Handle with Care: Implementation of the List Experiment and Crosswise Model in a Large-Scale Survey on Academic Misconduct *

Julia Jerke[1], David Johann[2], Heiko Rauhut[1], Kathrin Thomas[3], and Antonia Velicu[†1]

[1]*University of Zürich*

[2]*ETH Zürich*

[3]*University of Aberdeen*

November 11, 2020

Abstract

This research analyses the effectiveness of the List Experiment and Crosswise Model in measuring self-plagiarism and data-manipulation. Both methods were implemented in a large-scale survey of academics on social norms and academic misconduct. As the results lend little confidence about the effectiveness of the methods. Researchers are best advised to avoid them or, at least, to handle them with care.

[*] This project is part of the SNSF Starting Grant BSS-GIO 155981 of Heiko Rauhut ("Social norms, cooperation and conflict in scientific collaborations"). We thank the anonymous reviewers as well as Alexander Ehlert, Isabel Raabe, and Justus Rathmann for their concise comments and constructive feedback on our work.

[†] Co-authors in alphabetical order. Study Design: Julia Jerke, David Johann, Heiko Rauhut, Kathrin Thomas, Antonia Velicu. Coding and Analysis: Julia Jerke, David Johann, Kathrin Thomas, Antonia Velicu. First draft: Julia Jerke, Heiko Rauhut, Kathrin Thomas, Antonia Velicu. Revisions: David Johann, Kathrin Thomas, Antonia Velicu. Final approval of the paper: Julia Jerke, David Johann, Heiko Rauhut, Kathrin Thomas, Antonia Velicu. jerke@soziologie.uzh.ch, david.johann@library.ethz.ch, rauhut@soziologie.uzh.ch; corresponding authors: kathrin.thomas@abdn.ac.uk, velicu@soziologie.uzh.ch

Keywords: Indirect Question Format, Item Count Technique, Sensitive Characteristics, Science Research, Survey Experiment

1 Introduction

Eliciting accurate prevalence estimates of sensitive characteristics, such as drug usage, antisocial behaviour, or misconduct in a survey environment is prone to misreporting (Tourangeau et al., 2000; Tourangeau and Yan, 2007; Krumpal, 2013). Survey methodologists thus seek to develop question designs that allow capturing sensitive behaviour indirectly to circumvent issues related social desirability pressures. Two popular methods are the List Experiment (Miller, 1984; Droitcour et al., 1991, LE) and the Crosswise Model (Yu et al., 2008, CM).

Even though both formats are frequently used to measure sensitive characteristics (Droitcour et al., 1991; Gilens et al., 1998; Tsuchiya et al., 2007; Holbrook and Krosnick, 2010; Jann et al., 2012; Korndörfer et al., 2014; Roberts and John, 2014; Wolter and Laier, 2014; Hoffmann and Musch, 2016; Höglinger et al., 2016; Thomas et al., 2016; Johann and Thomas, 2017; Höglinger and Jann, 2018), an increasing literature raises concerns about the effectiveness of these methods (Droitcour et al., 1991; Landsheer et al., 1999; Coutts and Jann, 2011; Wolter, 2012; Wolter and Laier, 2014; Aronow et al., 2015; Kiewiet de Jonge, 2015; Coffman et al., 2017; Höglinger and Diekmann, 2017; Hoffmann et al., 2017; Jerke et al., 2019; Schnell and Thomas, 2020). The methods "impose a higher cognitive burden on respondents" (Jerke et al., 2019, 320); are prone to produce false negative and positive results (Wolter and Laier, 2014; Höglinger and Jann, 2018); and their effectiveness may depend on the sample quality (Schnell and Thomas, 2020).

