Substance or Noise?
On the Retest Stability of Responses to Direct and Indirect Questions

Julia Meisters, Adrian Hoffmann, Jochen Musch

Article - Version of Record

Suggested Citation:
Meisters, J., Hoffmann, A., & Musch, J. (2024). Substance or Noise?: On the Retest Stability of Responses to Direct and Indirect Questions. European Journal of Psychological Assessment, 41(5), 393–401. https://doi.org/10.1027/1015-5759/a000813

Original Article

Substance or Noise?

On the Retest Stability of Responses to Direct and Indirect Questions

Julia Meisters, Adrian Hoffmann, and Jochen Musch

Department of Experimental Psychology, Heinrich Heine University Duesseldorf, Germany

Abstract: Indirect questioning techniques aim to provide more valid prevalence estimates for sensitive attributes than conventional direct questions. Despite being an important prerequisite for high estimation validity, indirect questioning techniques' retest stability has rarely been addressed. For temporally stable attributes, high stability of both prevalence estimates and individual responses is expected; however, insufficient understanding of the instructions and random response behavior may compromise retest stability. The present study is the first to assess the retest stability of the Extended Crosswise Model (ECWM), a recent indirect questioning technique, and to compare it to the retest stability of a conventional direct question (DQ). With a retest interval of approximately 10 days, we asked N = 2,317 mothers twice whether they had smoked during a previous pregnancy. In both ECWM and DQ conditions, prevalence estimates were virtually identical over time, and most respondents answered consistently (ECWM: 89%, DQ: 95%). In the ECWM condition, inconsistent response behavior was slightly more prevalent and negatively associated with respondents' education. However, as these effects were small, the retest stability of both ECWM and DQ in surveys on sensitive attributes was evaluated as high.

Keywords: retest stability, indirect questioning techniques, extended crosswise model, validity, smoking during pregnancy

Sensitive attributes are hard to capture in survey research because many respondents hesitate to provide truthful answers due to fear of punishment or social disapproval. Indirect questioning techniques such as the Randomized Response Technique (RRT; Warner, 1965) may help to control social desirability bias. Based on a random distortion of responses, the RRT guarantees individual responses' confidentiality and encourages honest responses.

The Crosswise Model (CWM; Yu et al., 2008) is a recent improvement of the RRT frequently used in empirical research. In the CWM, two statements are shown to the respondents: A sensitive statement A with unknown prevalence π (e.g., "I have smoked during at least one pregnancy"), and a non-sensitive statement B with known prevalence p used for randomization (e.g., "I was born in November or December"). Respondents are asked to indicate whether they agree (a) with both statements or none of them; or (b) with exactly one of the statements (irrespective of which one). The following formula provides a maximum likelihood estimate for the prevalence π of the sensitive attribute on the sample level (Yu et al., 2008):

$$\hat{\pi}_{\mathrm{CWM}} = \frac{\frac{n'}{n} + p - 1}{2 \times p - 1}, \qquad p \neq \frac{1}{2}, \tag{1}$$

with n' representing the total number of "both/ none" responses and n being the sample size. Figure 1 shows a tree diagram of the CWM.

A major advantage of the CWM is that it is easier to understand than other indirect questioning techniques (Hoffmann et al., 2017). Moreover, none of its answer alternatives reveals whether a respondent carries a sensitive attribute, thereby granting complete confidentiality of the answer and making the CWM robust to faking (Hoffmann et al., 2021). However, the CWM is more complex than simple direct questioning and less efficient, thus requiring larger sample sizes (Ulrich et al., 2012). Moreover, respondents need considerably more time to answer questions in the CWM format than to answer simple direct questions (Meisters et al., 2020a). Therefore, its application is only justified if it increases the validity of prevalence estimates. The Extended Crosswise Model (ECWM; Heck et al., 2018) is a recent extension of the CWM to two groups with different randomization probabilities, p_1 and p_2. The ECWM allows testing model fit to reveal systematic response biases, such as a systematic preference for one of the response options, without loss of efficiency compared to the CWM.

