Survey Research Methods (2023)
Vol. 17, No. 1, pp. 1-12
doi:10.18148/srm/2023.v17i1.8010
European Survey Research Association

Puzzling Answers to Crosswise Questions: Examining Overall Prevalence Rates, Response Order Effects, and Learning Effects

Sandra Walzenbach and Thomas Hinz
University of Konstanz, Department of Sociology

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 into two potential explanatory mechanisms that have not yet received attention in empirical studies: response order effects and learning via repeated exposure. We do not find consistent response order 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.

Keywords: crosswise model; randomized response; social desirability bias; response order effects; learning effects; privacy concerns

1 Introduction

The crosswise model (CM) has lately received a lot of attention in sensitive question research. Proposed by Yu et al. (2008), the method was designed to free survey responses from social desirability bias. The idea was that by ensuring complete anonymity, respondents would no longer feel the pressure to present themselves in a favourable light and provide honest responses, even if they are sensitive. In contrast to its predecessors, the crosswise model does not require a random device, making it easy to implement in every survey mode. In short, it was hoped to overcome some of the key flaws associated with Randomized Response Techniques (RRT) that have been documented by scientific studies for several decades (Edgell et al., 1982; John et al., 2018; Locander et al., 1976; Van der Heijden et al., 2000; Wolter & Preisendörfer, 2013).

Many studies drew—and we believe that too many keep drawing—positive conclusions regarding the crosswise model. Concerns about its validity have been raised since 2014 (Höglinger et al., 2014; Walzenbach & Hinz, 2014) and made it into publications since 2017 (Höglinger & Diekmann, 2017; Jerke et al., 2019; Krause & Wahl, 2022; Kuhn & Vivyan, 2018; Walzenbach & Hinz, 2019). Nonetheless, the crosswise model is still implemented by researchers who believe in its attenuating effects on social desirability bias (Canan et al., 2021; Hopp & Speil, 2021; Mieth et al., 2021) or do not account for potential problems in their design (Jerke et al., 2022). Even two recent meta-analyses paint a rather positive picture: While one admittedly at least suspects problems with less educated respondents and with publication bias favouring significant results (Schnell & Thomas, 2021), one considers the crosswise model a "promising" method (Sagoe et al., 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—a result that is widely interpreted as a successful reduction in social desirability bias. The underlying reasoning is that the crosswise model must have worked if more people admitted a socially undesirable behaviour.

In contrast to these views, we argue that this is not a valid indicator for data quality. Instead, the CM estimator is systematically biased towards 50% whenever respondents disobey the instructions (inadvertently or deliberately) and tick answers randomly (for a detailed discussion see Höglinger & Diekmann, 2017, Appendix C). In other words, we observe the same response pattern when the crosswise model works and when it does not work. This problem applies to all cases in which a socially undesirable behaviour with low prevalence is assessed—that is, in the overwhelming majority of all existing CM experiments. In other words, close to all existing studies, including those analysed in meta-analyses, are based on a problematic comparison between direct question and crosswise model, leading them to unjustifiably positive conclusions. Maybe worse, it also keeps most authors from

Corresponding author: Sandra Walzenbach, University of Konstanz, Department of Sociology, Universitätsstraße 10, 78464 Konstanz, Germany (E-mail: sandra.walzenbach@uni-konstanz.de)

looking at the underlying mechanisms that cause bias in CM estimates.

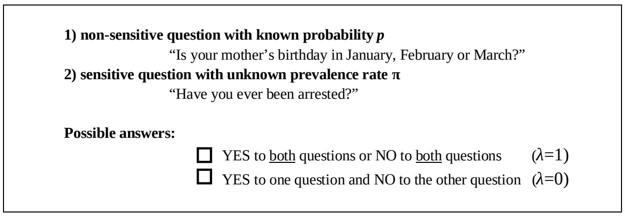
In light of these current gaps in CM research, this paper theoretically explains why the assumption that “more is better” is faulty in the overwhelming majority of all existing CM applications and provides insights into the applicability of the crosswise model in a general population sample. Over the time span of several years, we implemented five experiments on the validity of the crosswise model in a panel survey that targets all registered citizens in the town of Konstanz in Southern Germany. 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. In contrast to the overwhelming majority of other studies, we use a design that relies on socially desirable behaviours with low prevalence rate (explained in detail in Section 2.2). 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 response order effects and learning via repeated exposure to the CM procedure.

