Selection bias in voluntary random testing: Evidence from a COVID-19 antibody study

By Deniz Dutz, Michael Greenstone, Ali Hortaçsu, Santiago Lacouture, Magne Mogstad, Azeem M. Shaikh, Alexander Torgovitsky & Winnie van Dijk*

Consider the objective of determining the fraction of a specific population who have been infected with COVID-19. A common way to approach this problem is to use data on confirmed cases among those who have been tested. It is well appreciated that this may lead to biased estimates. Data are missing for individuals who have not been tested, and these individuals are likely to differ from those who have been tested for both observable and unobservable reasons.

One way to address this missing data problem is to combine the data on those who have been tested with statistical assumptions to construct bounds on the population infection rate. Manski and Molinari (2021) apply such an approach to observational data from Illinois, New York, and Italy. They conclude the resulting bounds are too wide to be economically informative, unless untenable assumptions are made. Thus, the authors argue, better data is needed, such as data obtained through random testing of the population of interest.

* Dutz: Department of Economics, University of Chicago (ddutz@uchicago.edu). Greenstone: Department of Economics, University of Chicago (mgreenst@uchicago.edu). Hortaçsu: Department of Economics, University of Chicago (hortacsu@uchicago.edu). Lacouture: Department of Economics, University of Chicago (slacouture@uchicago.edu). Mogstad: Department of Economics, University of Chicago (mogstad@uchicago.edu). Shaikh: Department of Economics, University of Chicago (amshaikh@uchicago.edu). Torgovitsky: Department of Economics, University of Chicago (torgovitsky@uchicago.edu). Van Dijk: Department of Economics, Harvard University and NBER (winnie_vandijk@fas.harvard.edu). This study is approved by the IRB at the University of Chicago (IRB20-0721). We gratefully acknowledge financial support from the University of Chicago Women's Board, the Sloan Foundation, the Smith Richardson Foundation, the Becker Friedman Institute, the Booz Allen Foundation, and the Flu Lab.

Although random testing would in principle solve the missing data problem, legal and ethical barriers prohibit mandatory participation in testing. In practice, studies typically can only invite and encourage individuals to be tested. The voluntary decision to participate is likely non-random and possibly correlated with the probability of infection. Thus, the missing data problem reoccurs, and test results may be contaminated with selection bias and unrepresentative of the infection rate in the population of interest.

This paper studies this missing data problem in the context of a COVID-19 serological study that invited a random – and thus representative – sample of Chicago households to participate. Unlike typical settings, the study experimentally varied financial incentives for participation. We illustrate how randomized incentives allow for the detection and characterization of selection bias using only participant data. While the empirical results we present are new, the methods we use build on the recent work by Dutz et al. (2021) and Dutz et al. (2022).

We first use neighborhood-level data on a COVID-19 risk index to document the presence of selection bias. Crucially, this variable is observed for all invited households – participants and nonparticipants – because it is measured at the neighborhood-level. Thus, it provides a ground truth, which can be used as a basis to assess the performance of alternative methods designed to account for selection bias. We find that participants are from neighborhoods with substantially lower COVID-19 risks, compared to nonparticipants.

We next demonstrate how to use exogenous variation in participation incentives to detect selection bias in voluntary random

testing, even in the standard case when the outcome of interest is not observed for non-participants. Specifically, we reject the null hypothesis of equality in the COVID-19 risk index among participants across incentive arms. This rejection allows us to conclude, using only participant data, that the study is contaminated with selection bias.

We then assess the performance of existing methods that account for selection bias in participant samples. We find that some of the methods produce bounds that contain the population quantities, but are very wide. Other methods produce bounds (or point estimates) that are inconsistent with the true population COVID-19 risk, suggesting that the underlying assumptions are invalid.

Building on Dutz et al. (2022), we investigate these methods' failures by taking a closer look at the determinants of participation. Specifically, we present evidence that there are two types of nonparticipants: "active" nonparticipants who saw the study invitation and declined to participate because the benefits were outweighed by the costs of participating, and "passive" nonparticipants who never saw the invitation but may have participated had they seen it. We find that in terms of the COVID-19 risk index, these two types of nonparticipants differ from the average participant in opposite ways.