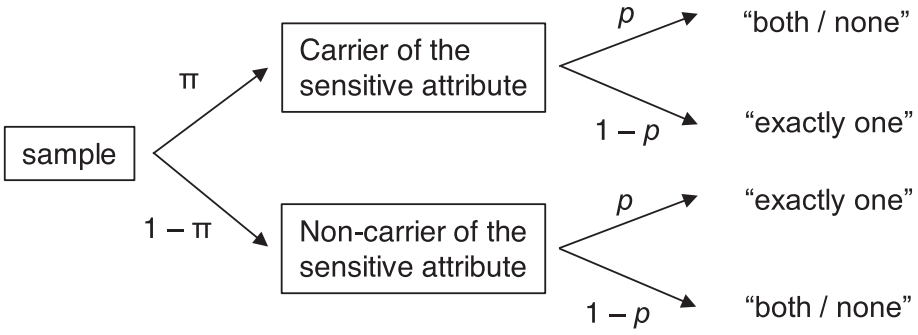

Figure 1. Tree diagram of the crosswise model. The parameter π represents the unknown prevalence of the sensitive attribute, which has to be estimated; *p* represents the known randomization probability.

Direct questioning (DQ) and (E)CWM prevalence estimates have repeatedly been compared (e.g. Hoffmann et al., 2020; Hoffmann & Musch, 2019; Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; Jann et al., 2012; Korndörfer et al., 2014; Mayer et al., 2021; Meisters et al., 2020b, 2022a; Mieth et al., 2021; Nakhaee et al., 2013; Thielmann et al., 2016). Meta-analyses found that CWM estimates tend to be more valid than DQ estimates, as they are often higher for socially undesirable and lower for socially desirable attributes (Sagoe et al., 2021; Schnell & Thomas, 2023). However, strong validation studies with a known true prevalence of the sensitive attribute and known actual status of individual respondents showed mixed results. Substantial rates of false positives (noncarriers of the sensitive attribute misclassified as carriers) and false negatives (carriers of the sensitive attribute misclassified as noncarriers) were observed both in the (E)CWM and DQ (e.g., Hoffmann et al., 2015; Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; Meisters et al., 2020a). False positives could potentially be caused by respondents being over-challenged by the relatively complex instructions of the (E)CWM and therefore deciding to choose an answer option at random. This would lead to a convergence of prevalence estimates toward 50% in the (E)CWM, thus providing an alternative explanation for studies that found higher prevalence estimates for socially undesirable and lower prevalence estimates for socially desirable attributes in the (E)CWM without knowing the true prevalence of the sensitive attribute (Höglinger & Diekmann, 2017; Schnapp, 2019; Walzenbach & Hinz, 2019). However, a recent study disentangled the influence of successful control of social desirability on the one hand and random responding on the other hand. The results suggested that the differences observed between ECWM and DQ estimates can be traced back to the improved control of social desirability afforded by the ECWM rather than the influence of random responses (Meisters et al., 2022a).

Retest Stability of Indirect Questioning Techniques

Indirect questioning techniques, such as the (E)CWM, are intended to provide valid measurements of sensitive attributes. A central prerequisite for the validity of an indirect questioning technique is the retest stability of its estimate. If a sensitive attribute is temporally stable, prevalence estimates for this attribute should be equal at two different measurement time points. Even more so, when both the sensitive and the non-sensitive attributes used for randomization are temporally stable, respondents should give the same answer at both measurement points. Despite its importance for the validity of indirect questioning techniques, retest stability has very rarely been addressed; only Pu et al. (2016) showed in a correlational analysis and for the Simmons model that two repeated sample surveys employing stratified cluster sampling correlated highly (r = .88).

The Present Study

The present study reports the first assessment of the retest stability of the ECWM, the most recent extension of the CWM that, unlike the original CWM, allows researchers to test whether observed responses fit the model used to determine prevalence estimates. For comparison, we also decided to assess the retest stability of a direct question whose temporal stability should not be taken for granted either; for example, test-retest correlations ranged from only .35–.78 for 18 single-item measures in Fisher et al. (2016).

With a retest interval of approximately 10 days, we asked female respondents about their sensitive smoking behavior during a previous pregnancy. The sample consisted of mothers who were not pregnant during the survey, so their status concerning the sensitive attribute remained constant. Smoking during pregnancy can severely impair the health

of the unborn child (Cnattingius, 2004; Mei-Dan et al., 2015), but its prevalence in Germany is still alarmingly high and was around 10.9% in a direct survey of mothers of 0-to-6-year-old children born between 2007 and 2016 (Kuntz et al., 2018). Since the awareness of the negative consequences of smoking during pregnancy steadily increases, socially desirable response behavior in direct surveys is likely (Cnattingius, 2004); therefore, prevalence estimates provided by the ECWM might be higher than prevalence estimates based on responses to direct questioning.