Last but not least, the study uses panel data from a general population sample to examine the applicability of the crosswise model. Validation studies with such samples are still extremely rare—but very important. There is a widespread belief that CM formats (and related Randomized Response Techniques) produce bias because they are too complicated for most respondent groups. If this is the case, testing it on university student samples or convenience samples of academics will necessarily not reveal the whole scale of the problem.

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 4.1), discuss potential response order effects (Section 4.2), and learning via repeated exposure to the question format (Section 4.3).

2 The Crosswise Model and Common Flaws in Previous Research

2.1 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 <u>both</u> questions or NO to <u>both</u> questions (λ=1)
□ YES to one question and NO to the other question (λ=0)

Figure 1

Basic Logic of the Crosswise Model

In a nutshell, the crosswise model (Yu et al., 2008) combines two dichotomous questions into one response task: the sensitive question of interest with an unknown prevalence rate π and a non-sensitive question with a known prevalence p (see example in Figure 1). Respondents only provide information as to whether their answers to these two questions are equal ($\lambda = 1$) or different ($\lambda = 0$). This means that there is no socially undesirable or revealing response option. Assuming that respondents answer more honestly to the crosswise model than a direct question, it should provide more accurate overall prevalence rates for the sensitive behaviour. That is, the detrimental effects of social desirability bias to data quality should be attenuated or even disappear.

Knowing that the first response category 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 a given λ and p by using the formula $\lambda = p\pi + (1-p)(1-\pi)$, which transforms to

$$\pi = \frac{\lambda + p - 1}{2p - 1} \quad .$$

The crosswise model is structurally equivalent to Warner’s original Randomized Response Technique (Warner, 1965), in which respondents are directed to a sensitive question or its negation by means of a random device. One crucial reason why the crosswise model was initially celebrated as an advancement to Warner’s approach is that it does not rely on any random device, which in theory makes it feasible to implement in self-administered web surveys.

Both, Warner’s original question format and the crosswise model, have in common that they add random noise and sacrifice statistical precision for the sake of greater respondent privacy. In practice, this is a serious trade-off. The variance of a CM estimate is given by

$$\mathrm{Var}(\pi) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \quad ,$$

which reduces to the variance of a direct question for $p = 0$ or $p = 1$, and exponentially grows as p approaches 0.5 (for technical details, see Yu et al. 2008). Speaking in practical terms, this means that for a sensitive item with prevalence $\pi = 0.1$ the variance of a crosswise model with $p = 1/6$ would be inflated by the factor 4.4 compared to a direct question, a crosswise model with $p = 1/4$ by the factor 9.3. For practitioners it is hence essential to know if this trade-off pays off in terms of more valid answers from their respondents.

2.2 Common flaws: The assumption that "more is better" is usually wrong

As argued above, 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 & Musch, 2016; Höglinger et al., 2016; Jann et al., 2012; Jerke et al., 2022; Korndörfer et al., 2014; Shamsipour et al., 2014). Typically, the crosswise estimate is compared to an experimental condition with a direct question and a higher CM than DQ prevalence is interpreted as a successful reduction in social desirability bias. The fundamental problem with this approach is that there is another mechanism that can produce the same results: Respondents might be unable or unwilling to comply with the procedure and tick answers randomly instead. Whenever this happens, the CM estimate is biased towards 50%. This means that respondents who are confused and/or do not comply with the procedure for other reasons produce the same response pattern as a crosswise model that successfully reduces social desirability bias.

To understand more generally which CM applications come with this problem, we need to take into account if socially desirable or undesirable behaviour is assessed, and if its prevalence is below or above 50%. As Figure 2 illustrates, the upper right and the lower left cell show the same response patterns for a crosswise model that reduces bias and for one that suffers from random ticking. That is, whenever we assess socially undesirable behaviours with low prevalence rates of under 50% (such as having been arrested) or socially desirable behaviours with high prevalence rates of above 50% (such as paying taxes). If in the CM condition more people admitted an arrest or fewer people claimed to pay their taxes compared to a direct question, these patterns could likewise reflect a successful reduction of social desirability bias or a bias towards 50% because respondents did not follow the procedure but ticked answers randomly. We just cannot know which one is true.

