NBER WORKING PAPER SERIES

WEIGHTING FOR EXTERNAL VALIDITY

Isaiah Andrews
Emily Oster

Working Paper 23826
http://www.nber.org/papers/w23826

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2017

Sophie Sun provided excellent research assistance. We thank Nick Bloom, Guido Imbens, Matthew Gentzkow, Larry Katz, Ben Olken, Jesse Shapiro, Andrei Shleifer and participants in seminars 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2017 by Isaiah Andrews and Emily Oster. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Weighting for External Validity
Isaiah Andrews and Emily Oster
NBER Working Paper No. 23826
September 2017
JEL No. C1,C18,C21

ABSTRACT

External validity is a fundamental challenge in treatment effect estimation. Even when researchers credibly identify average treatment effects – for example through randomized experiments – the results may not extrapolate to the population of interest for a given policy question. If the population and sample differ only in the distribution of observed variables this problem has a well-known solution: reweight the sample to match the population. In many cases, however, the population and sample differ along dimensions unobserved by the researcher. We provide a tractable framework for thinking about external validity in such cases. Our approach relies on the fact that when the sample is drawn from the same support as the population of interest there exist weights which, if known, would allow us to reweight the sample to match the population. These weights are larger in a stochastic sense when the sample is more selected, and their correlation with a given variable reflects the intensity of selection along this dimension. We suggest natural benchmarks for assessing external validity, discuss implementation, and apply our results to data from several recent experiments.

Isaiah Andrews
Department of Economics, E52-504
MIT
77 Massachusetts Avenue
Cambridge MA 02139
and NBER
iandrews@mit.edu

Emily Oster
Brown University
Department of Economics
64 Waterman Street
Providence, RI 02912
and NBER
emily_oster@brown.edu

1 Introduction

External validity is a major challenge in empirical social science. Modern identification strategies allow researchers to identify causal effects for well-defined subpopulations. In many contexts, however, the population of policy interest differs from that for which we have credible identification. In the context of instrumental variables, this tension is reflected in the LATE critique (Imbens and Angrist, 1994), which highlights that IV methods uncover causal effects only for those individuals whose behavior is changed by the instrument. In many randomized experiments a similar concern arises due to selection of the experimental sample, even if compliance with the randomly assigned treatment is perfect.

As an example, consider Bloom, Liang, Roberts and Ying (2015), who report results from an experimental evaluation of working from home in a Chinese firm. In the first stage of the evaluation, workers at the firm were asked to volunteer for an experiment in which they might have the opportunity to work from home. The study then randomized among eligible volunteers, and compliance was excellent. The study estimates large productivity gains from working from home. Given these results, one might reasonably ask whether the firm would be better off having more of their employees work from home - or even having them all work from home. To answer this question, we need to know the average treatment effect of working from home in the entire population of workers.

The population of volunteers for the experiment differs from the overall population of workers along some observable dimensions (for example, commute time and gender). It seems plausible that they also differ on some unobservable dimensions, for example ability to self-motivate, and thus that they may have different treatment effects. To the extent volunteers have systematically different treatment effects than non-volunteers, the average treatment effect estimated by the experiment will differ from that in the population of workers as a whole. This issue - that the experimental sample differs from the population of policy interest - is widespread in economics and other fields.[1]

If selection is driven entirely by observable variables, then one can reweight the sample to

[1] In medicine, for example, the efficacy of drugs is tested on study participants who may differ systematically from the general population of possible users. See Stuart, Cole, Bradshaw and Leaf (2011) for discussion. Within economics, Allcott (2015) shows that OPower treatment effects are larger in early sites than later sites, and adjustments for selection on observables do not close this gap.

obtain population-appropriate estimates (as in e.g. Stuart et al, 2011). However, when there are concerns about selection on unobservable factors, adjusting for selection on observables may be insufficient.

In this paper we provide a tractable framework for thinking about external validity of treatment effects estimated from a selected sample. Our approach rests on the idea of reweighting the sample to match the population. When the sample is selected on variables unobserved to the researcher, the resulting weights cannot be directly calculated. Nonetheless, our framework provides an intuitive way to bound the plausible range of selection bias.

The paper contains two key results. First, we develop a general representation result which provides intuition about the magnitude of the external validity bias under sample selection on either observed or unobserved variables. Second, we link the bias from selection on unobserved variables to the bias from selection on observed variables. Using this result, we show that a formal adjustment for selection on observables provides a benchmark for selection on unobservables under intuitive assumptions.