2 On the Logic of the List Experiment and Crosswise Model

The traditional LE is based on a split sample design, where the control group receives a short list of unobtrusive items; the treatment group the same list plus a sensitive item. Participants are asked to indicate *how many* of the items apply to them without disclosing which ones. The mean difference of the long and short list provides the prevalence estimate of the sensitive characteristic. Variations of the traditional LE exist (Blair and Imai, 2012; Glynn, 2013; Aronow et al., 2015; Li and Van den Noortgate, 2019), but have only been applied infrequently.

The CM features an unobtrusive and a sensitive question. As opposed to providing separate answers to each question, respondents are asked to give a joint answer to both. Either the response to the questions is the same or it is different. The prevalence of the sensitive characteristic can be estimated, if the researcher knows the probabilities of the unobtrusive question and both questions have binary response codes (Krumpal et al., 2015; Jerke et al., 2019). Variations of the CM, including changes to the selection of the unobtrusive question, have been proposed but not tested yet (Diekmann, 2012; Höglinger, 2016).

If validation with a true value is impossible (Landsheer et al., 1999), a Direct Question (DQ) can be asked in a separate split sample to test whether the LE or CM produce better estimates in comparison to the DQ. A better estimate according to the commonly applied "more-is-better"-assumption is a higher prevalence estimate for socially undesirable behaviour and, in reverse, a lower estimate for socially desirable characteristics (Umesh and Peterson, 1991). Yet, recent research indicated that this assumption might be undermined by respondents deliberately lying or cheating, e.g., by disregarding the rules (Höglinger et al., 2016; Höglinger and Jann, 2018; Jerke et al., 2019; Walzenbach and Hinz, 2019).

3 Method

The Zürich Survey of Academics (Rauhut et al., 2020) enquired about recent developments in academia in Austria, Germany, and Switzerland. The project conducted a census of academics in Austria and Switzerland and drew a 50%-probability sample of scholars in Germany. In total, 15,972 academics at 236 universities were interviewed (Austria: n=2,832; Germany: n=8,228; Switzerland: n=4,912). The overall response rate was 11.33% (Austria=10.08%; Germany=10.44%; Switzerland=14.44%).

The LE and CM were designed to measure socially undesirable academic misconducts: (1) Submitting the same results without indicating it (self-plagiarism) and (2) intentionally altering the data to confirm the research question (data-manipulation) (Fanelli, 2009). The exact question wording is presented in the Appendix A.

Both items are considered as sensitive: Roughly 95% indicated that they felt uncomfortable admitting to data-manipulation (Austria: 94.88%; Germany: 95.10%; Switzerland: 93.74%); two thirds said the same about self-plagiarism (Austria: 67.38%; Germany: 65.33%; Switzerland: 67.98%).

The wording and all items were pretested in expert discussions and cognitive interviews, indicating that commonly used unobtrusive questions with known probabilities, such as asking about someone's birthday, raised mistrust among respondents (Rauhut et al., 2020; Jerke et al., 2019). To circumvent this, a context-related statement question was asked in a separate split sample and serves an estimator for the unobtrusive question. This is a novel approach. To avoid the loss of privacy protection in the LE due to floor and ceiling effects, the unobtrusive items vary in their prevalence and some of the items correlate negatively (see Appendix B) (Glynn, 2013; Jerke et al., 2019).

Both designs were integrated towards the end of the overall questionnaire to avoid early breakoffs. The Computer Assisted Web Interviews were programmed to randomly assign respondents to treatment and control groups: Group 1 received both CM questions, Group 2 the LE's long lists, Group 3 the LE's short lists, Group 4 the DQs.

We calculated the prevalence estimate $\hat{\pi}$ along with the respective standard errors $SE_{\hat{\pi}}$ for both items and methods on the full sample. The underlying equations can be found in Appendix C. We excluded respondents indicating that they had never published, as academics without publication experience are unlikely to self-plagiarise. We also drop those saying that they have no experience with statistical data analysis, because as they are unable to manipulate quantitative data.