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 upper left and lower right cell in Figure 2). For these two scenarios, the CM estimate points into different directions compared to the direct question: The CM prevalence should be further away from 50% than the DQ prevalence if the model works and successfully reduces social desirability bias, but closer to 50% than the DQ prevalence if there is a problem with random ticking.

In what follows, we will come back to this distinction and present some experimental designs that allow us to distinguish random ticking from a successful reduction of social desirability bias. For the moment we conclude that most CM applications in the scientific literature do not allow researchers to detect possible problems at all. On the contrary, many of the reported CM estimates potentially suffer from undetectable bias because they usually elicit undesirable and rare behaviour.

2.3 What do we know about underlying mechanisms?

Reflecting the problem that CM estimates tend towards 50% whenever respondents do not comply with the procedure and tick answers randomly, a part of the recently published studies on the crosswise model have focused on examining false positive and/or false negative answers. Often, they have drawn very skeptical conclusions (e.g. Höglinger & Jann, 2018; Kuhn & Vivyan, 2018; Walzenbach & 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 et al., 2020). The cognitive burden is assumed to trigger satisficing (Krosnick, 1991; Simon, 1957). Survey methodologists typically use self-reported comprehension or education as indicators for cognitive load and risk of satisficing. However, empirical studies on the crosswise model 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 (Höglinger & Diekmann, 2017, Appendix C; Jerke et al., 2019; Meisters et al., 2020; Walzenbach & Hinz, 2019; Wolter & Diekmann, 2021).

Few studies have discussed or analysed concrete forms of satisficing in crosswise models. From a theoretical perspective, it would be straightforward to avoid cognitive burden by ignoring the question instructions altogether and ticking an answer randomly (as suggested by Höglinger and Diekmann, 2017 who calculate potential shares of random tickers to explain false positive answers in the appendix[1]). Alternatively, respondents could choose a response category based on its more salient position. According to the general survey literature one would expect a preference for the first category, whenever a survey is presented visually (Tourangeau et al., 2010, 304f). Empirical tests of this latter hypothesis

[1] The share of random tickers can easily be included into the formula for the CM estimate. If we assume a share of random tickers R with a λ of 0.5, while the remaining respondents $(1 - R)$ comply with the procedure, the new estimate is given by $\lambda = (p\pi + (1 - p)(1 - \pi)) \cdot (1 - R) + (R/2)$.