Methods

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Participants

Respondents were recruited from a commercial German online panel provider. The inclusion criteria were that respondents were female adults who had been pregnant before but were not pregnant at the time of the survey. Moreover, respondents had to be native German speakers to ensure they could understand the instructions.

The initial sample in the first wave consisted of N = 4,002 respondents who met the inclusion criteria; of them, 3,585 completed the first wave. The initial sample in the second wave consisted of N = 2,777 respondents who met the inclusion criteria. Overall, N = 2,317 women completed the survey in both waves. We randomly assigned 778 respondents (33.58%) to the DQ condition, 750 respondents (32.37%) to the ECWM condition with randomization probability p_1 and 789 respondents (34.05%) to the ECWM condition with randomization probability p_2. The final sample's age and education distribution can be found in Table 1 and did not differ between conditions. All respondents participated voluntarily and provided informed consent.

A priori power considerations based on Ulrich et al. (2012) indicated that a four-digit sample size would ensure sufficient statistical power ($1 - \beta \geq .80$) for the planned prevalence comparisons between prevalence estimates; a precise calculation of the required sample size was not possible, as the exact parameter constellation remained unknown until data collection was completed. Twice as many respondents had to be allocated to the indirect questioning condition compared to the DQ condition to compensate for the generally lower efficiency of indirect questioning techniques (Moshagen et al., 2012; Ulrich et al., 2012).

Design and Procedure

In wave 1, respondents provided demographic data (gender, age, native language, and highest educational attainment). Participation was limited to adult women with German as their native language. Only females indicating that they had previously been pregnant but were not currently pregnant were allowed to proceed because if they were currently pregnant, their status regarding the sensitive attribute might have changed during the study interval. Respondents then received the sensitive statement ("I have smoked during at least one pregnancy") and an anchor question ("I am on the island 'La Reunion' right now") as proposed by Atsusaka and Stevenson (2023) in randomized order, both either in the DQ or in the ECWM format. In the DQ conditions, respondents were presented with the sensitive statement and had to indicate whether they agreed with the statement or not. In the ECWM conditions, respondents were first explained the ECWM in detail and then had to answer comprehension questions and were provided with feedback on their responses to the comprehension questions to ensure they had understood the instructions. Then, respondents in the ECWM conditions were simultaneously presented with a sensitive statement and a non-sensitive statement: "I was born in November or December" in the group with randomization probability p_1 or "I was born between January and October" in the group with randomization probability p_2 (p_1 = .158, p_2 = .842; Pötzsch, 2012). Respondents were asked to provide a joint answer by choosing one of the two answer options "I agree with *both* statements or I agree with *none* of the statements" versus "I agree with *exactly one* of the statements (irrespective of which one)." Since both the sensitive and the non-sensitive randomization attributes were temporally stable, prevalence estimates at the aggregate level and individual responses could meaningfully be compared over time. Subsequently, respondents were asked to evaluate their subjective perception of the sensitive question and provide information on their response behavior by indicating their agreement with nine statements on a 7-point Likert scale ranging from 1 = *strongly disagree* to 7 = *strongly agree* (Meisters et al., 2022b). The exact wording of these statements along with observed means and standard errors are provided via the OSF in Table S1 (https://osf.io/gfasy/). In the DQ condition, respondents were also asked about their and their mother's birth month. All respondents were informed that they would be invited for a second wave 1 week later.

Table 1. Demographics for the total sample and by questioning technique

	Total sample (%)	DQ (%)	ECWM m(%)
Age (years)			
18–25	3	4	3
26–35	31	29	32
36–45	40	43	39
46–55	22	20	22
56–65	4	4	4
> 65	0	0	0
χ^2-test for the difference of the distribution of age between DQ and ECWM condition: $\chi^2(5) = 6.83$, $p = .234$, Cramer's $V = .05$			
Educational achievement			
No school leaving certificate	0	0	0
Lower secondary school leaving certificate	9	9	9
Secondary school leaving certificate	42	43	41
Subject-specific university entrance qualification	10	11	10
Higher education entrance qualification	15	16	15
Bachelor's degree	8	7	9
Master's degree	14	13	14
PhD	1	1	1
χ^2-test for the difference of the distribution of educational achievement between DQ and ECWM condition: $\chi^2(7) = 5.57$, $p = .591$, Cramer's $V = .05$			

Note. DQ = Direct Questioning; ECWM = Extended Crosswise Model.

