

Hidden Practices Remain Hidden? Lessons from a Failed Attempt to Measure Purchasing Sexual Services Using the Item Count Technique

Carmen M. Leon¹ · Eva Aizpurua²

¹School of Law, University of Castilla-La Mancha

²Centre for Social Survey Transformation, National Centre for Social Research

The Item Count Technique (ICT) is an experimental method aimed at eliciting truthful responses to sensitive questions. In its basic form, the ICT randomly assigns respondents to one of two conditions: a control group, which receives a list of neutral statements, and a treatment group, which receives the same list with an additional sensitive item. Respondents then indicate the total number of statements that apply to them without specifying which ones. The difference in the average counts between the two groups is used to estimate the prevalence of the sensitive behavior. This study examines discrepancies in reporting the purchase of sexual services by comparing direct questioning with the ICT using an online survey of 1603 panelists in Spain. The findings reveal that the ICT was ineffective in estimating the prevalence of purchasing sexual services, underscoring the need to thoroughly test proposed designs and to further develop strategies for obtaining accurate self-reported data on sensitive topics.

Keywords: list experiment; sensitive data; social desirability; purchasing sexual services; measurement error; online survey

1 Introduction

Sensitive questions often concern taboo topics such as sexual activities, illegal behaviors, or socially undesirable attitudes. Responses to such questions are particularly prone to measurement error arising from social desirability bias, whereby respondents underreport undesirable behaviors or overreport socially approved ones (Tourangeau & Yan, 2007). Because surveys are often the only feasible means of studying hidden behaviors, reducing this source

of error is essential for obtaining valid estimates of sensitive behaviors (Yan, 2021). This challenge has motivated the development of indirect questioning techniques designed to reduce socially desirable responses.

1.1 The Item Count Technique

The Item Count Technique (ICT), also known as the *list experiment* or *unmatched count technique*, is a widely used indirect questioning method designed to reduce social desirability bias when measuring sensitive behaviors and attitudes. Unlike direct questioning (DQ), respondents report only the total number of applicable statements rather than indicating which specific items apply to them. Respondents in the treatment group receive an additional sensitive item alongside a list of non-sensitive statements, whereas those in the control group receive only the non-sensitive items. Comparing the average number of endorsed items across groups allows researchers to estimate the prevalence of the sensitive behavior without requiring respondents to disclose it directly (Miller, 1984).

Supplementary Information The online version of this article (<https://doi.org/10.18148/srm/2026.v20i3.8532>) contains supplementary material.

Corresponding author: Carmen M. Leon, School of Law, University of Castilla-La Mancha, 1 Plaza de la Universidad, 02071 Albacete, Spain (Email: carmen.leon@uclm.es)

Corresponding author: Eva Aizpurua, Centre for Social Survey Transformation, National Centre for Social Research, 35 Northampton Square, London, EC1V 0AX, United Kingdom (Email: eva.aizpu@gmail.com)

Although numerous studies have evaluated the ICT, its performance has proven highly context-dependent. Blair et al. (2020) conducted a meta-analysis of ICT studies up to 2017, concluding that ICT generally outperforms DQ, particularly in mitigating overreporting of socially desirable attitudes and behaviors. Similarly, Ehler et al. (2021) reviewed 89 articles encompassing 124 distinct samples, finding a significant positive pooled effect of ICT on the validity of survey responses compared to DQ. However, they also reported substantial heterogeneity, suggesting that ICT effectiveness varies according to research design, item characteristics, and implementation procedures. Li and Van den Noortgate (2022), who synthesized 246 effect sizes from 54 studies, reached similar conclusions while identifying additional moderators, including cultural orientation (i.e., collectivism versus individualism), topic valence, and the number of control items. Importantly, this heterogeneity suggests that understanding when and why the ICT fails may be as informative as documenting successful applications.

Several factors have been proposed to explain this heterogeneity. One important source of variation is the context in which the ICT is implemented. For example, Ehler et al. (2021) found that the technique tends to perform more effectively in non-Western societies, suggesting that its effectiveness may vary across cultural contexts and populations. This finding highlights the need to evaluate the ICT across diverse settings rather than assuming that its performance generalizes across groups.

