

IN-BRIEF NOTES

# Large sampling errors when using the Unmatched Count Technique to estimate prevalence: A simulation study

Zachary Neal<sup>1</sup> 

<sup>1</sup> Psychology, Michigan State University

Keywords: sampling, error, unmatched count technique, simulation

<https://doi.org/10.29115/SP-2024-0002>

---

## Survey Practice

Vol. 17, 2024

---

The Unmatched Count Technique (UCT) is a method for ensuring respondent anonymity and thereby providing an unbiased estimate of the prevalence of a characteristic in a population. I illustrate that under realistic conditions UCT estimates can have ten times more sampling error than estimates derived from direct questions, and that UCT estimates can take nonsensical negative values. Therefore, the UCT should be used with caution.

The prevalence of a particular characteristic in a population is typically estimated by directly asking respondents if they have the characteristic (i.e., the direct question technique or DQT; Fowler 2013). However, DQT estimates of the prevalence of some characteristics can be biased, that is, they can differ systematically from the true value. For example, directly asking respondents whether they have committed a crime can yield downwardly biased estimates because respondents are reluctant to admit criminality. The Unmatched Count Technique (UCT) aims to avoid such bias by not asking respondents directly about a characteristic. It has been used to estimate the prevalence of characteristics about which respondents might underreport, including sexual behaviors (LaBrie and Earleywine 2000; Bahadivand et al. 2020), misconduct (e.g., Dalton, Wimbush, and Daily 1994), conservation behaviors (Hinsley et al. 2019; Spira et al. 2021; Olmedo et al. 2022), and violent extremism (Clemmow et al. 2020).

UCT estimates are less biased (Ehler, Wolter, and Junkermann 2021; Li and Van den Noortgate 2022), but they have more variance and therefore more sampling error (Raghavarao and Federer 1979), than DQT estimates. This means that, although *on average* UCT will correctly estimate prevalence, *in any given sample* it can yield substantially larger or smaller estimates, and can even yield nonsensically negative estimates. For example, as I show below, when used to estimate a characteristic whose true prevalence is 5%, the UCT is likely to yield an estimate somewhere between an impossible -11% and an implausible 22%. Having an unbiased estimate of prevalence is useful; however, an unbiased estimate with such large sampling error is not useful because it cannot be used to make informative inferences about the population. Because UCT prevalence estimates have very large sampling errors, and because UCT data cannot be used for anything except estimating prevalence, researchers should use UCT with caution.

## Methods

The DQT simply asks respondents a question like “Do you X”, where X is a focal characteristic, and prevalence is estimated as the percent answering yes. In contrast, the UCT asks control respondents how many of a series of control statements (e.g., “Do you like corn?”) they endorse, while it asks experimental respondents how many of these control statements *and* one focal statement (e.g., “Do you X?”) they endorse. Respondents report *how many*, but not *which*, statements they endorse, thereby preserving anonymity and encouraging honesty. The prevalence of focal characteristic X is estimated as the difference in the total number of endorsements by the control and experimental respondents, divided by the number of experimental respondents. To estimate the sampling error of DQT and UCT estimates, I simulate collecting data using these two techniques from 1000 samples of 1000 respondents each, where the true prevalence is  $p$ .

For DQT, I simulate a sample of 1000 respondents endorsing a single statement about a characteristic, where each respondent has probability  $p$  of endorsing. I repeat this 1000 times, measuring the sampling error of the DQT as the standard deviation of the 1000 DQT estimates (i.e., the standard error).

For UCT, I simulate a sample of 500 experimental respondents endorsing  $s$  statements, and 500 control respondents endorsing  $s - 1$  statements. The experimental respondents have a probability  $p$  of endorsing one of these statements, which represents a statement about a focal characteristic. All respondents have a 50% probability of endorsing all other statements, which represent control statements. Endorsements for pairs of statements have correlation  $r$ . I repeat this 1000 times, measuring the sampling error of the UCT as the standard deviation of the 1000 UCT estimates (i.e., the standard error).