In wave 2, respondents were only allowed to proceed if they indicated that they had not become pregnant and did not have a positive pregnancy test since the first part of the survey. Thus, we ensured that the respondent's status regarding the sensitive attribute did not change during the study interval. Respondents were assigned to the same condition as in wave 1 and again received the same sensitive and anchor questions in randomized order. In the ECWM conditions, respondents again received a detailed explanation of the questioning technique and had to answer questions to ensure they understood the instructions and received feedback on their responses. Moreover, respondents were again asked to evaluate their subjective perception of the sensitive question and to provide information on their response behavior using the nine statements mentioned above. Finally, respondents were thanked and debriefed.

Statistical Analyses

For our analyses, we used R (version 4.0.5; R Core Team, 2021) and the packages *RRreg* for prevalence estimation and comparison (version 0.7.1; Heck & Moshagen, 2018), *tidyr* for pivoting data (version 1.2.0; Wickham & Girlich, 2022), *propint* for tests of proportions (version 0.2.14; Papenberg, 2018) and *cocor* for comparisons of correlations (version 1.1-4; Diedenhofen & Musch, 2015). To assess whether the distribution of the empirically observed answer frequencies differed between waves 1 and 2, we used a McNemar test. Parameter estimates π for each questioning technique (DQ, ECWMp_1, ECWMp_2) were calculated based on the empirically observed answer frequencies. The model fit of the ECWM was assessed via the asymptotically χ^2-distributed log-likelihood statistic G^2. To this end, we tested the fit of an overall ECWM model with one degree of freedom, in which the prevalence parameters of the two ECWM groups were equalized and combined into a single parameter ($\pi_{ECWM\_1} = \pi_{ECWM\_2} = \pi_{ECWM}$). To compare prevalence estimates between different questioning technique groups (e.g. π_{ECWM} vs. π_{DQ}), we assessed the difference in model fit (ΔG^2) between an unrestricted baseline model in which the respective parameters could be estimated freely and a restricted alternative model in which parameters were set equal (e.g., $\pi_{ECWM} = \pi_{DQ}$). A significant difference in model fit indicated when a restriction was invalid because the respective parameters significantly differed from each other (e.g., $\pi_{ECWM} \neq \pi_{DQ}$). The raw data and the code for analysis are available via the OSF (https://osf.io/gfasy/).

In an effort to increase the validity of the ECWM prevalence estimates by correcting for the potentially harmful influence of random responses, we intended to use a procedure proposed by Atsusaka and Stevenson (2023). To do this, respondents received the sensitive question and an anchor question with a known prevalence of zero in randomized order. Responses to the anchor question allow for estimating the proportion of participants who responded randomly; this proportion can then be used to make a post hoc correction to the prevalence estimates obtained using the ECWM.

Results

In the following, we report the results of prevalence estimation in both waves, retest stability, and additional analyses, by experimental condition.

Although we intended to correct the ECWM prevalence estimates for random responses using an anchor question (Atsusaka & Stevenson, 2023), a necessary condition for this correction was not met by participants' response behavior. The results for the anchor question showed a proportion of random responses close to zero in both wave 1 (3%) and wave 2 (4%), respectively. Apparently, almost none of the participants answered randomly, making a post hoc correction for random responses ineffective. Consequently, we report uncorrected ECWM prevalence estimates in the following.

Prevalence Estimates and Retest Stability

The ECWM fitted the observed data well in both wave 1, $\Delta G^2(1) = 1.01$, $p = .314$, and wave 2, $\Delta G^2(1) = 0.32$, $p = .570$, respectively. For wave 1, estimates for the prevalence of having smoked during at least one pregnancy were 19.54% ($SE = 1.42\%$) in the DQ condition and 23.26% ($SE = 1.73\%$) in the ECWM condition. For wave 2, prevalence estimates were 19.79% ($SE = 1.43\%$) in the DQ condition and 23.26% ($SE = 1.73\%$) in the ECWM condition. Neither the DQ estimates nor the ECWM estimates differed between waves 1 and 2, DQ: $\Delta G^2(1) = 0.02$, $p = .899$; ECWM: $\Delta G^2(1) < .001$, $p > .999$. The distribution of the empirically observed answer frequencies did not differ between waves 1 and 2 for either condition, DQ: $\chi^2(1) = .03$, $p = .874$, ECWM: $\chi^2(1) < .001$, $p > .999$. Moreover, for both waves, prevalence estimates in the DQ and ECWM condition did not significantly differ, wave 1: $\Delta G^2(1) = 2.74$, $p = .098$; wave 2: $\Delta G^2(1) = 2.37$, $p = .124$.