Ceiling and floor effects represent another important source of bias in ICT applications. Ceiling effects may compromise respondent anonymity when participants endorse all items on the list, whereas floor effects reduce response variability and make it more difficult to estimate the prevalence of the sensitive behavior accurately (Groenitz, 2014). When a substantial proportion of respondents cluster at either extreme, differences between the control and treatment groups become harder to detect, reducing statistical precision and increasing the risk of biased prevalence estimates.

Finally, the cognitive demands of the ICT may also compromise data quality. Because respondents must process multiple statements while simultaneously understanding the indirect response format, the technique can be cognitively demanding, particularly in self-administered surveys (Kramon & Weghorst, 2019). Confusion may arise not only from the unfamiliar response format itself but also from respondents' perception that the survey is attempting to disguise its true purpose, potentially increasing mistrust and response error (Kuha & Jackson, 2014).

1.2 The Item Count Technique in Measuring Sensitive Sexual Behaviors

Empirical applications of the ICT to sensitive sexual behaviors remain relatively limited and have produced mixed findings. Starosta and Earleywine (2014) used the ICT to estimate the prevalence of risky sexual behaviors, concluding that the technique could reduce social desirability bias compared with traditional DQ, although effect sizes varied across behaviors. Roe-Sepowitz et al. (2019), in contrast, combined a double-list experiment with DQ to estimate the prevalence of purchasing sexual services among adult men in the United States and found that the ICT yielded a lower prevalence estimate ($\approx 2\%$) than the direct question ($\approx 4\%$). More recently, Jonsson (2023) incorporated a list experiment into a cross-national survey conducted in eight European countries, including Spain, to examine the purchase of sexual services. Although the increase in reporting obtained through the ICT was not statistically significant, indirect estimates were consistently higher than those obtained through DQ, and the substantive conclusions remained robust across both measurement approaches. Collectively, these studies illustrate that ICT performance is highly context-dependent, underscoring the importance of evaluating the assumptions underlying the technique in each application.

This variability is further reflected in reports of implausible or even negative prevalence estimates in some empirical applications (Rinken et al., 2021, 2024), highlighting important limitations in the technique's performance. At the same time, the published evidence may itself provide an incomplete picture of ICT effectiveness. Studies reporting successful applications of the ICT may be more likely to be published than those yielding null, implausible, or otherwise unsuccessful results. Consequently, publication bias may lead to an overly optimistic assessment of the robustness of the technique. Documenting unsuccessful implementations is therefore essential for identifying the conditions under which the ICT performs poorly, refining its design, and developing a more balanced evidence base on indirect questioning methods.

2 Current Study

This study evaluates the performance of the ICT in measuring the purchase of sexual services in Spain by comparing prevalence estimates obtained through DQ and the ICT. Consistent with the theoretical rationale underlying indirect questioning techniques, we hypothesized that the prevalence estimated through the ICT would exceed that obtained through DQ because its indirect response format

reduces the social desirability concerns that typically lead to underreporting of stigmatized behaviors.

Beyond comparing prevalence estimates, this study seeks to identify the methodological factors that may influence ICT performance when measuring a low-prevalence, highly stigmatized behavior. By examining an application in which the technique did not perform as expected, we aim to contribute to a better understanding of the conditions under which the ICT may fail and the methodological lessons that can be drawn from such unsuccessful implementations. In addition, the study extends the evidence on ICT performance to a Spanish-language survey context. Previous research suggests that language can influence how respondents interpret and answer survey questions, with implications for data quality (Sha & Gabel, 2020). Because most ICT applications have been conducted in English-speaking settings, evaluating the technique in a Spanish-language survey broadens the evidence base on its performance across different linguistic contexts.

3 Method

3.1 Data and Procedure

The online survey was administered between February 21 and 24, 2022, using the *Netquest* opt-in panel in Spain. Quotas were applied for gender, age, and habitat size to obtain a sample distribution similar to the Spanish population (see Appendix A). Among the panelists invited to participate, 62% completed the questionnaire and received points redeemable for gifts as a token of appreciation for their time. The achieved sample comprised 1603 residents in Spain (52% women), with ages ranging from 18 to 88 ($M = 45.7$, $SD = 15.6$).

The survey, conducted in Spanish, included up to 93 items measuring personal attitudes toward prostitution and prostitutes, as well as questions about the purchase of sexual services. The DQ and the ICT were administered toward the end of the questionnaire, after attitudinal measures and before the sociodemographic questions. Completion time was approximately 18 min ($mdn = 18.3$). The study's procedures were approved by the Social Research Ethics Committee of the University of Castilla-La Mancha on December 28, 2021 (protocol number CEIS-615709-Q7H1).