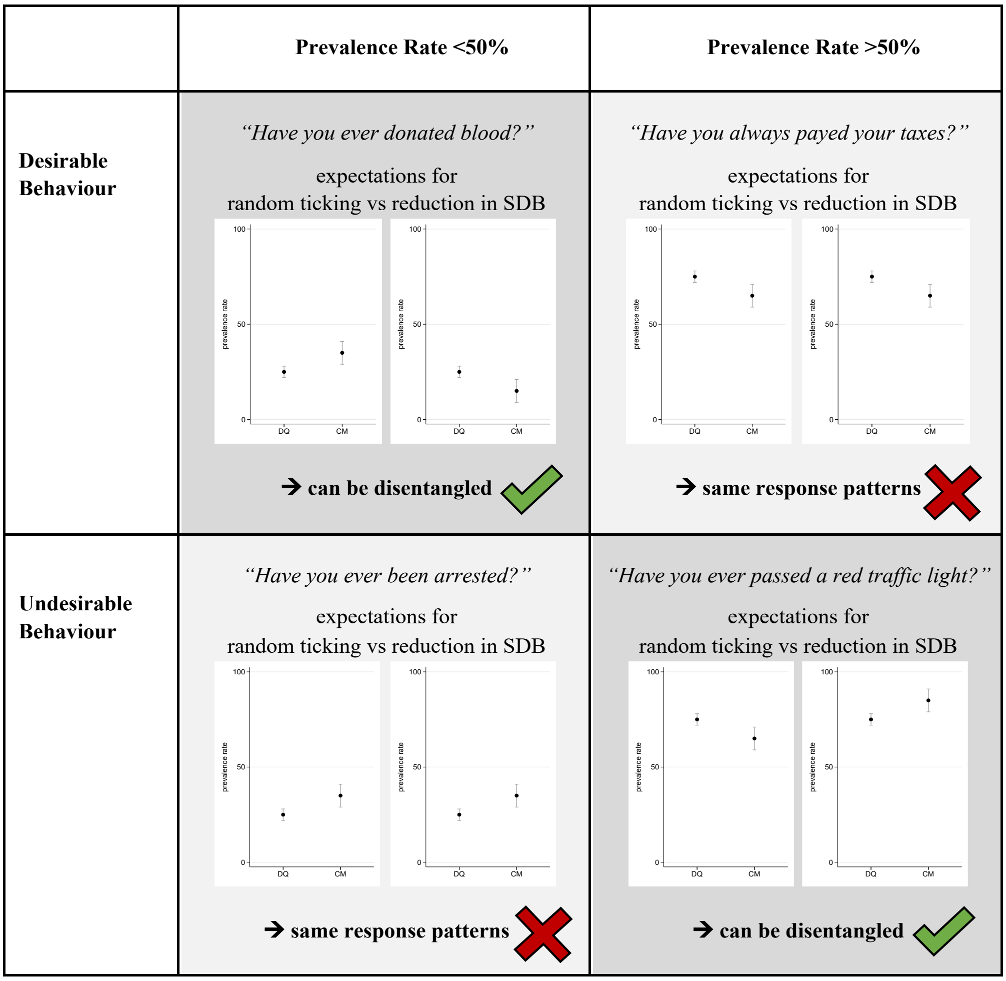


Figure 2

Random Ticking and Reduction of Social Desirability Bias (SDB) dependent on prevalence and type of behaviour

are rare and usually do not find any effect of response order (Höglinger & Diekmann, 2017, Appendix C; Wolter & Diekmann, 2021).

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

3.1 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 response order effects for bias is a follow-up research question that was inspired by the puzzling results of our experiment on blood donation (see Section 3.3 for details). In a previous paper, we tried to explain the method's failure to reduce bias by indicators that are traditionally related to satisficing. However, similar to other studies we could not find any significant correlations between respondent characteristics, such as age and education, and biased CM estimates. Instead, we consistently found upwardly biased CM estimates irrespective of respondent characteristics (for details, see Walzenbach & Hinz, 2019).

For non-sensitive items with prevalence $p < 0.5$, the estimated prevalence rate of the sensitive item increases with decreasing λ (lower share of respondents ticking the first response category), meaning that a general preference for the second response category could theoretically have explained our results. Although this pattern contradicts the more general view that first categories should be more salient in visual survey modes, it seems worth exploring response orders in an experimental setting.

The idea to analyse learning effects is very much based on theoretical arguments from previous research. If the problem really lies in the complexity of the format and/or lacking motivation for cognitive involvement, repeated exposure to the crosswise model should give respondents more time to process, multiple opportunities to thoroughly read, learn and understand.

As we are lucky enough to have panel data, we will go beyond what other studies have done by measuring comprehension and look at the aspect of complexity 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.

3.2 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[2], which targets the registered citizens in a town in the south of Germany. Data are 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 Online-Appendix A1 for more details on sampling strategy, cooperation rates and socio-demographics of respondents). 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 and be part of the realized sample. However, as we rely on experimental approaches with random assignment, we do not expect these characteristics to compromise our results in any way.

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.

3.3 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 Online-Appendix A2 for exact question wordings). In all of them, respondents were randomly assigned to a crosswise model or a direct question. This means that respondents might have answered a direct question in one wave and a crosswise model in another.

In addition to the common but somewhat error-prone strategy to validate the crosswise model's performance, experiments (2) and (5) assess desirable behaviour with low prevalence and undesirable behaviour with high prevalence. As explained in Section 2.2, this design allows us to distinguish a reduction in social desirability bias from the effects of random ticking.