Overall, 88.6% of the respondents in the ECWM conditions answered consistently by choosing the same answer option at waves 1 and 2, whereas 94.9% answered consistently in the DQ condition. In both conditions, the number of inconsistent answers differed significantly from zero, DQ: $d = .05$, 95% CI [.04, .07], $p < .001$; ECWM: $d = .11$, 95% CI [.10, .13], $p < .001$. The number of inconsistencies was significantly larger in the ECWM than in the DQ condition, $\chi^2(1) = 23.48$, $p < .001$, $\varphi = .10$. The correlation of responses at waves 1 and 2 was .837 in the DQ and .771 in the ECWM condition. These correlations significantly differed from each other, Fisher's $z = 4.29$, $p < .001$ (Fisher, 1925); the correlation of responses at waves 1 and 2 was higher in the DQ than in the ECWM condition. To account for possible differences in base rates when assessing response consistency, we additionally calculated odds ratios for choosing one of the response options at wave 2 given that the same response option had been chosen at wave 1. We found an odds ratio of 201.7 (95% CI [108.4, 397.8]) for DQ and an odds ratio of 49.7 (95% CI [36.4, 68.8]) for the ECWM. The absolute answer frequencies for DQ and ECWM in both waves can be found in Table 2.

Additional Analyses

In both waves, respondents in the DQ condition showed significantly higher mean values on the scale of their subjective perception of the sensitive question than respondents in the ECWM condition, indicating a slightly more positive evaluation of the questioning technique, wave 1: DQ: $M = 5.93$, $SE = 0.03$, ECWM: $M = 5.78$, $SE = 0.03$, $t(2{,}315) = -3.66$, $p < .001$; wave 2: DQ: $M = 6.06$, $SE = 0.03$, ECWM: $M = 5.86$, $SE = 0.03$, $t(2{,}315) = -4.97$, $p < .001$.

To identify possible causes for inconsistent answer behavior, we analyzed the Spearman rank correlation of answer consistency (1 = consistent vs. 0 = inconsistent) with the demographic variables age group and education, as well as the point-biserial correlation of answer consistency with the subjective evaluation of the questioning technique obtained via the 9-item scale by Meisters et al. (2022b), separately for the DQ and the ECWM condition. We found that in the DQ condition, all associations were nonsignificant (age: $\rho = .02$, $p = .555$, education: $\rho = .05$, $p = .208$, subjective evaluation wave 1: $r = .03$, $p = .334$, subjective evaluation wave 2: $r = .07$, $p = .065$), whereas in the ECWM condition, all associations except for the association with age were significant (age: $\rho = -.01$, $p = .812$, education: $\rho = .09$, $p < .001$, subjective evaluation wave 1: $r = .16$, $p < .001$, subjective evaluation wave 2: $r = .15$, $p < .001$). In the ECWM, respondents who answered inconsistently were less educated and evaluated the questioning technique worse than respondents who answered consistently. Slower and therefore presumably more careful responders answered somewhat more consistently; however, this relationship was very small ($r = .11$), occurred only in the ECWM condition in wave 1, and explained less than 2% of the variance in response consistency overall (results are available via the OSF: https://osf.io/gfasy/).

Discussion

The present study reports the first assessment of the retest stability of the ECWM, as well as a comparison with the retest stability of DQ. Using a temporally stable attribute for the sensitive question and a temporally stable non-sensitive attribute for the randomization ensured that no change of these attributes during the study could influence the results.

Table 2. Absolute answer frequencies for DQ and ECWM in both waves

	Wave 2			
	DQ		ECWM	
Wave 1	"false"	"true"	"exactly one"	"both/none"
"false"/"exactly one"	605	21	400	88
"true"/"both/none"	19	133	88	963

Note. DQ = Direct Questioning; ECWM = Extended Crosswise Model.