3.2 Experimental Design and Variables of Interest

This study employed a list experiment to examine the prevalence of purchasing sexual services. Respondents were randomly assigned to either the control ($n = 776$) or the treat-

ment ($n = 827$) condition. The control list comprised three statements: one with high expected prevalence ("I have had a beer in the last 30 days"), one with medium prevalence ("I have subscribed to a video-on-demand platform like Netflix or HBO in the last 6 months"), and one with low prevalence ("My credit card has been stolen in the last 12 months"). Respondents were asked how many of the listed behaviors they had engaged in, with purchasing sexual services ("I have paid for sex") included as a sensitive item alongside the control items in the treatment group.

To evaluate the effectiveness of this technique, respondents also answered a direct question about whether they had purchased sexual services. This question, administered before the ICT, used a dichotomous (yes/no) scale, with an option for "I prefer not to answer". The exact wording of the ICT and DQ questions is provided in Appendix B, while the expected population prevalence of the control items is reported in Appendix C.

Sociodemographic information such as *gender*, *age*, *education*, *monthly household income*, *political orientation*, *religiosity*, and *country of birth* was collected to ensure no significant differences in sample composition between the experimental groups due to random assignment or non-response. We also gathered data on the *device* respondents used to complete the survey.

3.3 Analytic Strategy

All statistical analyses were conducted using Stata 16 and R (version 4.2.3). We began by describing sample characteristics by experimental condition. Given the gendered nature of the topic and significant differences in group composition, we analyzed the responses to the direct question and the number of behaviors reported in the ICT separately for men and women. Finally, we estimated a Poisson regression model, with the experimental condition as the predictor while controlling for respondents' gender. Variance Inflation Factors (VIF) fell within appropriate limits, suggesting no multicollinearity problems ($1.01 \leq VIF \leq 1.41$).

Due to missing data in some variables, multiple imputation procedures were applied when estimating the Poisson regression. The variables with missing data were *monthly household income* (22%), *political orientation* (7%), *religiosity* (4%), and *purchasing sexual services* (2%). Multiple imputation was used to create and analyze 20 imputed datasets, with incomplete variables imputed under fully conditional specification. The choice of 20 datasets follows standard recommendations to ensure stability and efficiency of parameter estimates while also considering the proportion of missing cases in our dataset. Little's test provided no evidence against the assumption that data were missing completely at random conditional on fully observed

covariates ($\chi^2 = 157.58$, $df = 200$, $p = 0.988$). As a sensitivity analysis, we re-estimated the model using complete cases only ($n = 1177$); results were substantively identical to those obtained with multiple imputation (see Table 3A in Appendix E).

4 Results

4.1 Sample Composition

Table 1 presents the sociodemographic characteristics of the sample by experimental condition. Significant differences were observed only in gender ($\chi^2 = 6.42$, $df = 1$, $p = 0.011$), with a higher proportion of men in the control group.

This overrepresentation could influence ICT outcomes, as gender differences in engagement with control items (e.g., drinking beer) may lead to discrepancies in baseline counts, underscoring the importance of controlling for gender in subsequent analyses.

Most respondents (94%) were born in Spain, and the majority used a smartphone to complete the survey (65%). Approximately half of the sample had completed a college education (49%) and identified as religious (52%). Slightly over half (56%) reported a monthly income exceeding €1800. Politically, respondents leaned toward the center-left on average ($M = 4.05$, on a scale from 0—left to 10—right).