For two experiments, we have external data sources from which we derive the true aggregate-level prevalence rate for validation: For experiment (1), we refer to statistics on voter turnout available from the town council. For experiment (2), we use data on blood donors in Konstanz that were facilitated upon request by the Red Cross, which administers all local blood donation campaigns. Experiments (4) and (5) varied how the response categories in the crosswise model were presented to examine potential order effects on bias in estimates.

All experiments are listed in Table 1. They appear in the order of their implementation into the panel survey waves.

3.4 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 (in Section 4.1). For each experiment, Table 1 explicitly states the expected response patterns if the crosswise model predominantly reduces social desirability bias versus if the results are predominantly driven by random ticking. Where an external validation criterion is used, the CM estimate should be closer to the true value than the DQ estimate if the technique reduces social desirability bias as it should. Note that due to random assignment to the experimental conditions, the general conclusions 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 4.2). 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 4.3). For these analyses, we divide respondents into those who are answering a crosswise model for the first time, and those who already had been assigned to a crosswise model in one or more

[2] https://www.buergerbefragung-konstanz.de; for reasons of data protection, we unfortunately cannot deposit the data online. Researchers can access the data on site at Konstanz University upon request.

Table 1

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%
expectation if CM reduces SDB	CM < DQ	CM < DQ	CM > DQ	CM > DQ	CM > DQ
expectation if respondents tick randomly	CM < DQ	CM > DQ	CM > DQ	CM > DQ	CM < DQ
disentangling of mechanisms possible	no	yes	no	no	yes
external validation criterion available	yes	yes	no	no	no
experiment on order of response categories	no	no	no	yes	yes

previous panel waves and hence have experience with the question format. We first compare the CM estimates of these groups descriptively and then move on to some robustness checks that take non-random panel-attrition into account.

All reported significance tests for differences between estimates 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, with p indicating the prevalence of the non-sensitive item and λ denoting the share of respondents that ticked the first response category (for details see Jann et al., 2012). A linear probability model was chosen over a logistic regression because results are more robust. This is particularly important as we expect some non-compliance with the procedure.[3] We provide stata code for all presented analyses in the supplementary online material.

4 Results

4.1 Prevalence rates from five crosswise experiments

Figure 3 shows the estimated overall prevalence rates for all five experiments in the order of their implementation into the panel survey. For each experiment, the DQ condition is compared to the CM condition. Each estimate is shown with its 95% confidence interval. The dashed line highlights the 50%-prevalence threshold that the CM estimates tend towards in case of problems with random ticking.

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 from the official statistics (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, suggesting that (unsurprisingly) our sample is not suitable to estimate voter turnout in the target population. However, this is unproblematic as we are interested in the comparison of experimental conditions rather than absolute values: The crosswise model should yield an estimate that is

[3] Following a recommendation from the review process, we calculated all significance tests with a logistic regression model as a robustness check (Jann, 2005). Most of the time, the differences are negligible. For the overall prevalences of experiments 2 and 4 (reported in Figure 3), the differences between the CM and the DQ format become even more significant in the logistic model (especially the blood donation item yielded a p-value of 0.003 instead of 0.037). The opposite is true for some of the (negative) learning effects reported in Table 2, where the order 2 conditions of experiment 5 yielded p-values above the 0.05 mark (namely 0.084 and 0.18 instead of 0.015 and 0.047). However, we consider these results as less reliable than those based on a linear model.

closer to the true value than the direct question if it reduces 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%. 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 ($p = 0.037$). External validation data from the Red Cross suggests a true prevalence rate of below 4%[4] (for more details, see Walzenbach & Hinz, 2019).

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.

4.2 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 response order effect in experiment (4) (see first row of results in Figure 4). 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 that respondents partly apply a heuristic response strategy. As a consequence, the estimated prevalence rates stemming from the two different response orders 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 last row in Figure 4). 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 preference for the last instead of the first response category.

Although the differences between the DQ and CM conditions in experiments (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 seeing reproducible results on response order effects, our findings suggest that response behaviour is inconsistent and might be susceptible to minor differences, e.g. in question wording or survey setting.