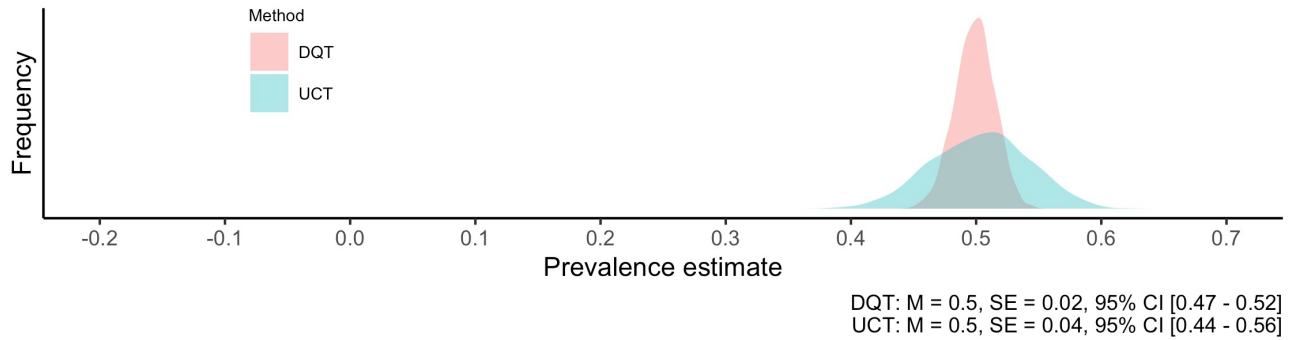
I examine the simulated sampling error of DQT and UCT estimates by varying the number of statements ( $s = 2, 4, \text{ or } 6$ ), true prevalence ( $p = 0.5, 0.1, \text{ or } 0.05$ ), and correlation among statements ( $r = 0, 0.1, \text{ or } 0.2$ ). The code necessary to reproduce these results, or to try other values, is available at <https://osf.io/6t7me/>.

## Results

[Figure 1](#) displays the simulated sampling distribution of DQT and UCT prevalence estimates under a range of conditions defined by the true prevalence ( $p$ ), number of statements used by the UCT ( $s$ ), and correlation among UCT statements ( $r$ ). For each set of conditions, each panel also reports the mean, standard error, and 95% confidence interval of each estimate’s sampling distribution.

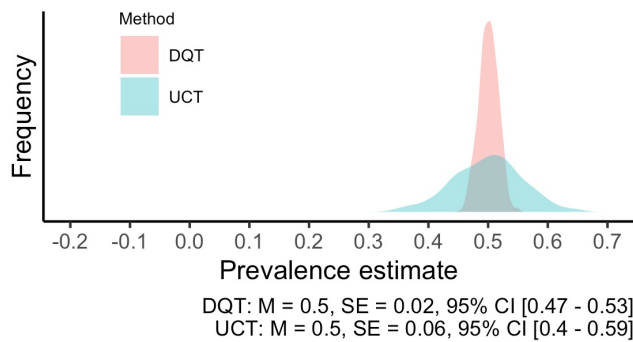
### A The best case scenario for UCT

( $p = 0.5$ ,  $s = 2$ ,  $r = 0$ )



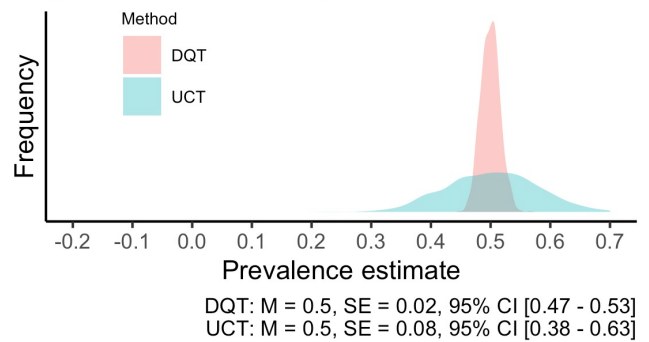
### B More statements

( $p = 0.5$ ,  $s = 4$ ,  $r = 0$ )