These findings underscore the importance of allowing for multiple dimensions of unobserved heterogeneity when accounting for selection bias. Similarly, Dutz et al. (2021) find that allowing for active and passive non-participants in a Norwegian survey produces bounds (or point estimates) that are narrower and closer to the truth than the other methods. A natural question for future research is whether such an approach can also prove useful in applications other than surveys, such as to account for selection bias in voluntary random testing. Unfortunately, the present study's sample size is too small to use this approach to draw inferences about the presence or levels of antibodies in the target population.

I. Study design and data

We use data from the Representative Community Survey Project's (RECOVER) COVID-19 serological study, which was carried out in Chicago between December 2020 and March 2021. The study was designed and implemented using best practices in collaboration with NORC, a leading national statistical agency, and the Wilson Antibody Biology Laboratory at the University of Chicago. This section discusses the main aspects of the study; see Dutz et al. (2022) for a more detailed description of the study and its implementation.

A. Design and implementation

The invited sample consisted of 882 Chicago addresses that NORC randomly sampled from United States Postal Service data. Sampled households were sent a package that contained a self-administered blood sample collection kit, and were asked to return the sample by mail to our partner research lab to be tested for COVID-19 antibodies.

Unlike typical settings, the study experimentally varied financial incentives for participation (i.e. returning a blood sample). Households in the sample were randomly assigned one of three levels of financial compensation for participating in the study: \$0, \$100, or \$500. While the first two levels of compensation are consistent with those in other serological studies, the latter is a remarkably high level of compensation for participation (see Appendix C of Dutz et al., 2022, for a comparison to other COVID-19 antibody studies). Additionally, NORC sent reminder postcards to invited households.

B. Data

The RECOVER data consists of the randomly-assigned compensation level, participation status, and addresses for each sampled household. For participants, we additionally observe the date they mailed in the blood sample.

We use addresses to link households to neighborhood-level (i.e., zipcode-level)

measures of COVID-19 risks, independently of whether the households participated in the study. This allows us to compare how participants differ from the invited sample. We focus on the COVID-19 local risk index, which measures the risk for and threat of COVID-19 infection on a 10-point scale.[1]

II. Results

For ease of exposition and to gain precision in the below analyses, we partition the three compensation levels into two incentive groups: "low" ($0 or $100) and "high" ($500). The "low" group's compensation levels are similar to those typically offered in serological studies (see Appendix C of Dutz et al., 2022).

A. *Participation rates*

By the end of the data collection period, 14.2% of invited households participated. The participation rate for the low incentive group was 11.5%. This rate is comparable to other serological studies that invited a random sample of households to be tested for COVID-19 antibodies (see Appendix C of Dutz et al., 2022). Offering high incentives for participation substantially and significantly increased participation rates to 29.1%, almost triple the rate of the low incentive group (p–value < 0.01).

B. *Selection bias*

When the invited sample is a random subsample of the population–as is the case in the RECOVER study–the mean of the invited sample is a consistent estimator of the population mean. However, this estimator is not feasible because the outcome of interest is only observed for participants.

If the decision to voluntarily participate in the study is correlated with the outcome, the unknown nonparticipant mean will differ from the participant mean. It is easy to see why this could occur when the outcome is a measure of COVID-19 risk. For example, households who choose to participate may have greater trust in the scientific community, which may be positively correlated with taking COVID-19 safety precautions. In that case, selection bias would cause the participant mean to be lower than that of nonparticipants; the result would be a (downward) biased estimate of the population mean.

To show the potential for this concern to play a role in biasing health statistics, we use the COVID-19 local risk index to measure selection bias for this outcome. The participant mean of this index is 4.3, which is a statistically significant 1 point lower than the invited sample's mean of 5.3 (p–value < 0.01). We conclude that the RECOVER study suffers from selection bias.

C. *Detecting selection bias when outcome data is only available for participants*

Data linked to the invited sample, like the neighborhood-level data we used above, do not allow us to directly test for selection bias in other outcomes of interest that are only observed for participants.