4.3 Results on learning effects due to repeated exposure

We argued that data quality should improve when respondents are repeatedly confronted with a crosswise model and have the chance to learn. Table 2 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 due to previous exposure.

The columns represent the prevalence rates from the direct question and the crosswise model for different levels of familiarity with the crosswise model (without/with previous experience[5]). 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

[4] From the Red Cross data, we could obtain information on the number of people who donated blood for the first time as well as on the total number of donations from repeated donors, but we do not know how often the repeated donors participated within the time period of interest. Dividing the number of blood donations registered in Konstanz in 2012 by the number of inhabitants over 18 at that time yields a prevalence rate of 4%. In the likely event that some people donated more than once, this estimate would decrease even further. We therefore assume that the true value ranges below 4%. However, it is possible that respondents donated blood outside of Konstanz. This seems unlikely for most citizens, but could in principal apply to those university students who regularly go to visit their families outside of Konstanz.

[5] 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.

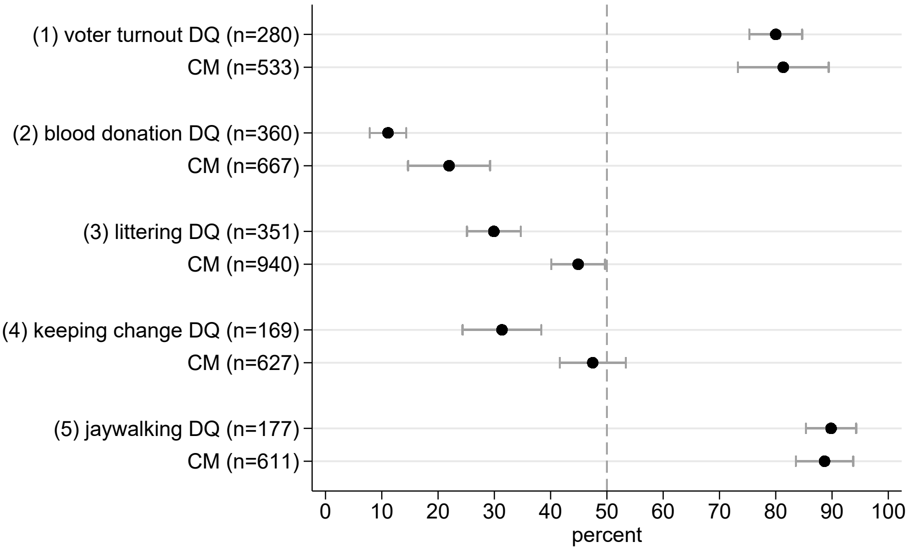


Figure 3

Estimated Prevalence Rates in DQ and CM conditions: Konstanz Citizen Survey (online-panel respondents from waves 4, 6, 7, 8)

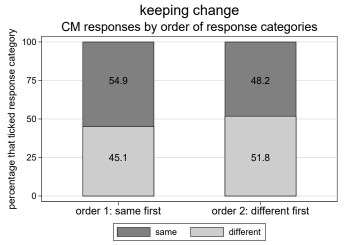
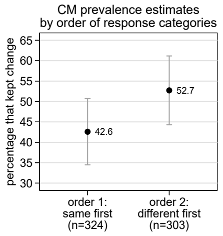
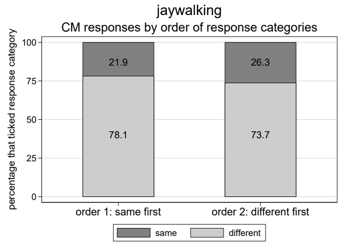
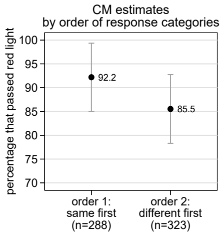
	ticked response categories	prevalence rates by response order
Experiment 4: keeping too much change		
Experiment 5: jaywalking		

Figure 4

Results on Order Effects for Experiments (4) and (5). Data: Konstanz Citizen Survey (wave 8). Terms and definitions: “same” refers to the response category “YES to both questions or NO to both questions”, “different” refers to the response category “YES to one question and NO to the other question”, “order 1: same first” refers to the experimental condition where same was the first response category, “order 2: different first” refers to the experimental condition where different came first.

provide two types of additional analyses to account for this. First, the last column of Table 2 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 Online-Appendix A3).