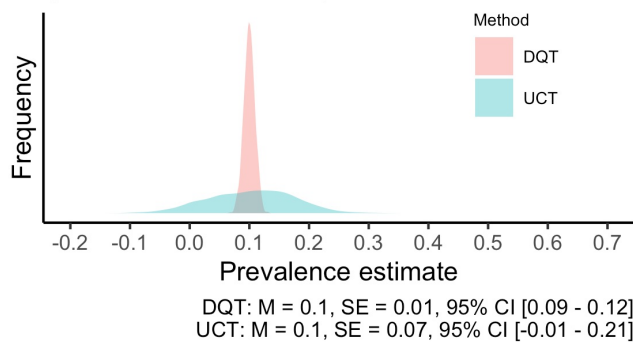
### C Even more statements

( $p = 0.5$ ,  $s = 6$ ,  $r = 0$ )



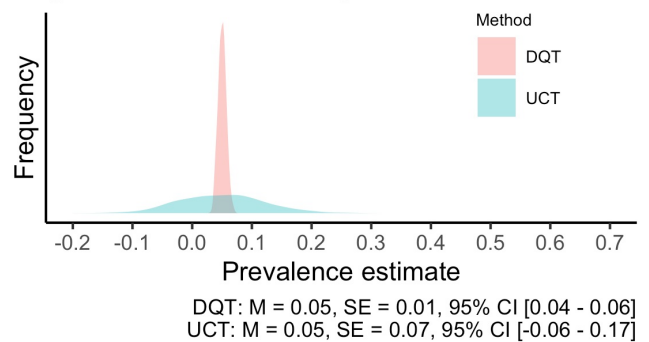
### D ...and lower true prevalence

( $p = 0.1$ ,  $s = 6$ ,  $r = 0$ )



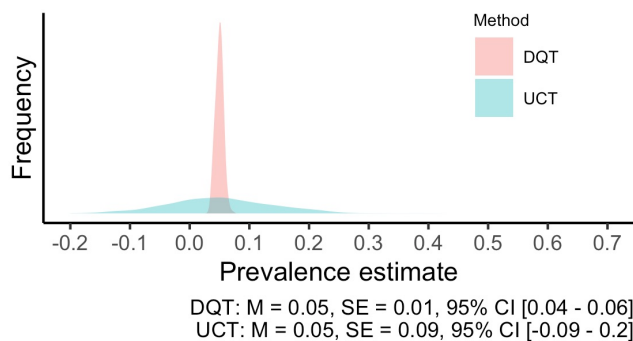
### E ...and even lower true prevalence

( $p = 0.05$ ,  $s = 6$ ,  $r = 0$ )



### F ...and correlated statements

( $p = 0.05$ ,  $s = 6$ ,  $r = 0.1$ )



### G ...and more highly correlated statements

( $p = 0.05$ ,  $s = 6$ ,  $r = 0.2$ )

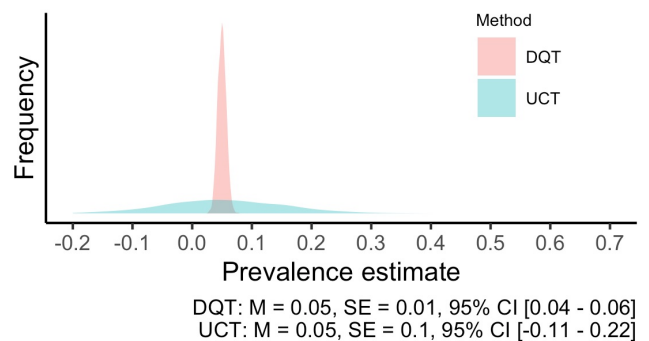


Figure 1. Simulated sampling distributions of prevalence estimates obtained using the direct question technique (DQT) and the unmatched count technique (UCT).