We begin in Section 2 with our theoretical framework. We assume that we observe a random sample from some population, which we label the trial population, and are interested in the mean of some function of the data in a target population. We call this function the target function, and call its mean in the target population the target moment. Later, we show that a suitable choice of target function yields the average treatment effect (ATE) in experimental settings. Under regularity conditions, including that the trial and target populations are drawn from the same support, we can reweight the trial population to match the target population. If the trial population is selected on unobserved variables, however, the necessary weights are unknown.

Our first result shows that the bias in the sample average of the target function, as an estimator for the target moment, is the product of three terms: (1) the standard deviation of the target function, (2) the correlation between the weights and the target function, and (3) the standard deviation of the weights. These measure, respectively, the variability of the target function, the intensity of selection on the target function, and the overall degree of sample selection. In the context of ATE estimation, this highlights that bias is larger when (1) there is substantial treatment effect heterogeneity, (2) individual-level treatment effects are highly correlated with selection into the sample, and (3) there is a lot of selection into the sample. We illustrate using data from the National Longitudinal Survey of Youth (NLSY), a setting in which we observe the true weights to re-balance the sample to match the US population. In this example, we take the target moments to be the means of several variables in the US population.

This decomposition is not directly useable in practice, since both the correlation between the weights and the target function and the standard deviation of the weights are unknown. To eliminate dependence on the standard deviation of the weights, we consider the ratio of bias in the target moment to bias in some benchmark moment whose value in the target population is known. The only unknown term affecting this ratio is the intensity of selection on the target moment relative to the benchmark moment. For example, if there are demographic variables which we observe in both the trial and target populations, we could use selection on these variables to benchmark selection on the target moment.

We again illustrate in the NLSY data, showing that we can compare bias across moments. For example, we compare the degree of selection on college graduation to the degree of selection on high school graduation or gender. Such comparisons highlight that selection on college graduation is similar to that on high school graduation, but much more intense than that on gender. While the ways in which the NLSY is selected are well-understood, these results show that our approach recovers sensible answers in this setting.

In the second part of the paper we turn to the problem of ATE estimation. In the treatment effect setting, a natural benchmark is adjustment for selection on observables (e.g. Hellerstein and Imbens, 1999; Hotz, Imbens and Mortimer, 2005). In Section 3 we show how the selection on observables adjustment relates to the overall bias in the trial population ATE under two models of selection.

We show, first, that under the assumption of selection on the treatment effect - that is, a model in which individuals are more likely to be in the sample if they have a higher (or lower) treatment effect - the ratio of the total bias to the selection on observables adjustment can be interpreted as measuring the degree of private information about the treatment effect. We denote this by Φ. A value of $\Phi = 2$, for example, indicates that there is as much private information as there is observed information about the treatment effect.

Second, we show more generally that under the assumption of selection on some variable, the ratio of total bias to selection on observables bias can be interpreted as measuring the relative importance of observables and unobservables in explaining the covariance between the selection covariate and the treatment effect. We denote the ratio in this case by Φ_A; it has a similar interpretation in terms of private information. We note that in a model with a large number of variables, a random subset of which are observed, Φ and Φ_A both have an interpretation as the ratio of total to observed variables. In both cases, these results link an intuitively appealing object - the selection on observables adjustment - to the bias from unobservable selection. Establishing this link, among other things, highlights the value of a formal adjustment for selection on observables.

In Section 4 of the paper we discuss implementation. We consider a general setting in which a researcher estimates an ATE in a trial population and is interested in the plausible values of this ATE in the target population. We propose two approaches to this problem. The first specifies a value of the target population ATE of particular relevance (e.g. zero average effect) and calculates the Φ or Φ_A required to yield this value. A larger Φ or Φ_A indicates a more robust result. In the second approach, the researcher specifies bounds on Φ and Φ_A and calculates the implied bounds on the target population ATE. Tighter bounds indicate a more robust result. In both cases, the first step is to adjust for selection on observables. We illustrate in a constructed example which provides intuition for the interpretation of Φ and Φ_A.

In Section 5 we turn to applications. We first briefly discuss the range of applications for which this approach is likely to apply, focusing in particular on the role of the common support assumption empirically. We apply our results to data drawn from Attanasio, Kugler and Meghir (2011), Bloom et al (2015), Dupas and Robinson (2013), and Olken, Onishi, and Wong (2014).

In both Attanasio et al (2011) and Bloom et al (2015) we suppose that individuals select into the experiment - studying a job training program in the first case and working from home in the second - based on their expected treatment effect. We correct for selection on observables and then ask how much private information would be necessary to overturn the result. In Attanasio et al (2011), correcting for selection on observables considerably attenuates the effect, while in Bloom et al (2015) this attenuation is much smaller. As a result, the effects in the latter paper appear more robust to concerns about selection.