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 when sociodemographic characteristics are controlled ($p = 0.047$). 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 Online-Appendix A3).

The findings related to learning effects 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.

4.4 Discussion

Summing up, this validation study casts serious doubts about the applicability of the crosswise model in general population samples. We presented empirical evidence from five experiments that were implemented in the Konstanz Citizen Survey 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 learning effects due to previous exposure 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 & Hinz, 2019; Wolter & Diekmann, 2021) but contradicts the argumentation of others (Schnell & 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. However, the fact that different response orders can trigger considerable differences in prevalence rates shows the crosswise model's 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 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 (Sagoe et al., 2021; Schnell & Thomas, 2021), which are based on the often problematic comparisons of the crosswise model to direct questions (as discussed above).

Generally, our study leaves many open questions concerning the mechanisms that cause the response patterns we observe. We examined response order effects and learning via 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 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

Table 2

Results on Learning via Repeated Exposure

	Crosswise models				Significance tests from regression	
	without experience		repeated exposure		w/o controls	with controls[a]
	%	*n*	%	*n*	*p*	*p*
(2) Blood donation	21	459	25	208	0.580	0.230
(5) Jaywalking: order 1	95	128	90	160	0.570	0.960
(5) Jaywalking: order 2	94	162	77	161	0.015	0.047

Data: Konstanz Citizen Survey (online-panel respondents from waves 4, 6, 7, 8)
[a] Sex, age, education

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

Canan, C. E., Chander, G., Moore, R., Alexander, G. C., & Lau, B. (2021). Estimating the prevalence of and characteristics associated with prescription opioid diversion among a clinic population living with HIV: Indirect and direct questioning techniques. *Drug and Alcohol Dependence*, *219*, 108398. https://doi.org/10.1016/j.drugalcdep.2020.108398

Coutts, E., Jann, B., Krumpal, I., & Näher, A.-F. (2011). Plagiarism in student papers: Prevalence estimates using special techniques for sensitive questions. *Jahrbücher für Nationalökonomie und Statistik*, *231*(05-06), 749–760. https://doi.org/10.1515/jbnst-2011-5-612

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods & Research*, *11*(1), 89–100. https://doi.org/10.1177/0049124182011001005

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques. A stochastic lie detector versus the crosswise model. *Behavior Research Methods*, *48*, 1032–1046. https://doi.org/https://doi.org/10.3758/s13428-015-0628-6

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the crosswise-model RRT. *Political Analysis*, *25*(1), 131–137. https://doi.org/https://doi.org/10.1017/pan.2016.5

Höglinger, M., Diekmann, A., & Jann, B. (2014). Telling the truth? Results from an experimental validation of sensitive question techniques. [Presentation at the Conference on Analytical Sociology, Venice].

Höglinger, M., & Jann, B. (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. https://doi.org/10.1371/journal.pone.0201770

Höglinger, M., Jann, B., & Diekmann, A. (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), 171–187. https://doi.org/10.18148/srm/2016.v10i3.6703

Hopp, C., & Speil, A. (2021). How prevalent is plagiarism among college students? Anonymity preserving evidence from Austrian undergraduates. *Accountability in Research-Policies and Quality Assurance*, *28*(3), 133–148. https://doi.org/10.1080/08989621.2020.1804880

Jann, B. (2005). "RRLOGIT: Stata module to estimate logistic regression for randomized response data" [Statistical Software Components S456203, Boston College Department of Economics, revised 12 May 2011.].

Jann, B. (2008). RRREG: Stata module to estimate linear probability model for randomized response data [Statistical Software Components S456962, Boston College Department of Economics, revised 12 May 2011.].