For robustness, we (1) re-ran the analysis on the full sample, (2) conducted cross-country analysis to uncover potential variation, and (3) re-estimated the results for self-plagiarism by respondents' perceived level of sensitivity, comparing those feeling highly uncomfortable admitting to the misconduct with those feeling less uncomfortable. This last check was only performed for self-plagiarism, as almost all respondents rated data-manipulation as highly sensitive, generating too little item variance. The rationale for this check is that the perceived item sensitivity may impact the effectiveness of the method when high-frequency behaviour is concerned (Wolter and Laier, 2014).

4 Results

Table 1 displays the prevalence estimates for self-plagiarism and data-manipulation in the DQ, LE, and CM condition along with their standard errors. We also calculated the difference between each indirect and direct question $\Delta_{LE/CM-DQ}$, its standard error (SE_{Δ}) and the 95%-confidence intervals. The results indicate that the LE fails to produce a higher prevalence estimate for both items in comparison with the DQ, failing the "more-is-better"-assumption. It even produced a negative prevalence estimate for self-plagiarism. The results for the CM are mixed: While it resulted a negative prevalence estimate for self-plagiarism, it generated a higher estimate than the DQ for data-manipulation, confirming the "more-is-better"-assumption.

	π	SE$_{\pi}$	Δ	SE$_{\Delta}$	95% CI	N
DQ$_{SP}$	2.99	0.31	-	-	[2.38; 3.60]	3,012
LE$_{SP}$	2.15	1.68	-0.84	1.71	[2.50; 4.18]	6,180
CM$_{SP}$	-7.06	2.51	-10.05	2.53	[-15.01; -5.08;]	3,008
DQ$_{DM}$	1.53	0.29	-	-	[0.96; 2.10]	1,765
LE$_{DM}$	-3.29	1.98	-4.82	2.00	[-8.73; -0.89]	3,654
CM$_{DM}$	9.02	2.46	7.49	2.48	[2.63; 12.35]	1,792

Table 1: Estimates of Self-Plagiarism (SP) and Data-Manipulation (DM)

The robustness checks confirm our results: The analyses reveal similar patterns for the full sample, across countries – with the exception of Austria, where the LE arguably works better following the "more-is-better"-assumption – and when splitting the sample by respondents' perception of sensitivity.[1]

[1] Detailed results of all robustness checks can be provided upon request.

5 Discussion and Conclusion

The LE and CM seem popular methods to help circumvent issues of misreporting and are believed to have the potential to better estimate sensitive characteristics. Yet, concerns have been raised indicating that the methods are mostly successful when low quality samples are used (Schnell and Thomas, 2020), that they are cognitively too challenging (Jerke et al., 2019), and produce substantive numbers of false positives and negatives (Höglinger and Jann, 2018). Effective implementations may also be tied to the level of sensitivity (Wolter and Laier, 2014; Thomas et al., 2016).

Our results show that both methods fail to generate higher prevalence estimates for self-plagiarism. While the LE does also not work for data-manipulation, the results indicate that the CM potentially performs well for data-manipulation. However, we are wary that this positive result may be biased by the occurrence of false positives (Höglinger et al., 2016; Höglinger and Jann, 2018).

Our results support a growing body of literature raising concerns about the effectiveness of the LE and CM. The occurrence of negative prevalence estimates in the LE conditions might be a relic of the design or related to different sample populations (Tsuchiya et al., 2007). However, as we employ a highly educated sample of academics, we may conclude that the effectiveness of LE is not an issue of respondents' cognitive skills. The same argument applies to the CM. (Jerke et al., 2019). We would carefully note that asking about academic misconduct in general may be too sensitive, as the associated sanctions can be severe. Any survey question format might fail to predict too sensitive characteristics. This article challenges Roberts and John (2014) who emphasize that future studies on scientific misconduct should consider using specialised questioning techniques, such as LE and CM. In line with prior research (e.g., Glynn, 2013; Gelman, 2014; Hinsley et al., 2019; Jerke et al., 2019; Schnell and Thomas, 2020), we conclude that researchers are best advised to avoid these methods or, at least, to handle them with care.

