A Simple Approximation for Evaluating External Validity Bias[*]

Isaiah Andrews
Harvard and NBER

Emily Oster
Brown University and NBER

December 20, 2018

Abstract

External validity is a major challenge in treatment effect estimation. Even in randomized trials, the experimental sample often differs from the population of policy interest. If the sample differs in ways that are unobservable to researchers, correcting for differences in observables fails to fully eliminate bias. To help researchers think about external validity bias, we develop a simple approximation that relates the total bias to the bias from observables and a measure for role of treatment effect heterogeneity in driving selection into the experimental sample.
Key Words: External Validity, Randomized Trials
JEL Codes: C1

1 Introduction

External validity has drawn attention in economics with the growth of randomized trials. Randomized trials provide an unbiased estimate of the average treatment effect in the experimental sample, but this may differ from the average treatment effect in the population of interest. We will refer to such differences as "external validity bias."

[*] This paper was previously circulated under the title "Weighting for External Validity." Sophie Sun provided excellent research assistance. We thank Nick Bloom, Guido Imbens, Matthew Gentzkow, Peter Hull, Larry Katz, Ben Olken, Jesse Shapiro, Andrei Shleifer and seminar participants at Brown University, Harvard University and University of Connecticut for helpful comments. Andrews gratefully acknowledges support from the Silverman (1978) Family Career Development chair at MIT, and from the National Science Foundation under grant number 1654234.

One reason such differences may arise is because individuals (or other treatment units) actively select into the experiment. For example, Bloom, Liang, Roberts and Ying (2015) report results from an evaluation of working from home in a Chinese firm. Workers at the firm were asked to volunteer for the experiment, and the study randomized among eligible volunteers. External validity bias arises here if the effects of the treatment on the volunteers differ from effects in the overall population.

Papers that report experimental results often comment qualitatively on differences between the sample and some population of interest, and sometimes provide a table comparing means between groups (e.g. Bloom, Liang, Roberts and Ying (2015); Attanasio, Kugler and Meghir (2011); Muralidharan et al (2018)). Similar means are taken as reassuring.

This comparison of means does not fully address external validity concerns. First, external validity bias depends on *both* the difference in characteristics across groups and the extent to which treatment effects vary based on these characteristics. Second, this approach does not rule out differences in *unobservable* characteristics across groups.[1] To partially address the first concern, we can formally adjust for observed differences in covariates (e.g. Hotz et al, 2005, Stuart et al, 2011, Dehejia, Pop-Eleches and Samii, 2015). This is rarely done in practice, however, and does not address differences on unobserved dimensions.[2]

Our goal is to provide a framework in which to consider selection on unobservables when studying external validity. Under a simple model for selection and treatment effects, if the degree of selection is small, external validity bias is approximately equal to the bias due to selection on observables multiplied by a measure of the role of treatment effect heterogeneity in driving selection. We suggest that researchers report formal corrections for selection on observables and then use our result to benchmark how much selection on unobservables would be required to overturn their findings. This approach does not provide a definitive estimate of external validity bias, but offers a

[1] For example, individuals could sort directly on their treatment effects, which in an instrumental variables context Heckman et al (2006) describe as “essential heterogeneity.”

[2] A notable exception is Alcott (2015).

tractable language to frame the question.

2 A Simple Approximation to External Validity Bias

2.1 Setup

We are interested in the effect of a binary treatment $D_i \in \{0, 1\}$ on an outcome Y_i. Adopting the standard potential outcomes framework (see e.g. Imbens and Rubin, 2015) we write the outcomes of unit i in the untreated and treated states as $Y_i(0)$, $Y_i(1)$, respectively.[3]

We observe an iid sample from a randomized experiment in a trial population with distribution P_S, where treatment D_i is randomly assigned. The experiment allows us to unbiasedly estimate the average treatment effect in the trial population, $E_{P_S}[TE_i] = E_{P_S}[Y_i(1) - Y_i(0)]$.

The trial population is a potentially non-representative subset of a larger target population, and we are interested in inference on the average treatment effect in the target population. Let S_i be a dummy equal to one if individual i in the target population is a member of the trial population. For P the distribution in the target population, our object of interest is $E_P[TE_i]$, while our experiment estimates $E_{P_S}[TE_i] = E_P[TE_i|S_i = 1]$. The "external validity bias" is $E_{P_S}[TE_i] - E_P[TE_i]$.

We do not observe the distribution of all variables in the target population, and so cannot in general correct this bias. We assume, however that we know the target-population-mean for a set of covariates C_i, $E_P[C_i]$, where C_i is also observed in the trial population.[4]

2.2 A Simple Model

We next adopt a simple model for treatment effects and selection.