In this subsection, we illustrate how to test for selection bias for outcomes that are only available for participants. We do so by applying Dutz et al.'s (2021) test for selection bias in the context of voluntary, random testing. The test's basis is that random assignment of financial compensation can be used to detect selection bias when nonparticipant data is missing due to voluntary, random testing. The random assignment of different incentives creates ex-ante identical groups with different participation rates. The test's null hypothesis of no selection bias implies that participant means across incentive groups are equal. A rejection of the null implies that there is selection bias. Barring knife-edge cases, this implies that the participant mean is not equal to the population mean.

Using only participant data for the COVID-19 index, we find that the low and high incentive participant means are substantially and significantly different from

[1] This index is constructed by the City Health Dashboard, a portal developed by NYU Langone Health, to measure the potential for COVID-19 infection and risk for more severe COVID-19 outcomes and risks at the zipcode-level.

each other: the mean for low incentive participants is 3.9 while the mean for high incentive participants is 5.2 (p–value < 0.01).[2] Thus, using only outcomes for participants, we conclude that the RECOVER study suffers from selection bias, as we also did in the previous subsection.

D. (Attempting to) Account for selection bias

This subsection assesses the performance of several approaches to correcting for selection bias to learn about the population mean.

A frequently-used approach is to assume selection bias is entirely due to selection on observables and to reweight participants by the probability of participating conditional on observables.

To illustrate this approach, we reweight participants based on their neighborhood racial composition. Selection bias persists: our study would continue to significantly understate the average COVID-19 risk index by half a point (p–value < 0.01). We conclude that there is selection on other quantities, including possibly unobservable quantities.

We next evaluate the performance of existing methods that allow for selection on unobservables. The top row of Figure 1 presents "worst-case" bounds that arise from imposing that the unobserved nonparticipant mean necessarily lies between the endpoints of the index, i.e., between 1 and 10 (Horowitz and Manski, 1998). Although the bounds contain the actual population mean, they are wide: the bounds imply that the population value could be as low as 1.5 and as high as 9.2.

To obtain more informative estimates, we consider approaches based on existing parametric and nonparametric selection models (Mogstad and Torgovitsky, 2018, and the references therein). For the sake of brevity, we defer a description of these approaches to the figure notes of Figure 1, and we instead note that they are commonly considered in the treatment effects literature and that our implementation mirrors that of Dutz et al. (2021).

The remaining three rows of Figure 1 present results of these attempts to correct for bias. The resulting bounds and point estimates all miss the population mean. For example, under a Heckman (1979) selection model, we would estimate the mean COVID-19 risk index to be 7.2, which is far off from the true value of 5.3. Hence, conventional approaches to selection correction do not perform well and lead to incorrect conclusions about the population mean in our setting.

III. Understanding the participation decision

One explanation for the failure of the above approaches that model participation is that they presume nonparticipants declined to participate because the offered incentive was too low. However, there are two types of nonparticipants: active nonparticipants who saw the study invitation and hesitated to participate because the incentive was too low, and passive nonparticipants who were never aware of the invitation, but might have participated had they been successfully contacted. The above approaches may fail if these distinct types of nonparticipants differ in ways that correlate differently with the outcome of interest.

Using participation data from the RECOVER study, Dutz et al. (2022) show how to quantify the extent to which nonparticipation is due to "participation hesitancy" and "non-contact" and conclude that low participation is driven by both types of nonparticipants.

Figure 2 presents evidence that these two types of nonparticipants differ in their COVID-19 risks. Specifically, we split participants by whether they mailed in the blood sample before or after receiving the January 11th reminder postcard. The darker bar of Figure 2 shows the average difference between the high and low incentive groups among participants who participated before the reminder. It is apparent that households that participated because

[2] Although the high incentive participant mean happens to be similar to the population mean, a higher response rate is no guarantee that the participant mean is closer to that of the population (Dutz et al., 2021).

of the high incentives–who we interpret as active nonparticipants at low incentives–have higher COVID-19 risk scores.

The lighter bar reports the average difference between participants who participated after the reminder and participants who participated before the reminder, within the low incentive group. Strikingly, households who participate after the reminder–who we interpret as passive nonparticipants prior to the reminder–have lower COVID-19 risk scores. Thus, Figure 2 supports the possibility of two forms of selection that work in opposite directions.