Answer frequencies and, thus, prevalence estimates for waves 1 and 2 were almost identical for both the ECWM and DQ. At the individual level, response consistency was also high (ECWM: 89%, DQ: 95%), as were the correlation of answers in waves 1 and 2 (ECWM: .717, DQ: .837), and the odds ratios for choosing the same response options in wave 2 as in wave 1 (ECWM: 50, DQ: 202). Collectively, these results suggest that most participants chose the same response options across waves in both questioning technique conditions. The inconsistencies rate was somewhat higher in the ECWM than in the DQ condition, which aligns with previous findings indicating that indirect questioning techniques are more prone to erroneous responses than DQ (Hoffmann et al., 2017). Although some responses changed, this occurred in either direction for the ECWM and DQ; changes from option A to option B were as frequent as changes from option B to option A. Changes thus seemed nonsystematic, and prevalence estimates in waves 1 and 2 remained identical. In addition, the good model fit of the ECWM to the observed data in both waves rules out a systematic preference for one of the response options generated by a misunderstanding of the instructions as a cause of the high retest stability of the ECWM. Overall, the relatively high retest stability found in the present study suggests that response behavior in the ECWM is consistent over time; it is inconsistent with the assumption that respondents answered randomly. In addition, the analysis of an anchor question confirmed that almost none of the respondents answered randomly, strengthening the validity of the ECWM.

Inconsistent answer behavior was unrelated to demographic variables in the DQ condition. It was, however, associated with lower education and a more negative evaluation of the questioning technique in the ECWM condition. These findings are consistent with a previous study indicating that more errors are made in indirect questioning among lower-educated respondents and respondents who do not understand or trust the procedure (Meisters et al., 2020a).

As for the sensitive attribute of smoking during a previous pregnancy, the ECWM provided prevalence estimates that were somewhat, but not significantly higher than, DQ estimates. Approximately 20% of the mothers surveyed in the current study reported having smoked during at least one pregnancy. This rate is relatively high compared to a previous study reporting a prevalence of about 11% in a German sample (Kuntz et al., 2018). These results suggest that some respondents did not perceive this behavior as sensitive or felt their privacy was sufficiently protected even in the DQ condition. Considering the increasing awareness of the negative consequences and the resulting social undesirability of smoking during pregnancy (Cnattingius, 2004), it seems more likely that the respondents trusted the researcher's promise to keep all responses confidential. In the present study, as in previous studies using the same scale that includes an item on respondents' confidence in their privacy protection (Meisters et al., 2022b), the subjective evaluation of DQ by the respondents was found to be quite positive.

Limitations

Retest stability in the present study was not perfect for either questioning technique. The less-than-perfect reliability was, thus, not a flaw limited to indirect questioning but a problem of self-reports in general – including DQ. In the current study, even in the DQ condition, the rate of inconsistent answers was 5%, and in a previous study, the rate of incorrect answers in the DQ condition was about 10% (Hoffmann et al., 2017); thus, a non-zero proportion of participants always seems to respond in a nonsensical or inattentive way. In a study by Bishop et al. (1980), many respondents took a clear position on a purely fictitious issue; and in many studies, respondents failed the Instructional Manipulation Check proposed by Oppenheimer et al. (2009). These examples illustrate that self-reports can always be distorted by inattentive responses and should be interpreted cautiously. The ECWM also involves a more complex procedure than a DQ, which might cause difficulties in comprehension, especially for lower-educated participants (Hoffmann et al., 2017; Meisters et al., 2020a). More processing steps are required to answer an indirect question than a direct question, making indirect questioning techniques more susceptible to errors. As for the current study, we cannot tell the exact reason for the inconsistent responses. However, these inconsistent responses were rare and occurred non-systematically in either direction, leading to identical prevalence estimates in waves 1 and 2 for both

DQ and the ECWM and high retest stability for both questioning techniques.

The sample examined in the present study consisted exclusively of mothers. Future studies should investigate the stability of indirect questioning techniques in more heterogeneous populations. However, attention should always be paid to the selection of the sensitive attribute, as it is essential for the validity and conclusiveness of fundamental studies of retest stability that the attribute under investigation does not change between measurement times.

Although this does not affect the validity of the main finding regarding retest stability, our non-representative sample does not allow generalization of the results to the population of all mothers. It is worth noting, however, that biasing self-selection effects are rather unlikely, as the survey topic was not disclosed at the time of recruitment.

Finally, the relatively high attrition between waves 1 and 2 may have been partly systematic. For example, high conscientiousness may have facilitated participation in both waves as well as consistent responding, potentially inflating the estimated retest stability. Unfortunately, this hypothesis could not be tested because personality data were not collected.

Conclusion

In the current study, we conducted the first experimental assessment of the retest stability of an indirect questioning technique, the ECWM, and compared it to the retest stability of DQ. Using two measurement points, we found that prevalence estimates for a temporally stable attribute were almost identical, and most individual responses matched perfectly (ECWM: 89%, DQ: 95%). The current study, therefore, supports the notion that indirect questioning techniques may provide satisfying retest stabilities that are not substantially lower than for other assessment methods.