Assumption 1 *For a set of unobservables U_i,*

$$TE_i = \alpha + C_i'\gamma + U_i'\delta + \varepsilon_i$$

[3] We assume throughout that all random variables considered have bounded fourth moments.

[4] For example, C_i could contain demographic or geographic variables.

where $E_P[\varepsilon_i|C_i, U_i] = 0$ *and* $Cov\,(C_i, U_i) = 0$.

Without additional restrictions on the unobservables U_i, this assumption is without loss of generality. In practice, however, we will typically want to assume that U_i consists of particular known (but unobserved) variables, which makes this restriction substantive.

Assumption 2 *For the same set of unobservables* U_i*,*

$$S_i = 1\left\{C_i'\kappa + U_i'\tau - v_i \leq 0\right\}$$

where v_i *is independent of* $(C_i, U_i, \varepsilon_i)$ *under* P *and has support support equal to* $\mathbb{R}$*. The distribution function* F_v *of* v_i *is twice continuously differentiable with a bounded second derivative.*

Assumption 2 is equivalent to assuming that $E\left[S_i|C_i, U_i\right] = F_v\left(C_i'\kappa + U_i'\tau\right)$ where $0 < F_v\left(C_i'\kappa + U_i'\tau\right) < 1$ and $\frac{\partial^2}{\partial v^2}F_v\left(v\right)$ is bounded. This implies that all values (C_i, U_i) that arise in the target population also sometimes arise in the trial population.

2.3 Corrections for Selection on Observables

Assumptions 1 and 2 imply that

$$E_{P_S}\left[TE_i|C_i, U_i\right] = E_P\left[TE_i|C_i, U_i\right] = \alpha + C_i'\gamma + U_i'\delta, \tag{1}$$

so the conditional average treatment effect given covariates and unobservables is the same in the trial and target populations. The external validity bias is thus

$$E_{P_S}[TE_i] - E_P[TE_i] = \left(E_{P_S}\left[C_i\right] - E_P\left[C_i\right]\right)'\gamma + \left(E_{P_S}\left[U_i\right] - E_P\left[U_i\right]\right)'\delta. \tag{2}$$

Hence, the external validity bias depends on (a) the shift in the mean of (C_i, U_i) between the trial and target populations and (b) the importance of (C_i, U_i) for predicting treatment effects.[5]

[5] Many antecedents for (2) exist in the literature. See for example Nyugen et al (2017).

If $\delta = 0$, so the unobservables do not predict treatment effects, then external validity bias depends only on the difference in means for the covariates, $E_{P_S}[C_i] - E_P[C_i]$, and the coefficient γ. As discussed in the introduction the difference of means is sometimes reported, but the coefficient γ is rarely discussed. We can, however, estimate γ as the difference in coefficients $\hat{\gamma} = \hat{\gamma}_1 - \hat{\gamma}_0$ for $(\hat{\gamma}_0, \hat{\gamma}_1)$ calculated from the regression

$$Y_i = (1 - D_i)\,\alpha_0 + (1 - D_i)\,C_i'\gamma_0 + D_i\alpha_1 + D_iC_i'\gamma_1 + u_i$$

of Y_i on C_i in the treatment and control groups. If we assume that $\delta = 0$ we can easily estimate (and correct) external validity bias.

2.4 Small-Selection Approximation

If we do not assume $\delta = 0$, the external validity bias depends on terms we cannot estimate. To make progress we consider settings where the degree of selection is small, and in particular consider behavior as $(\kappa, \tau) \to 0$.[6] We then relate the external validity bias to the bias estimated by assuming that $\delta = 0$.

Let γ_S denote the probability limit of our estimate $\hat{\gamma}$ obtained from regression (1) in the trial population. The probability limit of our bias estimate based on observables is $(E_{P_S}[C_i] - E_P[C_i])'\,\gamma_S$. This estimated bias bears an intuitive relationship to the true bias when the degree of selection is small.

Proposition 1 *Under Assumptions 1 and 2, for $(\kappa, \tau) = \lambda \cdot (\tilde{\kappa}, \tilde{\tau})$ and $(\tilde{\kappa}, \tilde{\tau})$ fixed, as $\lambda \to 0$*

$$\frac{E_{P_S}[TE_i] - E_P[TE_i]}{(E_{P_S}[C_i] - E_P[C_i])'\,\gamma_S} - \frac{\gamma'\Sigma_C\kappa + \delta'\Sigma_U\tau}{\gamma'\Sigma_C\kappa} \to 0,$$

provided $\gamma'\Sigma_C\tilde{\kappa} \neq 0$, where $\Sigma_C = Var_P(C_i)$ and $\Sigma_U = Var_P(U_i)$.

