, Mortaza, Sima Masudi, Sakineh Nouri Saeidlou, Fatemeh Rezaigoyjeloo, Fariba Babaie, Zahra Abdollahi, and Fatemeh Safaralizadeh. 2019. 'Implementation Evaluation of an Iron Supplementation Programme in High-School Students: The Crosswise Model'. *Public Health Nutrition* 22(14):2635–42. doi: 10.1017/S1368980019001575.

Coutts, Elisabeth, Ben Jann, Ivar Krumpal, and Anatol-Fiete Näher. 2011. 'Plagiarism in Student Papers: Prevalence Estimates Using Special Techniques for Sensitive Questions'. *Jahrbücher Für Nationalökonomie Und Statistik* 231(05–06):749–60.

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 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(8):e0201770. doi: 10.1371/journal.pone.0201770.

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. 2008. 'RRREG: Stata Module to Estimate Linear Probability Model for Randomized Response Data'.

Jann, Ben, Julia Jerke, and Ivar Krumpal. 2012. 'Asking Sensitive Questions Using the Crosswise Model. An Experimental Survey Measuring Plagiarism'. *Public Opinion Quarterly* 76(1):32–49. doi: 10.1093/poq/nfr036.

Jerke, Julia, David Johann, Heiko Rauhut, and Kathrin Thomas. 2019. 'Too Sophisticated Even for Highly Educated Survey Respondents? A Qualitative Assessment of Indirect Question Formats for Sensitive Questions'. *Survey Research Methods* 13(3):319–51. doi: 10.18148/srm/2019.v13i3.7453.

Jerke, Julia, David Johann, Heiko Rauhut, Kathrin Thomas, and Antonia Velicu. 2021. 'Handle with Care: Implementation of the List Experiment and Crosswise Model in a Large-Scale Survey on Academic Misconduct'. *Field Methods* 1525822X20985629. doi: 10.1177/1525822X20985629.

Johann, David, and Kathrin Thomas. 2017. 'Testing the Validity of the Crosswise Model: A Study on Attitudes Towards Muslims'. *Survey Methods: Insights from the Field (SMIF)*.

John, Leslie K., George Loewenstein, Alessandro Acquisti, and Joachim Vosgerau. 2018. 'When and Why Randomized Response Techniques (Fail to) Elicit the Truth'. *Organizational Behavior and Human Decision Processes* 148:101–23. doi: 10.1016/j.obhdp.2018.07.004.

Khosravi, Ahmad, Seyed Abbas Mousavi, Reza Chaman, Faride Khosravi, Mohammad Amiri, and Mansour Shamsipour. 2015. 'Crosswise Model to Assess Sensitive Issues: A Study on Prevalence of Drug Abuse Among University Students of Iran'. *International Journal of High Risk Behaviors and Addiction* 4(2). doi: 10.5812/ijhrba.24388v2.

Korndörfer, Martin, Ivar Krumpal, and Stefan C. Schmukle. 2014. 'Measuring and Explaining Tax Evasion: Improving Self-Reports Using the Crosswise Model'. *Journal of Economic Psychology* 45(1):18–32.

Krosnick, Jon A. 1991. 'Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys'. *Applied Cognitive Psychology* 5(3):213–36. doi: 10.1002/acp.2350050305.

Kuhn, Patrick M., and Nick Vivyan. 2018. 'Reducing Turnout Misreporting in Online Surveys'. *Public Opinion Quarterly* 82(2):300–321. doi: 10.1093/poq/nfy017.

Meisters, Julia, Adrian Hoffmann, and Jochen Musch. 2020. 'Can Detailed Instructions and Comprehension Checks Increase the Validity of Crosswise Model Estimates?' *PLOS ONE* 15(6):e0235403. doi: 10.1371/journal.pone.0235403.

Mieth, Laura, Maike M. Mayer, Adrian Hoffmann, Axel Buchner, and Raoul Bell. 2021. 'Do They Really Wash Their Hands? Prevalence Estimates for Personal Hygiene Behaviour during the COVID-19 Pandemic Based on Indirect Questions'. *BMC Public Health* 21(1):12. doi: 10.1186/s12889-020-10109-5.