References

Aronow, P. M., Coppock, A., Crawford, F. W., and Green, D. P. (2015). Combining list experiment and direct question estimates of sensitive behavior prevalence. *Journal of Survey Statistics and Methodology*, 3(1):43–66.

Blair, G. and Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis*, 20(1):47–77.

Coffman, K. B., Coffman, L. C., and Ericson, K. M. M. (2017). The size of the lgbt population and the magnitude of antigay sentiment are substantially underestimated. *Management Science*, 63(10):3168–3186.

Coutts, E. and Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). *Sociological Methods & Research*, 40(1):169–193.

Diekmann, A. (2012). Making use of Benford's Law for the Randomized Response Technique. *Sociological Methods & Research*, 41(2):325–334.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., and Ezzati, T. M. (1991). The Item Count Technique as a method of indirect questioning: A review of its development and a case study application. In Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S., editors, *Measurement Errors in Surveys*, pages 185–210. New York, NY: John Wiley & Sons.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*, 4(5):e5738.

Gelman, A. (2014). Thinking of doing a list experiment? Here's a list of reasons why you should think again. *Statistical Modeling, Causal Inference, and Social Science*, pages `https://statmodeling.stat.columbia.edu/2014/04/23/thinking--list--experiment--heres--list--reasons--think/` (last accessed: 13 September 2020).

Gilens, M., Sniderman, P. M., and Kuklinski, J. H. (1998). Affirmative action and the politics of realignment. *British Journal of Political Science*, 28(1):159–183.

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly*, 77(S1):159–172.

Hinsley, A., Keane, A., St. John, F. A., Ibbett, H., and Nuno, A. (2019). Asking sensitive questions using the unmatched count technique: Applications and guidelines for conservation. *Methods in Ecology and Evolution*, 10(3):308–319.

Hoffmann, A., de Puiseau, B. W., Schmidt, A. F., and Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods*, 49(4):1470–1483.

Hoffmann, A. and Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods*, 48(3):1032–1046.

Höglinger, M. (2016). *Revealing the truth? Validating the Randomized Response Technique for surveying sensitive topics*. PhD thesis, ETH Zurich.

Höglinger, M. and Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the Crosswise-Model RRT. *Political Analysis*, 25(1):131–137.

Höglinger, M. and 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.

Höglinger, M., Jann, B., and 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.

Holbrook, A. L. and Krosnick, J. A. (2010). Measuring voter turnout by using the Randomized

Response Technique: Evidence calling into question the methods validity. *Public Opinion Quarterly*, 74(2):328–343.

Jann, B., Jerke, J., and Krumpal, I. (2012). Asking sensitive questions using the Crosswise Model: An experimental survey measuring plagiarism. *Public Opinion Quarterly*, 76(1):32–49.

Jerke, J., Johann, D., Rauhut, H., and 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.

Johann, D. and Thomas, K. (2017). Testing the properties of the Crosswise Model in reducing social desirability bias on attitudes towards muslims: A validation study. *Survey Methods: Insights from the Field*. Retrieved from: https://surveyinsights.org/?p=8887, Last Accessed 28 June 2020.

Kiewiet de Jonge, C. P. (2015). Who lies about electoral gifts? experimental evidence from latin america. *Public Opinion Quarterly*, 79(3):710–739.

Korndörfer, M., Krumpal, I., and Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the Crosswise Model. *Journal of Economic Psychology*, 45:18–32.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47(4):2025–2047.

Krumpal, I., Jann, B., Auspurg, K., and von Hermanni, H. (2015). Asking sensitive questions: A critical account of the Randomized Response Technique and related methods. In Engel, U., Jann, B., Lynn, P., Scherpenzeel, A., and Sturgis, P., editors, *Improving Survey Methods: Lessons from Recent Research*, pages 122–136. New York, NY: Routledge.