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model. An experimental survey measuring plagiarism. *Public Opinion Quarterly*, *76*(1), 32–49. https://doi.org/10.1093/poq/nfr036

Jerke, J., Johann, D., Rauhut, H., & Thomas, K. (2019). Too sophisticated even for highly educated survey respondents? A qualitative assessment of indirect question formats for sensitive questions. *Survey Research Methods*, *13*(3), 319–351. https://doi.org/10.18148/srm/2019.v13i3.7453

Jerke, J., Johann, D., Rauhut, H., Thomas, K., & Velicu, A. (2022). Handle with care: Implementation of the list experiment and crosswise model in a large-scale survey on academic misconduct. *Field Methods*, *34*(1), 69–81. https://doi.org/10.1177/1525822X20985629

John, L. K., Loewenstein, G., Acquisti, A., & Vosgerau, J. (2018). When and why randomized response techniques (fail to) elicit the truth. *Organizational Behavior and Human Decision Processes*, *148*, 101–123. https://doi.org/10.1016/j.obhdp.2018.07.004

Khosravi, A., Mousavi, S. A., Chaman, R., Khosravi, F., Amiri, M., & Shamsipour, M. (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). https://doi.org/10.5812/ijhrba.24388v2

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology*, *45*(1), 18–32. https://doi.org/10.1016/j.joep.2014.08.001

Krause, T., & Wahl, A. (2022). Non-compliance with indirect questioning techniques: An aggregate and individual level validation. *Survey Research Methods*, *16*(1), 45–60. https://doi.org/10.18148/SRM/2022.V16I1.7824

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kuhn, P. M., & Vivyan, N. (2018). Reducing turnout misreporting in online surveys. *Public Opinion Quarterly*, *82*(2), 300–321. https://doi.org/10.1093/poq/nfy017

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, *71*(2), 269–275. https://doi.org/10.2307/2285297

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? *PLoS ONE*, *15*(6), e0235403. https://doi.org/10.1371/journal.pone.0235403

Mieth, L., Mayer, M. M., Hoffmann, A., Buchner, A., & Bell, R. (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. https://doi.org/10.1186/s12889-020-10109-5

Sagoe, D., Cruyff, M., Spendiff, O., Chegeni, R., de Hon, O., Saugy, M., van der Heijden, P. G. M., & Petróczi, A. (2021). Functionality of the crosswise model for assessing sensitive or transgressive behavior: A systematic review and meta-analysis. *Frontiers in Psychology*, *12*, 655592. https://doi.org/10.3389/fpsyg.2021.655592

Schnell, R., & Thomas, K. (2021). A meta-analysis of studies on the performance of the crosswise model. *Sociological Methods & Research*, 004912412199552. https://doi.org/10.1177/0049124121995520

Shamsipour, M., Yunesian, M., Fotouhi, A., Jann, B., Rahimi-Movaghar, A., Asghari, F., & Akhlaghi, A. A. (2014). Estimating the prevalence of illicit drug use among students using the crosswise model. *Substance Use & Misuse*, *49*(10), 1303–1310. https://doi.org/10.3109/10826084.2014.897730

Simon, H. A. (1957). *Models of man: Social and rational. Mathematical essays on rational human behavior in society setting*. Wiley.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2010). *The psychology of survey response*. Cambridge University Press.

Van der Heijden, P. G., van Gils, G., Bouts, J., & Hox, J. J. (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*(4), 505–537. https://doi.org/10.1177/0049124100028004005

Walzenbach, S., & Hinz, T. (2014). Pouring water into wine. The advantages of the crosswise model asking sensitive questions revisited. Presentation at the conference on analytical sociology, venice [Presentation at the Conference on Analytical Sociology, Venice]. https://www.soziologie.uni-muenchen.de/venedig/index.html

Walzenbach, S., & Hinz, T. (2019). Pouring water into wine: Revisiting the advantages of the crosswise model for asking sensitive questions. *Survey Methods: Insights from the Field (SMIF)*. https://doi.org/10.13094/SMIF-2019-00002

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69. https://doi.org/10.1080/01621459.1965.10480775

Wolter, F., & Diekmann, A. (2021). False positives and the "more-is-better" assumption in sensitive question research. *Public Opinion Quarterly*, *85*(3), 836–863. https://doi.org/10.1093/poq/nfab043

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research*, *42*(3), 321–353. https://doi.org/10.1177/0049124113500474

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*(3), 251–263. https://doi.org/10.1007/s00184-007-0131-x