Table 1

Sociodemographic Characteristics of the Sample by Experimental Condition

Variable	Full sample (N = 1603)	Control group (n = 776)	Treatment group (n = 827)
<i>Gender</i> ($\chi^2 (1) = 6.42$, $p = 0.011$)			
Men	48% (776)	52% (401)	45% (375)
Women	52% (827)	48% (375)	55% (452)
<i>Age</i> ($t (1597.78) = 0.98$, $p = 0.329$)			
	45.66 (15.56)	46.05 (15.41)	45.29 (15.71)
<i>Education</i> ($\chi^2 (3) = 2.73$, $p = 0.436$)			
School graduate or less	12% (193)	12% (95)	12% (98)
High school or technical school	39% (632)	41% (318)	38% (314)
College graduate	37% (588)	35% (269)	39% (319)
Master or doctorate	12% (190)	12% (94)	12% (96)
<i>Monthly household income</i> ($\chi^2 (3) = 1.24$, $p = 0.744$)			
≤ €1200	22% (280)	21% (128)	23% (152)
€1201–€1800	22% (272)	23% (137)	21% (135)
€1801–€2400	21% (266)	21% (128)	21% (138)
> €2400	35% (432)	34% (205)	35% (227)
<i>Political orientation</i> (range 0—left to 10—right) $t (1476.63) = 0.81$, $p = 0.417$			
	4.05 (2.76)	4.11 (2.80)	4.00 (2.72)
<i>Religiosity</i> ($\chi^2 (1) = 1.06$, $p = 0.303$)			
Religious	52% (802)	51% (381)	53% (421)
Non-religious	48% (736)	49% (369)	47% (367)
<i>Country of birth</i> ($\chi^2 (1) = 0.05$, $p = 0.816$)			
Spain	94% (1500)	93% (725)	94% (775)
Other	6% (103)	7% (51)	6% (52)
<i>Device</i> ($\chi^2 (2) = 1.00$, $p = 0.608$)			
Smartphone	65% (1038)	64% (498)	65% (540)
Desktop	33% (525)	34% (261)	32% (264)
Tablet	3% (40)	2% (17)	3% (23)

Percentages may not total 100% due to rounding

4.2 Prevalence of Purchasing Sexual Services (DQ)

Responses to the direct question about purchasing sexual services revealed that 10% of respondents reported having paid for sex. Significant differences were found between men and women ($\chi^2 = 176.13$, $df = 2$, $p \leq 0.001$), with a medium effect size ($V = 0.33$). Nearly one in five men (20%) admitted to paying for sex, compared to 1% of women. Men were also more likely than women to select the “prefer not to answer” option (3% *versus* 1%).

4.3 Prevalence of Purchasing Sexual Services (ICT)

The results indicated no significant differences in the average number of statements reported between experimental groups ($t = 1.59$, $df = 1600.03$, $p = 0.113$). For men, the average number of behaviors selected in the control group was 1.20 (SD = 0.81), slightly decreasing to 1.18 (SD = 0.91) in the treatment group, where a higher average was expected due to the inclusion of an additional sensitive behavior. Similarly, for women, the average number of behaviors selected was 1.06 (SD = 0.78) in the control group, decreasing slightly to 0.97 (SD = 0.83) in the treatment group. These findings suggest an estimated prevalence of purchasing sexual services of 0% for both men and women.

Additionally, we estimated the prevalence of purchasing sexual services using the *combinedListDirect* function from the R package *list* (Blair & Imai, 2010), which integrates data from the *list* experiment and direct questioning. The estimated prevalence was -0.012 (SE = 0.04). This negative estimate reiterates that, on average, respondents in the treatment group reported fewer items than those in the control group.

To contextualize the magnitude of this underestimation, we calculated what the ICT estimates would have looked like had they aligned with the DQ results. If the ICT had captured the 20% prevalence observed among men in the DQ, we would have expected the treatment group mean to increase by approximately 0.195 (from 1.20 to around 1.395). Instead, we observed a decrease of 0.02. Similarly, for women, alignment with the 1% DQ prevalence would have implied a treatment group mean increase of approximately 0.007, yet we observed a decrease of 0.09. While the ICT estimates would not be expected to match these calculations exactly due to variability, the observed pattern—where treatment group means were consistently lower than control group means—suggests that the ICT failed to detect the sensitive behavior, yielding estimates inconsistent with any positive prevalence.

4.4 ICT Performance

To further evaluate the validity of the ICT, we applied the *ict.test* function from the *list* package to examine whether the inclusion of the sensitive item altered responses to the control items. In the context of ICT, this phenomenon, referred to as a *design effect*, occurs when respondents answer the control items differently simply because the sensitive item is present, violating the assumption that responses to the control items should remain unaffected by its inclusion (Blair & Imai, 2012). The Bonferroni-corrected p -value was 0.013, providing evidence of such an effect. In other words, the presence of the sensitive item influenced responses to the control items, compromising the validity of the ICT estimates.