References

Atsusaka, Y., & Stevenson, R. T. (2023). A bias-corrected estimator for the crosswise model with inattentive respondents. *Political Analysis, 31*(1), 134–148. https://doi.org/10.1017/pan.2021.43

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1980). Experiments in filtering political opinions. *Political Behavior, 2*, 339–369. https://doi.org/10.1007/BF00990173

Cnattingius, S. (2004). The epidemiology of smoking during pregnancy: Smoking prevalence, maternal characteristics, and pregnancy outcomes. *Nicotine & Tobacco Research, 6*(Suppl), S125–S140. https://doi.org/10.1080/14622200410001669187

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One, 10*(6), Article e0131499. https://doi.org/10.1371/journal.pone.0121945

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23. https://doi.org/10.1037/a0039139

Fisher, R. A. (1925). The correlation coefficient. In R. A. Fisher (Ed.), *Statistical methods for research workers* (pp. 138–175). Oliver and Boyd. https://psychclassics.yorku.ca/Fisher/Methods/chap6.htm

Heck, D. W., & Moshagen, M. (2018). RRreg: An R package for correlation and regression analyses of randomized response data. *Journal of Statistical Software, 85*(2), 1–29. https://doi.org/10.18637/jss.v085.i02

Heck, D. W., Hoffmann, A., & Moshagen, M. (2018). Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior Research Methods, 50*(5), 1895–1905. https://doi.org/10.3758/s13428-017-0957-8

Hoffmann, A., Diedenhofen, B., Verschuere, B. J., & Musch, J. (2015). A strong validation of the crosswise model using experimentally induced cheating behavior. *Experimental Psychology, 62*(6), 403–414. https://doi.org/10.1027/1618-3169/a000304

Hoffmann, A., Meisters, J., & Musch, J. (2020). On the validity of non-randomized response techniques: An experimental comparison of the crosswise model and the triangular model. *Behavior Research Methods, 52*(4), 1768–1782. https://doi.org/10.3758/s13428-020-01349-9

Hoffmann, A., Meisters, J., & Musch, J. (2021). Nothing but the truth? Effects of faking on the validity of the crosswise model. *PLoS One, 16*(10), Article e0258603. https://doi.org/10.1371/journal.pone.0258603

Hoffmann, A., & Musch, J. (2019). Prejudice against women leaders: Insights from an indirect questioning approach. *Sex Roles, 80*(11–12), 681–692. https://doi.org/10.1007/s11199-018-0969-6

Hoffmann, A., Waubert de Puiseau, B., Schmidt, A. F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods, 49*(4), 1470–1483. https://doi.org/10.3758/s13428-016-0804-3

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/10.1017/pan.2016.5

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), Article e0201770. https://doi.org/10.1371/journal.pone.0201770

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model. *Public Opinion Quarterly, 76*(1), 32–49. https://doi.org/10.1093/Poq/Nfr036

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*, 18–32. https://doi.org/10.1016/j.joep.2014.08.001

Kuntz, B., Zeiher, J., Starker, A., Prütz, F., & Lampert, T. (2018). Rauchen in der Schwangerschaft–Querschnittergebnisse aus KiGGS Welle 2 und Trends [Smoking during pregnancy – Cross-sectional results from KiGGS wave 2 and trends]. *Journal of Health Monitoring, 3*(1), 47–54. https://doi.org/10.17886/RKI-GBE-2018-009

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS One, 16*(12), Article e0261673. https://doi.org/10.1371/journal.pone.0261673

Mei-Dan, E., Walfisch, A., Weisz, B., Hallak, M., Brown, R., & Shrim, A. (2015). The unborn smoker: Association between smoking

during pregnancy and adverse perinatal outcomes. *Journal of Perinatal Medicine, 43*(5), 553–558. https://doi.org/10.1515/jpm-2014-0299

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

Meisters, J., Hoffmann, A., & Musch, J. (2020b). Controlling social desirability bias: An experimental investigation of the extended crosswise model. *PLoS One, 15*(12), Article e0243384. https://doi.org/10.1371/journal.pone.0243384

Meisters, J., Hoffmann, A., & Musch, J. (2022a). More than random responding: Empirical evidence for the validity of the (extended) crosswise model. *Behavior Research Methods 55*, 716–729. https://doi.org/10.3758/s13428-022-01819-2