Landsheer, J. A., Van Der Heijden, P., and Van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, 33(1):1–12.

Li, J. and Van den Noortgate, W. (2019). A meta-analysis of the relative effectiveness of the item count technique compared to direct questioning. *Sociological Methods & Research*, Online First:1–40.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. Unpublished Ph.D Thesis. George Washington University, Department of Sociology, Washington, DC.

Rauhut, H., Johann, D., Jerke, J., Rathmann, J., and Velicu, A. (2020). The Zurich Survey of Academics: Methods, design, and data. Zürich: University of Zürich.

Roberts, D. L. and John, F. A. S. (2014). Estimating the prevalence of researcher misconduct: a study of uk academics within biological sciences. *PeerJ*, 2:e562.

Schnell, R. and Thomas, K. (2020). A meta-analysis of studies on the performance of the Crosswise Model. *Under Review*.

Thomas, K., Johann, D., Kritzinger, S., Plescia, C., and Zeglovits, E. (2016). Estimating sensitive behavior: The ICT and high-incidence electoral behavior. *International Journal of Public Opinion Research*, 29(1):157–171.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The psychology of the survey response*. Cambridge: Cambridge University Press.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5):859—883.

Tsuchiya, T., Hirai, Y., and Ono, S. (2007). A study of the properties of the Item Count Technique. *Public Opinion Quarterly*, 71(2):253–272.

Umesh, U. N. and Peterson, R. A. (1991). A critical evaluation of the Randomized Response method applications, validation, and research agenda. *Sociological Methods & Research*, 20(1):104–138.

Walzenbach, S. and Hinz, T. (2019). Pouring water into wine: Revisiting the advantages of the crosswise model for asking sensitive questions. *Survey Methods: Insights from the Field*, pages 1–16.

Wolter, F. (2012). *Heikle Fragen in Interviews: Eine theoretische und empirische Validierung der Randomized Response-Technik*. Wiesbaden: VS Verlag.

Wolter, F. and Laier, B. (2014). The effectiveness of the Item Count Technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency. *Survey Research Methods*, 8(3):153–168.

Yu, J.-W., Tian, G.-L., and Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, 67(3):251–263.

Appendix A: Question Wording

Direct Question

V83. [IF Vsplit <= 32] Now we're interested in your experiences of certain behaviours. Please state which of the following statements apply to you and which do not.

1. I have submitted the same results to two or more journals without indicating this
2. I have intentionally manipulated empirical data to confirm my research question

□ Does not apply

□ Applies

□ no answer

List Experiment

Self-Plagiarism (Sensitive Item here in Italics, Long List Only)

Q86. [IF Vsplit = 65-96] You'll now be shown a list. This contains statements which apply to some academics, but not to others. Please indicate how many of these statements apply to you. Please do not say which statements apply to you, only how many.

- In the last 12 months I've worked on at least one research proposal [70.74%]
- *I have submitted the same results to two or more journals without indicating this*
- I share my office with at least one other person [60.49%]
- I speak at least three foreign languages fluently [25.99%]
- In the last semester I gave more than two lectures [41.00%]

Your privacy is protected, since we do not know your answers to the individual questions. Please note how many of the above-mentioned statements apply to you.

Number:

Data-Manipulation (Sensitive Item here in Italics, Long List Only)

V87. [IF Vsplit = 65-96] Once more, you'll be shown a list. Again, this contains statements that apply to some academics, but not to others. Please indicate how many of these statements apply to you. Please do not say which statements apply to you, only how many.

- In the last 12 months I've submitted at least one manuscript to a journal [84.61%]
- *I have intentionally manipulated empirical data to confirm my research question*
- In a typical working week I eat lunch with colleagues [69.65%]

- I have purchased subscriptions to the print version of at least two academic journals [16.28%]
- In the last 12 months I've attended more than four conferences [33.25%]

Your privacy is protected, since we do not know your answers to the individual questions. Please note how many of the above-mentioned statements apply to you.