Sagoe, Dominic, Maarten Cruyff, Owen Spendiff, Razieh Chegeni, Olivier de Hon, Martial Saugy, Peter G. M. van der Heijden, and Andrea Petróczi. 2021. 'Functionality of the Crosswise Model for Assessing Sensitive or Transgressive Behavior: A Systematic Review and Meta-Analysis'. *Frontiers in Psychology* 12:655592. doi: 10.3389/fpsyg.2021.655592.

Schnell, Rainer, and Kathrin Thomas. 2021. 'A Meta-Analysis of Studies on the Performance of the Crosswise Model'. *Sociological Methods & Research* 004912412199552. doi: 10.1177/0049124121995520.

Shamsipour, Mansour, Masoud Yunesian, Akbar Fotouhi, Ben Jann, Afarin Rahimi-Movaghar, Fariba Asghari, and Ali Asghar Akhlaghi. 2014. 'Estimating the Prevalence of Illicit Drug Use among Students Using the Crosswise Model'. *Substance Use & Misuse* 49(10):1303–10. doi: 10.3109/10826084.2014.897730.

Simon, Herbert A. 1957. *Models of Man: Social and Rational. Mathematical Essays on Rational Human Behavior in Society Setting.* New York: Wiley.

Vakilian, Katayon, Syyed Abbas Mousavi, and Afsaneh Keramat. 2019. 'Child Sexual Abuse Based on the Crosswise Model: A Cross-Sectional Study on 18–24-Year-Old Iranian Students'. *Family Medicine* 21(3):249–52.

Walzenbach, Sandra, and Thomas Hinz. 2019. 'Pouring Water into Wine: Revisiting the Advantages of the Crosswise Model for Asking Sensitive Questions'. *Survey Methods: Insights from the Field (SMIF)*. doi: 10.13094/SMIF-2019-00002.

Warner, Stanley L. 1965. 'Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias'. *Journal of the American Statistical Association* 60(309):63–69. doi: 10.1080/01621459.1965.10480775.

Wolter, Felix, and Andreas Diekmann. 2021. 'False Positives and the "More-Is-Better" Assumption in Sensitive Question Research'. *Public Opinion Quarterly* 85(3):836–63. doi: 10.1093/poq/nfab043.

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(3):251–63. doi: 10.1007/s00184-007-0131-x.

APPENDIX

A1) Cooperation Rate and Sampling Strategy

Table A1 presents cooperation rates (COOP1 according to AAPOR standards), that is, completed interviews per eligible and contacted individuals. This is done for each wave of the Konstanz Citizen Survey that had a CM experiment implemented. Since cooperation is generally much higher among already registered panel members (52-61%) than in the refreshment samples (20-24%), these are displayed separately.

Table A1. Cooperation Rate and Sampling Strategy

		n (eligible and contacted)	completed interviews
Wave 4 (2011)	Panel members	1249	666 (53%)
	Refreshment sample	1548	369 (24%) (244 online, 125 paper)
Wave 6 (2013)	Panel members	1280	782 (61%)
	Refreshment sample	2770 (1844 online first, 936 paper only)	581 (21%) 259 online, 322 paper
Wave 7 (2014)	Panel members	1577	869 (55%)
	Refreshment Sample	3009	635 (21%) 506 online, 129 paper
Wave 8 (2015)	Panel members	1885	982 (52%)
	Refreshment Sample	3279	640 (20%)

To ensure data quality, the target population and the covered respondent pool are regularly compared in terms of all demographic characteristics for which registry data is available: sex, age, citizenship and area of residence within the city. To account for lower response rates and higher panel attrition among young people and immigrants, citizens aged 18-30 and people with a non-German nationality were oversampled in the refresher samples since wave 4 (probability to be selected was doubled). Wave 8 was special in that its substantive focus was on neighborhoods. As a consequence, also inhabitants of underrepresented districts were oversampled.

