Avoiding Post-Treatment Bias in Audit Experiments

Alexander Coppock, Yale University*

January 30, 2018

Audit experiments are used to measure discrimination in a large number of domains (Employment: Bertrand and Mullainathan (2004); Legislator responsiveness: Butler and Broockman (2011); Housing: Fang et al. (2018)). Audit studies all have in common that they estimate the average difference in response rates depending on randomly varied characteristics (such as the race or gender) of a requester. Scholars conducting audit experiments often seek to extend their analyses beyond the effect on response to the effects on the *quality* of the response. Response is a consequence of treatment; answering these important questions well is complicated by post-treatment bias (Montgomery et al., 2018). In this note, I consider a common form of post-treatment bias that occurs in audit experiments.

As an instructive example, consider White et al. (2015), an audit experiment in which election officials were sent emails from putatively Non-Latino White or Latino names asking "I've been hearing a lot about voter ID laws on the news. What do I need to do to vote?" Whereas Non-Latino White names received a response 70.5% of the time, Latino names were responded to 64.8% of the time, for a statistically and substantively significant difference of negative 5.7 percentage points. In a secondary analysis, the authors further estimate the effect on the friendliness of the emails *conditional on response*.

Response is a post-treatment outcome; conditioning on post-treatment outcomes "de-randomizes" an experiment in the sense that the resulting treatment and control groups no longer have potential outcomes that are in expectation equivalent. Seen another way, conditioning on a post-treatment outcome induces confounding. This problem is relatively widespread. Seven of the 20 legislative audit experiments analyzed in Costa (2017) and nine of the 29 employment audit studies analyzed in Quillian et al. (2017) inappropriately condition on response.

In this setting, a subject might be one of the four types in the table below. $R_i(Z)$ is the response potential outcome depending on whether subject i is assigned to a putatively non-Latino White name ($Z = 0$) or a putatively Latino name ($Z = 1$). Together, $R_i(1)$ and $R_i(0)$ indicate whether a subject is an Always-Responder, an If-Treated-Responder, an If-Untreated-Responder, or a Never-Responder. The friendliness potential outcome $Y_i(Z)$ is *undefined* if a subject does not respond, implying that the average treatment effect of the Latino name on friendliness *does not exist* for subjects who do not respond in one condition or the other. The average effect of treatment on Always-Responders ($\mathrm{E}\left[Y_i(1) - Y_i(0) | R_i(0) = R_i(1) = 1\right]$) does exist, but estimating it is not straightforward because we do not have complete information on who is an Always-Responder.

*The data, code, and any additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at doi:10.7910/DVN/6NVI9C. I would like to thank Ariel White, Noah Nathan, Julie Faller, Saad Gulzar, and Peter Aronow for helpful comments.

Table 1: Types of Subjects

Type	$R_i(0)$	$R_i(1)$	$Y_i(0)$	$Y_i(1)$
Always-Responder	1	1	$Y_i(0)$	$Y_i(1)$
If-Treated-Responder	0	1	NA	$Y_i(1)$
If-Untreated-Responder	1	0	$Y_i(0)$	NA
Never-Responder	0	0	NA	NA

Analysts have three main choices:

Bounds. Zhang and Rubin (2003) develop bounds around the average effect for subjects whose outcomes are never "truncated by death," regardless of treatment assignment; their result can be immediately applied to the audit study case (see Aronow et al. (2018) for an application and extension of these bounds in Political Science). The estimates correspond to the most pessimistic and most optimistic scenarios for the average treatment effect among Always-Responders. These bounds often have very large (or even infinite) width, so their scientific utility will vary depending on the application.

Find Always-Responders. If we were to assume that a particular group of subjects consisted entirely of Always-Responders, we could directly estimate the effect of treatment on the quality of response in that group. One check on the plausibility of the "Always-Responders" assumption is that the response rate in both the treatment and control groups must equal 100%. The assumption can of course still be incorrect as some treated units might not have responded if untreated (or vice-versa). Bendick et al. (1994) implicitly invokes an "Always-Responders" assumption in an analysis that conditions the dataset to include only firms who offer jobs to both White and Black confederates before estimating the average effect of race on the salary offered among this group.