To further understand how this design effect manifested across respondents, we conducted a placebo test examining ICT performance separately by DQ response (Aronow et al., 2015). Among respondents who admitted to purchasing sexual services ($n = 157$), the expected treatment-control difference should be approximately 1. Instead, we observed a difference of 0.52 (SE = 0.14, $p < 0.001$), suggesting that even self-admitted purchasers only partially included this item in their treatment group counts or deflated their control item responses. Among those who answered “no” ($n = 1413$), the treatment-control difference was -0.12 (SE = 0.04, $p = 0.005$), which theoretically should be close to zero. These results reveal two distinct but related violations of ICT assumptions. First, respondents denying the behavior paradoxically reported fewer items when the sensitive statement was included, providing evidence of a design effect whereby the sensitive item’s presence altered responses to control items. Second, even among the 157 respondents who had openly admitted purchasing sexual services in the DQ, only about half appeared to include this item in the *list* experiment (observed difference = 0.52 *versus* expected 1). This latter finding indicates that even respondents who had already disclosed the behavior in the DQ did not report it consistently in the ICT.

The Poisson regression results, detailed in Appendix D, further supported the bivariate findings, indicating that women reported fewer behaviors on average compared to men (IRR = 0.82, SE = 0.04, $p < 0.001$). Additionally, the treatment condition had no significant effect on the number of behaviors selected by respondents (IRR = 0.95, SE = 0.04, $p = 0.190$).

5 Discussion

This study documents an unsuccessful application of the ICT to measure the purchase of sexual services in Spain. Contrary to expectations, the ICT failed to produce plau-

sible prevalence estimates, yielding values that were lower than those obtained through DQ and even negative when estimated using the combined list-direct indicator. Rather than treating this outcome as a mere null finding, we use it to identify the methodological conditions that may have contributed to this failure and to derive practical recommendations for future applications of the ICT.

Several methodological factors might help explain why the ICT failed to perform as expected in the current study. First, the design of the control list may have limited the technique's effectiveness. Control items must be common enough to be relatable but not sensitive, so that the inclusion of the sensitive behavior does not distort baseline responses. Although the selected items met these criteria, the control list contained only three statements, which may have limited its ability to conceal the sensitive item. While increasing the number of items can improve concealment, it also increases the sampling error of ICT estimates (Neal, 2024), creating a difficult trade-off. In addition, the list did not include mutually exclusive items, as recommended in previous research (Blair & Imai, 2012). Because the ICT assumes that adding the sensitive item does not alter responses to the control items, the significant design effect observed in our study suggests that this assumption was violated, with the presence of the sensitive item influencing responses to the non-sensitive ones.

A second factor concerns the limited variability of the control list. The low average number of selected behaviors ($M = 1.10$) and the high proportion of respondents in both groups selecting zero items (21–28%; see Appendix F) indicate restricted variability in the control items (Groenitz, 2014). In our study, one control item was rare (i.e., credit card theft), one moderately prevalent (i.e., recent subscription to a video-on-demand platform), and one highly prevalent (i.e., drinking beer). This configuration may not have provided respondents with a sufficiently even distribution of opportunities to report the listed behaviors, leading many participants to report no applicable items. Such constrained variability could have reduced the ability of the ICT to distinguish between the treatment and control conditions and weakened its capacity to mask the sensitive item effectively. Pretesting control lists to ensure an appropriate range of expected prevalence and including four items, or up to five when variability in the control items is limited (Glynn, 2013; Neal, 2024), while balancing content across demographic groups, may therefore improve the discriminatory capacity of the technique.

The inherently lower statistical efficiency of the ICT may have further contributed to its poor performance. As noted by Neal (2024), ICT estimates typically exhibit substantially larger sampling errors than DQ estimates because prevalence is calculated indirectly as the difference between the average counts reported by the treatment and

control groups. Consequently, sampling variability from both groups contributes to the variance of the final estimate. Given the low prevalence of the target behavior and the limited heterogeneity of the control list, statistical power may have been insufficient to detect meaningful differences, increasing the likelihood of null or even negative estimates. These findings reinforce concerns that stable ICT estimation for low-prevalence, highly stigmatized behaviors may require substantially larger sample sizes than conventional direct questioning.

The nature of the behavior being measured may also have contributed to the observed failure. Li and Van den Noortgate (2022) found that the ICT tends to be more effective at reducing socially desirable responses than at increasing disclosure of socially undesirable behaviors. When behaviors are highly stigmatized, respondents may remain reluctant to acknowledge them even when using an indirect questioning technique.