Note that members of the refreshment samples who did not want to register for the online panel had an option to fill in a paper questionnaire in wave 4, 6 and 7 but were not automatically followed-up on in the next year. Also, the paper questionnaires did usually not contain the experiments on the crosswise model – with the exception of wave 6 where the number of paper respondents is too low for meaningful analyses.

A2) Question Wording of Crosswise Models

(1) Voter turnout
Is your mother‘s birthday in January, February or March?
Have you participated in the vote on „Stuttgart 21“?

- ☐ YES to both questions or NO to both questions
- ☐ YES to one question and NO to the other question

(2) Blood donation
Is your father’s birthday in January, February or March?
Have you donated blood in the past 12 months?

- ☐ YES to both questions or NO to both questions
- ☐ YES to one question and NO to the other question

(3) Littering
Is your mother‘s birthday in June or July?
Have you ever thrown anything on the streets or into nature instead of using a bin?

- ☐ YES to both questions or NO to both questions
- ☐ YES to one question and NO to the other question

(4) Keeping Change
Is your father‘s birthday in September or October?
Have you ever kept too much change that was accidently given to you?

response order 1: “same first”
- ☐ YES to both questions or NO to both questions
- ☐ YES to one question and NO to the other question

response order 2: “different first”
- ☐ YES to one question and NO to the other question
- ☐ YES to both questions or NO to both questions

(5) Jaywalking
Is your father‘s birthday in September or October?
Have you ever passed a red traffic light when there wasn’t any car in sight?

response order 1: “same first”
- ☐ YES to both questions or NO to both questions
- ☐ YES to one question and NO to the other question

response order 2: “different first”
- ☐ YES to one question and NO to the other question
- ☐ YES to both questions or NO to both questions

A3) Panel Conditioning / Learning Effects by Sociodemographic Background

Since respondents with different experience levels differ in sample composition (due to non-random panel attrition), regression models of CM prevalence on experience were run separately for respondents with comparable sociodemographic background. Results are presented for combinations of characteristics where case numbers allowed a comparison of at least 10 respondents without previous experience to at least 10 respondents with previous experience. The findings show no significant learning effects for any of the groups, with most coefficients pointing towards the opposite direction.

Table A3. Learning effects by sociodemographic background

	respondent characteristics			case numbers	regression results (experience effect on CM)	
	Sex	age	education	experience (no – yes)	coefficient	significance
blood donation	Male	31-59	low	19 - 23	+0.36	0.22
(negative effect of experience expected if people learn)	Male	31-59	high	55 - 50	+0.11	0.57
	Male	60+	low	15 - 15	0	1.00
	Male	60+	high	36 - 21	-0.1	0.72
	female	31-59	low	30 - 18	+0.02	0.93
	female	31-59	high	44 - 28	+0.34	0.15
	female	60+	low	17 - 12	+0.36	0.32
jaywalking (all)	Male	18-30	middle	25 - 13	+0.41	0.59
(positve effect of experience expected if people learn)	Male	18-30	high	23 - 13	-0.22	0.30
	Male	31-59	low	18 - 26	-0.13	0.54
	Male	31-59	high	26 - 54	-0.13	0.41
	Male	60+	low	10 - 18	-0.38	0.21
	female	18-30	middle	37 - 16	+0.11	0.45
	female	18-30	high	35 - 24	-0.03	0.85
	female	31-59	low	13 - 26	-0.58	0.02
	female	31-59	middle	11 - 15	+0.11	0.68
	female	31-59	high	33 - 25	+0.05	0.78
	female	60+	low	12 - 20	-0.4	0.10
jaywalking	female	18-30	middle	20 - 11	+0.01	0.94
(order1)	female	18-30	high	14 - 10	-0.34	0.15
	female	31-59	high	12 - 13	-0.13	0.73
jaywalking	Male	31-59	low	11 - 12	-0.1	0.71
(order2)	Male	31-59	high	18 - 30	-0.17	0.42
	female	18-30	high	21 - 14	+0.18	0.46
	female	31-59	high	21 - 12	-0.16	0.46

education levels: low = below high school, middle = high school diploma, high = university degree