This is our main result, and links the (estimable) selection-on-observables bias to the true external validity bias.

[6] One could alternatively make progress by imposing other assumptions. For example, Nyugen et al (2017) bound $(E_{P_S}[U_i] - E_P[U_i])'\,\delta$, while Gechter (2016) restricts the level of dependence between the individual outcomes in the treated and untreated states.

Validity of Approximation Proposition 1 discusses behavior as $(\kappa, \tau) \to 0$. This can be interpreted as an approximation result, and shows that

$$E_{P_S}[TE_i] - E_P[TE_i] \approx \frac{\gamma'\Sigma_C\kappa + \delta'\Sigma_U\tau}{\gamma'\Sigma_C\kappa}\left(E_{P_S}[C_i] - E_P[C_i]\right)'\gamma_S$$

in the sense that the difference is of lower order for (κ, τ) small. Since in practice we are interested in settings with a nonzero degree of selection, it is reasonable to ask when this approximation will be reliable. The proof of Proposition 1 in Appendix A proceeds by (i) taking a first-order Taylor approximation of $F_v\left(C_i'\kappa + U_i'\tau\right)$ and (ii) approximating γ_S by γ. We expect that the result of Proposition 1 will provide a reasonable approximation so long as (a) $F_v\left(C_i'\kappa + U_i'\tau\right)$ is not overly nonlinear over the region containing most realizations of (C_i, U_i) and (b) γ_S is close to γ.

Interpretation: The key unknown term in Proposition 1 is the selection ratio

$$\Psi = \frac{\gamma'\Sigma_C\kappa + \delta'\Sigma_U\tau}{\gamma'\Sigma_C\kappa}.$$

This ratio measures the relative importance of treatment effect heterogeneity in explaining the observed and unobserved drivers of selection. In particular,

$$\Psi = 1 + \frac{Cov_P\left(TE_i, \tau'U_i\right)}{Cov_P\left(TE_i, \kappa'C_i\right)},$$

where we can interpret $\kappa'C_i$ and $\tau'U_i$ as the observed and unobserved drivers of selection.

To develop intuition, we consider four special cases.

Special Case 1 $\delta = 0$: unobservables are unrelated to the treatment effects, so $\Psi = 1$ and the correction for observable differences discussed above is valid.

Special Case 2 $\tau = 0$: unobservables may predict treatment effects but play no role in selection. We again have $\Psi = 1$, so the correction for observable differences is (approximately) valid.

Special Case 2 $\delta \neq 0$, $\tau \neq 0$, but $\delta'\Sigma_U\tau = 0$: unobservables predict both treatment effects and selection, but the unobserved drivers of selection and treatment effects are unrelated. Hence, $\Psi = 1$ and the correction for observable differences is (approximately) valid.

Special Case 4 $(\gamma, \delta) \propto (\kappa, \tau)$: the same combinations of observables and unobservables matter for both selection and treatment effects. Hence,

$$\Psi = \frac{R^2_{C,U}}{R^2_C}$$

for R^2_X the R-squared from the regression of TE_i on X_i. Here, Ψ can be interpreted as the proportional increase in R^2 from including the unobservables U_i in an (infeasible) regression of TE_i on covariates. This implies that $\Psi \geq 1$, so the correction for observable differences is a lower bound on the true bias.

The fourth special case and the general case are likely to be of the most interest, since they do not assume away selection on unobservables. The result in the fourth special case delivers sharper conclusions, since we get a lower bound on the external validity bias, but the result in the general case is more widely applicable.

3 Illustrative Application

To illustrate, we apply our results to data from Bloom et al (2015). Workers at a Chinese call center were given an opportunity to volunteer for a work-from-home program. Approximately 50% volunteered, and treatment was randomized among eligible volunteers. The results suggest substantial productivity gains from working from home.

A follow-up question is whether it would be productivity-enhancing to have many or all eligible call center employees work from home. If the ATE estimated in the experiment is valid for the entire workforce, the answer is likely yes. Given the sample construction, however, it seems plausible that the ATE for the experimental sample is not representative of the whole population.

Target Population Data The natural target population is the set of all eligible workers. Bloom et al (2015) collect some basic characteristics for this population, which are compared to characteristics of the volunteers in Table 1. There are some differences: the volunteers have longer commutes, are more likely to be male, and are more likely to have children.