Another methodological consideration concerns the administration of both the DQ and the ICT within the same survey. All respondents first completed the DQ and subsequently the ICT. Although this design does not compromise the anonymity provided by the list experiment, prior exposure to the sensitive item may have heightened respondents' awareness of the survey's purpose and influenced responses to both the control and sensitive items. Aronow et al. (2015) showed that administering both measures to the same respondents may produce question order effects and recommended either counterbalancing the order of administration or assigning respondents to separate experimental conditions. Future research should therefore examine whether administering the DQ and the ICT to independent samples yields more robust estimates.

Respondent satisficing may have provided an additional explanation for the observed findings. Both the DQ and the ICT appeared toward the end of a questionnaire comprising up to 93 items. As respondents progress through lengthy surveys, fatigue and declining motivation may reduce the cognitive effort devoted to answering questions, particularly those requiring more complex processing (Krosnick, 1991). Because the ICT requires respondents to evaluate multiple statements while simultaneously applying an indirect response format, satisficing may have increased response error and contributed to the observed design effect. Future studies should examine whether administering the ICT earlier in the questionnaire improves its performance.

Finally, the unfamiliar nature of the ICT may itself have affected respondents' comprehension and responses. Kuha and Jackson (2014) noted that respondents may find unconventional survey techniques confusing, particularly when questions depart from standard formats. This concern is especially relevant for the ICT, which presents a list of seemingly unrelated items without explaining why respon-

dents are asked to report only the total number that applies to them. In our study, the rationale underlying the technique was intentionally not explained to respondents (see Appendix C for the exact wording), which may have contributed to uncertainty or mistrust. Extensive pretesting, including cognitive interviews and pilot studies, together with clear but neutral instructions, may therefore improve respondent understanding without undermining the indirect nature of the technique.

Although the use of an opt-in online panel is unlikely to fully explain the outcome observed, it may have influenced the performance of the ICT. Compared with respondents recruited through probability-based surveys, panel members voluntarily enroll in survey research and are typically more familiar with questionnaire formats and repeated participation. This familiarity may affect how they respond to both direct and indirect questioning techniques, including their perceptions of anonymity and disclosure when answering sensitive questions. Future research should therefore examine whether similar patterns of ICT performance emerge in probability-based samples and respondents recruited in cross-sectional studies.

6 Conclusion

Overall, this unsuccessful implementation provides several practical lessons for researchers applying the ICT to sensitive topics. Data quality may be improved through careful design of control lists, rigorous pretesting, and clear instructions, particularly in self-administered surveys. Including a placebo item, that is, a sensitive item with a known or expected prevalence, may provide a useful diagnostic check of ICT performance. Implausible estimates for the placebo item can indicate that respondents did not engage with the task as intended or that unintended design effects are present (Riambau & Ostwald, 2021). Researchers should also consider routinely including direct questions alongside the ICT, where appropriate, to enable diagnostic placebo tests (Aronow et al., 2015), and, where feasible, combining the ICT with complementary indirect questioning techniques. Alternative ICT designs, such as double-list experiments, in which respondents complete two complementary lists to improve the precision of prevalence estimates, may further improve statistical efficiency by reducing the variance of ICT estimates while preserving respondent anonymity (Diaz, 2024; Glynn, 2013; Lépine et al., 2020). By translating a methodological failure into actionable evidence, this study underscores the value of transparently reporting unsuccessful implementations, thereby helping to refine indirect questioning methods and improve our understanding of the conditions under which the ICT performs well and those under which it does not.