Redefine the outcome. Analysts can change the outcome variable to be $Y_i^*(Z)$, which is equal to $Y_i(Z)$ if $R_i(Z) = 1$ and 0 otherwise. Crucially, this means that emails never sent are "not friendly." The average effect of treatment on this new dependent variable $\mathrm{E}\left[Y_i^*(1) - Y_i^*(0)\right]$ is well-defined. Kalla et al. (2018) use this approach; White et al. (2015) report in their footnote 29 that they ran this analysis as well.

In my reanalysis of White et al. (2015), I provide examples of all three approaches. Following the procedure in Zhang and Rubin (2003), I estimate the lower bound to be -66 points and the upper bound to be 65 points. These bounds themselves are subject to sampling variability, which I estimate via the nonparametric bootstrap. To find Always-Responders, I took advantage of the original experiment's matched-pair design. I subset the dataset to the 719 matched pairs in which both the treated and untreated pair member responded. Under the unverifiable assumption that both pair members would also have responded if the treatment assignment had been switched, I estimate the ATE among this subgroup to be -5.5 points (SE = 2.5 points). Finally, using the redefined outcome, I estimate the ATE to be a 6 point decrease in friendliness (SE = 1.7 points). As it happens, this estimate is statistically significant while the naive estimate (-3.5 percentage, SE = 2.0 points) is not. Table 2 displays all four estimates.

Table 2: Reanalysis of White, Nathan, and Faller (2015)

Estimand	Estimator	Estimate	95% CI
Undefined	Naive difference-in-means	-0.035	[-0.075, 0.005]
ATE on redefined outcome	Difference-in-means	-0.061	[-0.095, -0.027]
ATE among matched pairs in which both members are Always-Responders	Difference-in-means among matched pairs in which both members respond	-0.056	[-0.104, -0.007]
ATE among all Always-Responders	Bounds	[-0.658, 0.645]	[-0.73, 0.726]

In this note, I have outlined how in audit experiments, some causal quantities we seek to estimate do not exist. We can either attempt to recover estimates of the effect for a subgroup of units (the Always-Responders) or we can redefine the outcome so that the average treatment effect is defined. The approaches outlined here have applications beyond audit experiments. Rondeau and List (2008) seek to estimate the effect of a treatment on the size of donations; they inappropriately condition on units making any donation. Björkman and Svensson (2009) seek to estimate the effects of a monitoring intervention on child health; they inappropriately condition on infant survival. The applicability of each of the three "solutions" to the problem will depend on the substantive area, but conditioning on post-treatment variables should be avoided in all cases.

References

Aronow, P. M., J. Baron, and L. Pinson (2018). A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis*. In press.

Bendick, M., C. W. Jackson, and V. A. Reinoso (1994). Measuring Employment Discrimination through Controlled Experiments. *The Review of Black Political Economy 23*(1), 25–48.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg more employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review 94*(4), 991–1013.

Björkman, M. and J. Svensson (2009). Power to the People: Evidence from a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda. *Quarterly Journal of Economics 124*(2), 735–769.

Butler, D. M. and D. E. Broockman (2011). Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators. *American Journal of Political Science 55*(3), 463–477.

Coppock, A. (2018). Replication Data for: Avoiding Post-Treatment Bias in Audit Experiments. *Harvard Dataverse, v. 4.8.4. doi:10.7910/DVN/6NVI9C*.

Costa, M. (2017). How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials. *Journal of Experimental Political Science 4*(3), 241–254.

Fang, A. H., A. M. Guess, and M. Humphreys (2018). Can the Government Deter Discrimination? Evidence from a Randomized Intervention in New York City. *Journal of Politics*. In press.

Kalla, J., F. Rosenbluth, and D. L. Teele (2018). Are You My Mentor? A Field Experiment on Gender, Ethnicity, and Political Self-Starters. *The Journal of Politics*. In press.

Montgomery, J. M., B. Nyhan, and M. Torres (2018). How Conditioning on Post-treatment Variables Can Ruin Your Experiment and What to Do About It. *American Journal of Political Science*. In press.

Quillian, L., D. Pager, O. Hexel, and A. H. Midtbøen (2017). Meta-analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring Over Time. *Proceedings of the National Academy of Sciences 114*(41), 10870–10875.

Rondeau, D. and J. A. List (2008). Matching and Challenge Gifts to Charity: Evidence from Laboratory and Natural Field Experiments. *Experimental Economics 11*(3), 253–267.

White, A. R., N. L. Nathan, and J. K. Faller (2015). What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials. *American Political Science Review 109*(1), 129–142.

Zhang, J. L. and D. B. Rubin (2003). Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.