Correcting for Observables We first correct for selection on observables. We estimate γ as the difference $\hat{\gamma} = \hat{\gamma}_1 - \hat{\gamma}_0$ in coefficients from the regression (1), and estimate the selection on observables bias as $\left(\widehat{E}_{P_S}\left[C_i\right] - \widehat{E}_P\left[C_i\right]\right)' \hat{\gamma}_S$, where we use $\widehat{E}$ to denote the sample average.[7] Results are reported in Table 2, which shows that correcting for observable differences slightly *increases* the estimated effect, from 0.271 to 0.289.

Accounting for Unobservables We next consider the scope for further bias due to selection on unobservables. We bound the target population average treatment effect under the assumption that $\Psi \in [1, 2]$, so bias due to unobservables operates in the same direction as, and is no larger than, bias due to observables. Estimates are reported in column three of Table 2. These are similar to the baseline results.

We then ask what value $\Psi(0)$ of the selection ratio Ψ would yield an average treatment effect of zero in the target population. This value is equal -14.7, so the bias from unobservables would have to be much larger than the estimated bias from observables, and operate in the opposite direction, in order to overturn the main result.

Both approaches suggest that the results of Bloom et al (2015) are robust to a wide range of assumptions about the role of unobservables.

[7] We take C_i to include all of the variables reported in Table 1 and, for non-binary variables, their squares.

Table 1: **Observable Characteristics, Bloom et al (2015)**

Variable	Population: Mean (SD)	Sample: Mean (SD)
Age	24.4 (3.30)	24.7 (3.65)
Gross Wage	3.13 (0.84)	3.09 (0.78)
Any Children	0.155 (0.362)	0.201 (0.402)
Married	0.265 (0.442)	0.310 (0.463)
Male	0.385 (0.487)	0.438 (0.497)
At Least Tertiary Educ	0.456 (0.498)	0.399 (.490)
Commute Time (Min)	96.9 (61.1)	111.7 (62.7)
Job Tenure	32.4 (19.7)	31.2 (20.6)

Notes: This table reports moments for the sample and target population in Bloom et al (2015).

Table 2: **Application: Bloom et al (2015)**

Outcome	Baseline Effect	Observable Adjusted	Bounds, $\Psi \in [1, 2]$	$\Psi(0)$
Job Performance	0.271	0.289	[0.289, 0.309]	-14.7
	(0.22, 0.32)	(0.23,0.34)		

Notes: Bootstrapped 95% confidence intervals are reported below the baseline and observables-adjusted estimates. One can also calculate confidence sets for the last two columns, but for brevity we do not explore this possibility here.

4 Conclusion

This paper considers the problem of external validity, and derives an approximation which relates the total external validity bias to the bias from observables.

Our application to Bloom et al (2015) is representative of a broader class of applications in which participants select in to a study (e.g. Attanasio, Kugler and Meghir (2011), Gelber, Ibsen and Kessler (2016), Muralidharan et al (2018)). Our approach applies more broadly, however, including to settings where researchers select a set of areas or treatment units for their experiments (i.e. Muralidharan and Sundararaman (2011); Olken et al (2014); Alcott (2015)).[8] The only additional data requirement to implement our approach is knowledge of some characteristics of the target population. In many cases one could use demographic variables, where moments in the target population

[8] Note that when the selection occurs at a different level than treatment, S_i will not be iid across units i but our results continue to apply provided we define C_i and U_i to vary at same level as the selection decision.

may be available from public datasets.

Appendix A: Proof of Proposition 1

Note that under Assumptions 1 and 2,

$$E_{P_S}\left[TE_i\right]-E_P\left[TE_i\right]=E_P\left[\left(W_i-1\right)TE_i\right]=Cov_P\left(W_i,TE_i\right)$$

for $W_i=\frac{F_v\left(C_i'\kappa+U_i'\tau\right)}{E_P\left[F_v\left(C_i'\kappa+U_i'\tau\right)\right]}$.

By the mean value theorem $F_v\left(C_i'\kappa+U_i'\tau\right)=F\left(0\right)+f_v\left(v_i^*\right)\left(C_i'\kappa+U_i'\tau\right)$ for $f_v\left(\cdot\right)$ the density of v_i and v_i^* an intermediate value. Since $f_v\left(\cdot\right)$ is continuously differentiable with a bounded derivative it is Lipschitz, and $\left|f_v\left(v\right)-f_v\left(0\right)\right|\leq Kv$ for some constant K and all v. As a result, $\left|F_v\left(0\right)+f_v\left(0\right)\left(C_i'\kappa+U_i'\tau\right)-F_v\left(C_i'\kappa+U_i'\tau\right)\right|\leq K\cdot\left(C_i'\kappa+U_i'\tau\right)^2$. Hence, for $\left(\kappa,\tau\right)=\lambda\cdot\left(\tilde{\kappa},\tilde{\tau}\right)$, $Cov_P\left(W_i,TE_i\right)$ is equal to