Note:  $p$  = true prevalence,  $s$  = number of statements,  $r$  = correlation among statements

In [Figure 1A](#), the true prevalence is 50% ( $p = 0.5$ ), and the UCT uses two ( $s = 2$ ) uncorrelated ( $r = 0$ ) statements. These conditions represent the best case scenario for a UCT estimate, that is, the conditions when a UCT estimate has the smallest sampling error compared to a DQT estimate. However, even under these “best case” conditions, the UCT estimate’s sampling error (0.04) is still twice that of the DQT estimate (0.02). Moreover, these conditions are implausible. First, the UCT must use more than two statements to preserve respondent anonymity. Second, the UCT is used to estimate the prevalence of rare and potentially stigmatized, not common (i.e.  $p = 0.5$ ), characteristics. Third, it is nearly impossible for respondents’ answers to statements to be perfectly uncorrelated. The following panels explore how the UCT estimate’s sampling error changes as these conditions become more realistic.

To preserve respondent anonymity, the UCT requires asking one focal statement and multiple control statements. As the number of control statements increases, the UCT estimate’s sampling error also increases. [Figures 1B and 1C](#) retain the conditions shown in [Figure 1A](#), except that the UCT uses more statements. In [Figure 1B](#) the UCT uses four statements and its sampling error (0.06) is three times larger than that of the DQT estimate (0.02), while in [Figure 1C](#) the UCT uses six statements and its sampling error (0.08) is four times larger than that of the DQT estimate (0.02).

The UCT is primarily used to estimate the prevalence of rare characteristics about which respondents might be reluctant to report, for example, because they are stigmatized. As the true prevalence of the characteristic becomes rarer, the UCT estimate shifts to the left and its sampling error increases compared to the DQT. [Figures 1D and 1E](#) retain the conditions shown in [Figure 1C](#), except that the true prevalence is lower. In [Figure 1D](#) the true prevalence is 10% (i.e.,  $p = 0.1$ ), while in [Figure 1E](#) the true prevalence is 5% (i.e.,  $p = 0.05$ ). In both cases the UCT estimate’s sampling error (0.07) is seven times larger than that of the DQT estimate (0.01). Moreover, the UCT estimate’s large sampling error yields a wide 95% confidence interval, which includes impossible negative prevalence estimates as low as -6%.

Although the control statements used in the UCT should not be associated with each other or with the focal statement, it is impossible to choose statements that are perfectly uncorrelated. In practice these statements will exhibit some correlation, and as the correlation among statements increases, the UCT estimate’s sampling error increases. [Figures 1F and 1G](#) retain the conditions shown in [Figure 1E](#), except that there is some correlation among the UCT statements. In [Figure 1F](#) the UCT statements are correlated at  $r = 0.1$  and the UCT estimate’s sampling error (0.09) is nine times larger than that of the DQT estimate (0.01), while in [Figure 1G](#) the UCT statement are correlated at  $r = 0.2$  and the UCT estimate’s sampling error (0.1) is ten times larger than that of the DQT estimate (0.01).

## Conclusion

UCT yields less biased prevalence estimates than DQT (Ehler, Wolter, and Junkermann 2021; Li and Van den Noortgate 2022). However, this reduced bias comes at the cost of substantially increased sampling error. Under realistic conditions – six weakly correlated statements to estimate the prevalence of a rare characteristic (see [Figure 1G](#), where  $p = 0.05$ ,  $s = 6$ ,  $r = 0.2$ ) – the 95% confidence interval of a UCT estimate would lead a researcher to conclude that the true prevalence (here, 5%) is somewhere between an impossible -11% and an implausible 22%. More generally, under practical conditions UCT estimates have so much sampling error that inferences about the prevalence of a characteristic in a population are uninformative, and in many samples it may even yield nonsensical negative prevalence estimates.

UCT prevalence estimates can be both imprecise and invalid, but UCT data also cannot be used for anything except estimating prevalence. Specifically, because it is unknown which respondents endorsed which statements, these data cannot be used in analysis at the individual level. For example, consider a survey that collects respondents' gender and uses UCT to estimate the prevalence of criminality. Using these data it is not possible to test whether criminality is more prevalent among men or women because it is unknown whether any given man or women participating in the survey reported having committed a crime.