References

- Aronow, P.M., Coppock, A., Crawford, F.W., & 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. <https://doi.org/10.1093/jssam/smu023>.
- Blair, G., & Imai, K. (2010). *list: Statistical methods for the item count technique and list experiment*. <https://CRAN.R-project.org/package=list>
- Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis*, 20, 47–77. <https://doi.org/10.1093/pan/mpr048>.
- Blair, G., Coppock, A., & Moor, M. (2020). When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments. *American Political Science Review*, 114, 1297–1315. <https://doi.org/10.1017/s0003055420000374>.
- Diaz, G. (2024). Assessing the validity of prevalence estimates in double list experiments. *Journal of Experimental Political Science*, 11(2), 162–174. <https://doi.org/10.1017/XPS.2023.24>.
- Ehler, I., Wolter, F., & Junkermann, J. (2021). Sensitive questions in surveys. A comprehensive meta-analysis of experimental survey studies on the performance of the item count technique. *Public Opinion Quarterly*, 85(1), 6–27. <https://doi.org/10.1093/poq/nfab002>.
- Glynn, A.N. (2013). What can we learn with statistical truth serum? *Public Opinion Quarterly*, 77(S1), 159–172. <https://doi.org/10.1093/poq/nfs070>.
- Groenitz, H. (2014). Improvements and extensions of the item count technique. *Electronic Journal of Statistics*, 8, 2321–2351. <https://doi.org/10.1214/14-EJS951>.
- Jonsson, S. (2023). Does criminalizing the purchase of sex reduce sex-buying? Evidence from a European survey on prostitution. *European Journal of Law and Economics*, 56(1), 91–115. <https://doi.org/10.1007/s10657-023-09778-5>.
- Kramon, E., & Weghorst, K. (2019). (Mis)Measuring sensitive attitudes with the list experiment: Solutions to list experiment breakdown in Kenya. *Public Opinion Quarterly*, 83(S1), 236–263. <https://doi.org/10.1093/poq/nfz009>.
- 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>.
- Kuha, J., & Jackson, J. (2014). The item count method for sensitive survey questions: Modelling criminal behaviour. *Journal of the Royal Statistical Society*,

- Series C (Applied Statistics)*, 63(2), 321–341. <https://doi.org/10.1111/rssc.12018>.
- Lépine, A., Treibich, C., & D'Exelle, B. (2020). Nothing but the truth: Consistency and efficiency of the list experiment method for the measurement of sensitive health behaviours. *Social Science & Medicine*, 266, 113326. <https://doi.org/10.1016/j.socscimed.2020.113326>.
- Li, J., & Van den Noortgate, W. (2022). A meta-analysis of the relative effectiveness of the item count technique compared to direct questioning. *Sociological Methods & Research*, 51(2), 760–799. <https://doi.org/10.1177/0049124119882468>.
- Miller, J. D. (1984). A new survey technique for studying deviant behavior. [Doctoral dissertation], George Washington University.
- Neal, Z. P. (2024). Large sampling errors when using the unmatched count technique to estimate prevalence: A simulation study. *Survey Practice*. <https://www.surveypractice.org/article/92297-large-sampling-errors-when-using-the-unmatched-count-technique-to-estimate-prevalence-a-simulation-study>.
- R Core Team (2023). *R: A language and environment for statistical computing (Version 4.2.0)*. www.R-project.org Computer software.
- Riambau, G., & Ostwald, K. (2021). Placebo statements in list experiments: Evidence from a face-to-face survey in Singapore. *Political Science Research and Methods*, 9(1), 172–179. <https://doi.org/10.1017/psrm.2020.18>.
- Rinken, S., Pasadas-del-Amo, S., Rueda, M., & Cobo, B. (2021). No magic bullet: Estimating anti-immigrant sentiment and social desirability bias with the item-count technique. *Quality & Quantity*, 55, 2139–2159. <https://doi.org/10.1007/s11135-021-01098-7>.
- Rinken, S., Pasadas-del-Amo, S., & Trujillo-Carmona, M. (2024). A scent of strategy: Response error in a list experiment on anti-immigrant sentiment. *Methods, data, analyses*, 18(2), 249–262. <https://doi.org/10.12758/mda.2024.05>.
- Roe-Sepowitz, D., Bontrager, S., Pickett, J. T., & Kosloski, A. E. (2019). Estimating the sex buying behavior of adult males in the United States: List experiment and direct question estimates. *Journal of Criminal Justice*, 63, 41–48. <https://doi.org/10.1016/j.jcrimjus.2019.04.005>.
- Sha, M., & Gabel, T. (2020). *The essential role of language in survey research*. RTI International.
- Starosta, A. J., & Earleywine, M. (2014). Assessing base rates of sexual behavior using the unmatched count technique. *Health Psychology and Behavioral Medicine*, 2(1), 198–210. <https://doi.org/10.1080/21642850.2014.886957>.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883. <https://doi.org/10.1037/0033-2909.133.5.859>.
- Yan, T. (2021). Consequences of asking sensitive questions in surveys. *Annual Review of Statistics and Its Application*, 8, 109–127. <https://doi.org/10.1146/annurev-statistics-040720-033353>.