$$Cov_P\left(\frac{F_v\left(0\right)+f_v\left(0\right)\left(C_i'\kappa+U_i'\tau\right)}{E_P\left[F_v\left(0\right)+f_v\left(0\right)\left(C_i'\kappa+U_i'\tau\right)\right]},TE_i\right)+O\left(\lambda^2\right)=\lambda c_1Cov_P\left(C_i'\kappa+U_i'\tau,TE_i\right)+O\left(\lambda^2\right)$$

where $c_1=f_v\left(0\right)/E_P\left[F_v\left(0\right)\right]\neq0$. By Assumption 1, $Cov_P\left(C_i'\kappa+U_i'\tau,TE_i\right)=\gamma'\Sigma_C\kappa+\delta'\Sigma_U\tau$, so $Cov_P\left(W_i,TE_i\right)=\lambda c_1\left(\gamma'\Sigma_C\tilde{\kappa}+\delta'\Sigma_U\tilde{\tau}\right)+O\left(\lambda^2\right)$. By the same argument

$$E_{P_S}\left[C_i\right]-E_P\left[C_i\right]=Cov_P\left(W_i,C_i\right)=\lambda c_1Cov_P\left(C_i'\kappa+U_i'\tau,C_i\right)=\lambda c_1\tilde{\kappa}'\Sigma_C+O\left(\lambda^2\right),$$

and

$$\frac{E_{P_S}\left[TE_i\right]-E_P\left[TE_i\right]}{\left(E_{P_S}\left[C_i\right]-E_P\left[C_i\right]\right)'\gamma}=\frac{\gamma'\Sigma_C\tilde{\kappa}+\delta'\Sigma_U\tilde{\tau}}{\gamma'\Sigma_C\tilde{\kappa}}+O\left(\lambda^2\right),$$

where the denominator on the right hand side is nonzero by assumption.

This nearly completes the proof, except that the proposition replaces the γ on the left hand side by γ_S. Note, however, that random assignment implies $\gamma_S=Var_{P_S}\left(C_i\right)^{-1}Cov_{P_S}\left(C_i,TE_i\right)$. By arguments along the same lines as above, $Var_{P_S}\left(C_i\right)\to Var_P\left(C_i\right)$ and $Cov_{P_S}\left(C_i,TE_i\right)\to Cov_P\left(C_i,TE_i\right)$ as $\lambda\to0$. Hence

$$\left(E_{P_S}\left[C_i\right]-E_P\left[C_i\right]\right)'\left(\gamma_S-\gamma\right)=o\left(\lambda\right),$$

from which the result follows immediately. □

References

Allcott, Hunt, "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

Attanasio, Orazio, Adriana Kugler, and Costas Meghir, "Subsidizing Vocational Training for Disadvantaged Youth in Colombia: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, July 2011, *3* (3), 188–220.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying, "Does Working From Home Work? Evidence From A Chinese Experiment," *The Quarterly Journal of Economics*, 2015, *165*, 218.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii, "From Local to Global: External Validity in a Fertility Natural Experiment," Working Paper 21459, National Bureau of Economic Research August 2015.

Gechter, Michael, "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India," *manuscript, Pennsylvania State University*, 2016.

Gelber, Alexander, Adam Isen, and Judd B Kessler, "The Effects of Youth Employment: Evidence from New York City Lotteries," *The Quarterly Journal of Economics*, 2016, *131* (1), 423–460.

Heckman, James, Sergio Urzua, and Edward Vytlacil, "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 2006, *88* (3), 389–432.

Hotz, Joseph, Guido W. Imbens, and Julie H. Mortimer, "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, 2005, *125* (1-2), 241–270.

Imbens, Guido and Don Rubin, *Causal Inference for Statistics, Social Science and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press, 2015.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro Ganimian, "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," 2018. Working Paper.

____ **and Venkatesh Sundararaman**, "Teacher Performance Pay: Experimental Evidence from India," *The Journal of Political Economy*, 2011, *119* (1), 39–77.

Nguyen, Trang Quynh, Cyrus Ebnesajjad, Stephen R. Cole, and Elizabeth A. Stuart, "Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects," *The Annals of Applied Statistics*, 2017.

Olken, Benjamin A, Junko Onishi, and Susan Wong, "Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia," *American Economic Journal: Applied Economics*, 2014, *6* (4), 1–34.

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf, "The use of propensity scores to assess the generalizability of results from randomized trials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2011, *174* (2), 369–386.