The UCT is only capable of estimating prevalence; however, the illustrations in [Figure 1](#) show that it can yield both imprecise and invalid prevalence estimates. Therefore, researchers should use UCT with caution. In cases where DQT estimates are likely to be biased, and therefore UCT estimates are preferred, researchers can reduce the estimate's sampling error by using the smallest number of statements that will preserve respondents' sense of anonymity, using uncorrelated statements, and using larger samples. However, even when taking these steps, inferences of population prevalence derived from UCT estimates should be interpreted with caution.

.....

## *Lead author's contact information*

Zachary P. Neal  
Michigan State University  
[zpneal@msu.edu](mailto:zpneal@msu.edu)

Submitted: July 12, 2023 EST, Accepted: January 15, 2024 EST

## REFERENCES

- Bahadivand, Samira, Amin Doosti-Irani, Manoochehr Karami, Mostafa Qorbani, and Younes Mohammadi. 2020. "Prevalence of High-Risk Behaviors in Reproductive Age Women in Alborz Province in 2019 Using Unmatched Count Technique." *BMC Women's Health* 20 (1): 1–6. <https://doi.org/10.1186/s12905-020-01056-9>.
- Clemmow, Caitlin, Sandy Schumann, Nadine L. Salman, and Paul Gill. 2020. "The Base Rate Study: Developing Base Rates for Risk Factors and Indicators for Engagement in Violent Extremism." *Journal of Forensic Sciences* 65 (3): 865–81. <https://doi.org/10.1111/1556-4029.14282>.
- Dalton, Dan R., James C. Wimbush, and Catherine M. Daily. 1994. "Using the Unmatched Count Technique (UCT) to Estimate Base Rates for Sensitive Behavior." *Personnel Psychology* 47 (4): 817–29. <https://doi.org/10.1111/j.1744-6570.1994.tb01578.x>.
- Ehler, Ingmar, Felix Wolter, and Justus Junkermann. 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>.
- Fowler, F.J. 2013. *Survey Research Methods*. Thousand Oaks, CA: Sage.
- Hinsley, Amy, Aidan Keane, Freya A. V. St. John, Harriet Ibbett, and Ana Nuno. 2019. "Asking Sensitive Questions Using the Unmatched Count Technique: Applications and Guidelines for Conservation." *Methods in Ecology and Evolution* 10 (3): 308–19. <https://doi.org/10.1111/2041-210x.13137>.
- LaBrie, Joseph W., and Mitchell Earleywine. 2000. "Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique." *The Journal of Sex Research* 37 (4): 321–26. <https://doi.org/10.1080/00224490009552054>.
- Li, Jiayuan, and Wim Van den Noortgate. 2022. "A Meta-Analysis of the Relative Effectiveness of the Item Count Technique Compared to Direct Questioning." *Sociological Methods & Research* 51 (2): 760–99. <https://doi.org/10.1177/0049124119882468>.
- Olmedo, Alegria, Diogo Veríssimo, E.J. Milner-Gulland, Amy Hinsley, Huong Thi Thu Dao, and Daniel W.S. Challender. 2022. "Uncovering Prevalence of Pangolin Consumption Using a Technique for Investigating Sensitive Behaviour." *Oryx* 56 (3): 412–20. <https://doi.org/10.1017/s0030605320001040>.
- Raghavarao, D., and W. T. Federer. 1979. "Block Total Response as an Alternative to the Randomized Response Method in Surveys." *Journal of the Royal Statistical Society: Series B (Methodological)* 41 (1): 40–45. <https://doi.org/10.1111/j.2517-6161.1979.tb01055.x>.
- Spira, Charlotte, Rivo Raveloarison, Morgane Cournarie, Samantha Strindberg, Tim O'Brien, and Michelle Wieland. 2021. "Assessing the Prevalence of Protected Species Consumption by Rural Communities in Makira Natural Park, Madagascar, through the Unmatched Count Technique." *Conservation Science and Practice* 3 (7): e441. <https://doi.org/10.1111/csp2.441>.