These findings point to the potential for using selection models with multiple dimensions of unobserved heterogeneity to correct for selection bias. Dutz et al. (2021) develop a method to draw inference about the population mean that uses such a model and apply it in the context of a survey about labor market conditions. Using only participant data, they find that the method produces bounds and point estimates that, unlike existing methods, are narrow and contain the population mean across a wide range of outcomes. A natural question for future research is whether such an approach can also prove useful in settings other than surveys, such as to account for selection bias in voluntary random testing.

REFERENCES

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie van Dijk. 2021. "Selection in Surveys." National Bureau of Economic Research.

Dutz, Deniz, Michael Greenstone, Ali Hortaçsu, Santiago Lacouture, Magne Mogstad, Azeem Shaikh, Alexander Torgovitsky, and Winnie van Dijk. 2022. "What drives (gaps in) scientific study participation? Evidence from a COVID-19 antibody survey." Becker Friedman Institute Working Paper.

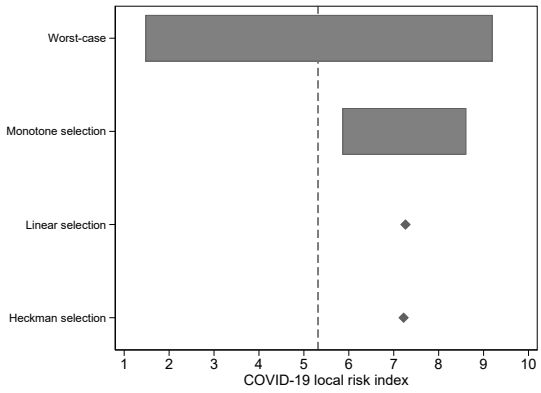
Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica: Journal of the econometric society*, 47(1): 153–161.

Horowitz, Joel L, and Charles F Manski. 1998. "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations." *Journal of Econometrics*, 84(1): 37–58.

Manski, C., and F. Molinari. 2021. "Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem." *Journal of Econometrics*, 220(S2): 181–192.

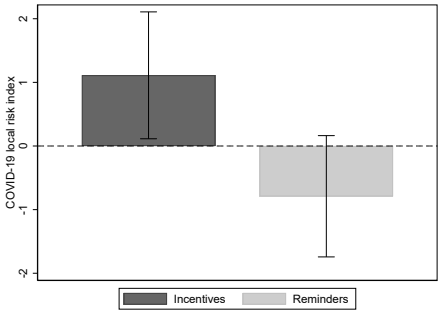
Mogstad, Magne, and Alexander Torgovitsky. 2018. "Identification and Extrapolation of Causal Effects with Instrumental Variables." *Annual Review of Economics*, 10: 577–613.

Figure 1. Estimated bounds and point estimates under various assumptions



Notes: This figure shows estimated bounds and point estimates of the population mean using participant data under various models and assumptions. The true population mean is presented as a vertical dashed line. "Worst-case" bounds arise from imposing that the outcome Y_i^* is bounded between 1 and 10 (see Horowitz and Manski, 1998). The other three results are obtained under the additional assumptions that incentives Z_i are randomly assigned and that participation R_i equals $\mathbb{1}[U_i \leq p(Z_i)]$, where U_i is a uniform $[0,1]$ latent variable, and $p(z) \equiv \mathbb{P}[R_i = 1|Z_i = z]$. Defining $m(u) \equiv \mathbb{E}[Y_i^*|U_i = u]$, the population mean is the integral of $m(u)$ over $[0,1]$, so that assumptions on $m(u)$ can help tighten inference on the population mean. 'Monotone selection' imposes that $m(u)$ is monotone, 'Linear selection' imposes that $m(u)$ is linear in u, and 'Heckman selection' imposes that $m(u)$ is linear in $\Phi^{-1}(u)$, the standard normal quantile function. For more details, see Section 5 and Online Appendix H of Dutz et al. (2021).

Figure 2. Selection patterns by incentive and reminder



Notes: The darker bar shows the average difference in the COVID-19 local risk index between high and low incentive groups among participants who mailed in the blood sample before receiving the January 11th reminder postcard. The lighter bar shows the average difference between participants who participated after the reminder and participants who participated before the reminder, within the low incentive group. 90% CIs are depicted for each difference.