Number:

Crosswise Model

Self-Plagiarism

V84. [IF Split = 33-64] We will now show you two statements that apply to some academics, but not to others. First, please consider whether the two statements apply to you or not, but do not write this down. Then please select the answer option (A) or (B), using the following rule:

If both statements apply to you or both statements do not apply to you, please select (A).

If one statement applies to you but the other does not, please select (B).

- Statement 1: In the last 12 months I have attended more than four conferences
- Statement 2: I have submitted the same results to two or more journals without indicating this

Your privacy is protected, since we do not know your answers to the individual questions. What is your answer?

□ (A) Both statements apply to me, or neither of the statements applies to me

□ (B) One of the statements applies to me, the other does not

Data-Manipulation

V85. [IF Split = 33-64] We will now show you two statements that apply to some academics, but not to others. First, please consider whether the two statements apply to you or not, but do not write this down. Then please select the answer option (A) or (B), using the following rule:

If both statements apply to you or both statements do not apply to you, please select (A).

If one statement applies to you but the other does not, please select (B).

- Statement 1: In the last 12 months I've worked on at least one research proposal
- Statement 2: I have intentionally manipulated empirical data to confirm my research question

Your privacy is protected, since we do not know your answers to the individual questions. What is your answer?

□ (A) Both statements apply to me, or neither of the statements applies to me

□ (B) One of the statements applies to me, the other does not

Appendix B: Correlations Unobtrusive Items LE

Table B1 Correlations Unobtrusive Items LE Self-Plagiarism (Q86)

	Item 2	**Item 3**	**Item 4**
Item 1	-0.1102	0.0298	0.1603
Item 2		-0.0248	-0.1870
Item 3			0.0440

Table B2 Correlations Unobtrusive Items LE Data-Manipulation (Q87)

	Item 2	**Item 3**	**Item 4**
Item 1	0.0347	-0.0119	0.0886
Item 2		-0.1280	-0.0834
Item 3			0.1646

Appendix C: Prevalence Estimates and Variances

List Experiment

The prevalence π of the sensitive item is estimated by

$$\hat{\pi}_{LE} = \bar{X}_{\text{long list}} - \bar{X}_{\text{short list}} \tag{1}$$

The sampling variance $Var(\hat{\pi}_{LE})$ is estimated by

$$Var(\hat{\pi}_{LE}) = Var(\bar{X}_{\text{long list}}) + Var(\bar{X}_{\text{short list}}). \tag{2}$$

The standard error $SE_(\hat{\pi}_{LE})$ is estimated by

$$SE(\hat{\pi}_{LE}) = \frac{\sqrt{Var(\hat{\pi}_{LE})}}{\sqrt{N_{\text{long list}} + N_{\text{short list}}}} \tag{3}$$

Crosswise Model

Prevalence π of the sensitive item is estimated by

$$\hat{\pi}_{CM} = \frac{\hat{\lambda} + p - 1}{2p - 1}, p \neq 0.5 \ , \tag{4}$$

where p is the known population prevalence of the non-sensitive item (Yu et al., 2008) and $\hat{\lambda}$ is the proportion of respondents stating that their response to the unobtrusive and the sensitive question is the same (Answer A).

The sampling variance $Var(\hat{\pi}_{CM})$ is estimated by

$$Var(\hat{\pi}_{CM}) = \frac{\hat{\lambda}(1-\hat{\lambda})}{n(2p-1)^2} = \frac{\hat{\pi}_{CM}(1-\hat{\pi}_{CM})}{n} + \frac{p(1-p)}{n(2p-1)^2}, p \neq 0.5 \quad . \tag{5}$$

The standard error $SE_(\hat{\pi}_{CM})$ is estimated by

$$SE(\hat{\pi}_{CM}) = \frac{\sqrt{Var(\hat{\pi}_{CM})}}{\sqrt{N_{CM}}} \tag{6}$$