Meisters, J., Hoffmann, A., & Musch, J. (2022b). A new approach to detecting cheating in sensitive surveys: The cheating detection triangular model. *Sociological Methods & Research*. https://doi.org/10.1177/00491241211055764

Meisters, J., Hoffmann, A., & Musch, J. (2023). *Substance or noise? On the retest stability of responses to direct and indirect questions* [Data set]. https://doi.org/10.17605/OSF.IO/GFASY

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*, Article 12. https://doi.org/10.1186/s12889-020-10109-5

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44*(1), 222–231. https://doi.org/10.3758/s13428-011-0144-2

Nakhaee, M. R., Pakravan, F., & Nakhaee, N. (2013). Prevalence of use of anabolic steroids by bodybuilders using three methods in a city of Iran. *Addiction & Health, 5*(3–4), 77–82.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Papenberg, M. (2018). *propint: Testing for interactions in proportions*. Version 0.2.14. https://github.com/m-Py/propint

Pötzsch, O. (2012). *Geburten in Deutschland* [Births in Germany]. https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/Bevoelkerungsbewegung/BroschuereGeburtenDeutschland0120007129004.pdf

Pu, X., Gao, G., Fan, Y., & Wang, M. (2016). Parameter estimation in stratified cluster sampling under randomized response models for Sensitive Question Survey. *PLOS One, 11*(2), Article e0148267. https://doi.org/10.1371/journal.pone.0148267

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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*, Article 655592. https://doi.org/10.3389/fpsyg.2021.655592

Schnapp, P. (2019). Sensitive question techniques and careless responding: Adjusting the crosswise model for random answers. *Methods, Data, Analyses, 13*(2), 307–320. https://doi.org/10.12758/mda.2019.03

Schnell, R., & Thomas, K. (2023). A meta-analysis of studies on the performance of the crosswise model. *Sociological Methods & Research, 52*(3), 1493–1518. https://doi.org/10.1177/0049124121995520

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. *Judgment and Decision Making, 11*(5), 527–536. https://doi.org/10.1017/S1930297500004605

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods, 17*(4), 623–641. https://doi.org/10.1037/A0029314

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*. 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.

Wickham, H., & Girlich, M. (2022). *tidyr: Tidy Messy Data*. Version 1.2.0. https://CRAN.R-project.org/package=tidyr

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

History

Received February 24, 2023
Revision received September 22, 2023
Accepted October 22, 2023
Published online April 2, 2024
EJPA Section / Category Methodological topics in assessment

Publication Ethics

The survey was carried out in accordance with the revised Declaration of Helsinki (World Medical Association, 2013) and the ethical guidelines of the German Society for Psychology (Berufsverband Deutscher Psychologinnen und Psychologen & Deutsche Gesellschaft fur Psychologie, 2016). In Germany, there is no binding obligation that research projects can only be carried out after approval by an ethics committee. Participation in the present study could not have any negative consequences for the respondents, and anonymity was ensured at all times. The respondents participated voluntarily and after informed consent was obtained. There was no risk that participation could cause any physical or mental damage or discomfort to participants beyond their normal everyday experiences. Therefore, ethics committee approval was not required according to the "Ethical Research Principles and Test Methods in the Social and Economic Sciences" formulated by the Ethics Research Working Group of the German Data Forum (RatSWD, 2017) and the "Ethical Recommendations of the German Psychological Society" (DGPs, 2018).

Authorship

Julia Meisters, Conceptualization, Formal analysis, Writing – Original Draft, Writing – Review & Editing; Adrian Hoffmann, Conceptualization, Writing – Review & Editing, Funding Acquisition; Jochen Musch, Conceptualization, Writing – Review & Editing, Funding Acquisition. All authors approved the final version of the article.

Open Science

Open Data: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook if relevant (Meisters et al., 2023).
Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (Meisters et al., 2023).
Open Analytic Code: We confirm that all the scripts, code and outputs needed to reproduce the results are provided (Meisters et al., 2023).

Preregistration of Studies and Analysis Plans: This study was not preregistered.
Data and materials for the current study are available via the OSF: https://osf.io/gfasy/.

Funding
This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Grant number 439602023. Open access publication enabled by Heinrich Heine University Duesseldorf.

Julia Meisters
Department of Experimental Psychology
Heinrich Heine University Duesseldorf
Universitätsstrasse 1
40225 Duesseldorf
Germany
julia.meisters@uni-duesseldorf.def