In Dupas and Robinson (2013), which estimates the effect of a variety of savings treatments on multiple health outcomes, our approach reveals that some treatment-outcome pairs are much more robust than others. Finally, Olken et al (2014) report results from an experiment in Indonesia which shows very little impact on the outcomes they consider. We find, however, that adjusting for selection on observables - and then further allowing for comparable selection on unobservables - dramatically increases the results. This suggests that the small effects estimated in this paper may reflect the choice of experimental population, rather than ineffectiveness of the treatment.

A key input to these applications is a formal correction for selection on observables. Although such corrections are well known in the econometrics and statistics literatures, they are not commonly used in applied work. By linking the formal selection on observables adjustment to the overall bias allowing for selection on unobservables, we provide an additional argument for correcting these observable differences formally, rather than simply discussing the differences between trial and target population informally.

This paper contributes to the literature on external validity of ATE estimates. We relate closely to the large literature studying selection on observables (e.g. Hellerstein and Imbens, 1999; Hotz et al, 2005; Cole and Stuart, 2010; Stuart et al, 2011; Imai and Ratkovic, 2014; Dehejia, Pop-Eleches and Samii, 2015; Hartman, Grieve, Ramsahai and Sekhon, 2015), as well as to the literature on propensity score reweighting (e.g. Hahn 1998 and Hirano, Imbens and Ridder 2003). Both Alcott (2015) and Chyn (2016) highlight the issue of selection on unobservables, although they do not provide a method to address it. Gechter (2015) considers the same problem we do, and suggests bounds which result from assumptions on the level of dependence between the individual outcomes in the treated and untreated states. This is a different object than that bounded by our approach, and his results are complementary to ours.

We also relate, more distantly, to the recent literature on external validity in instrumental variables (Feller et al, 2016; Kowalski, 2016; Kline and Walters, 2016*;* Brinch*,* Mogstad and Wiswall*,* 2017) and regression discontinuity (Bertanha and Imbens, 2014; Angrist and Rokkanen, 2015; Rokkanen, 2015) settings. These connections are discussed in Section 6.

While we consider a different problem, our approach is conceptually similar to that of Rosenbaum (2002, Chapter 4), Altonji, Elder and Taber (2005), and Oster (*forthcoming*). Like these authors we ask how intense selection on some unobservable dimension would have to be, relative to selection on observable dimensions, to overturn a given result. Since the true value for the relative intensity of selection is not identified from the available data, precisely identifying the effect of interest is impossible without additional assumptions. Our goal is to give researchers a transparent, interpretable framework through which to consider the range of plausible assumptions and the implications for their results.

2 Sample Selection and Reweighting

To develop our framework, suppose that we have a sample of observations X_i from the trial population. We denote distributions in the trial and target populations by P_S and P, respectively, and assume that we are interested in the mean of a target function $t\left(X_i\right)$ of X_i in the target population, $E_P\left[t\left(X_i\right)\right]$. We will call this the target moment. By contrast, the sample mean of the target function estimates $E_{P_S}\left[t\left(X_i\right)\right]$. For simplicity we assume an infinite sample in developing our theoretical results, so the distribution of X_i under P_S is known. Results on inference, which account for sampling uncertainty, are developed in Section 4.1 below.

We assume that the support of X_i in the target population is a subset of its support in the trial population and, more restrictively, that the distribution in the target population is absolutely continuous with respect to that in the trial population. We maintain this assumption for the remainder of the paper.

Assumption 1 *The distribution P_X of X_i under P is absolutely continuous with respect to the distribution $P_{X,S}$ of X_i under P_S.*

Absolute continuity requires that for any set A, $Pr_{P_S}\left\{X_i \in A\right\} = 0$ implies $Pr_P\left\{X_i \in A\right\} = 0$, and thus that all events which occur with zero probability in the trial population likewise occur with zero probability in the target population. This is a strong assumption, but is implied by probabilistic assignment as assumed in the literature on treatment effect estimation (see, for example, Definition 3.5 in Imbens and Rubin 2015). Nonetheless, Assumption 1 will fail if there are some values of X_i in the target population which are never observed in the trial population. If this occurs, the reweighting approach developed in this paper no longer applies. Even in this case, under limited deviations from absolute continuity one could build on our results to derive bounds, though we do not pursue this possibility.

Leading Case: Trial Population a Subset of Target Population In many contexts the trial population is a subset of the target population. To discuss this case formally, define a variable S_i in the target population which indicates whether individual i is also part of the trial population

$$S_i = \begin{cases} 1 & \text{if } i \text{ is part of the trial population} \\ 0 & \text{otherwise.} \end{cases}$$

If the distribution P_X has density $p_X(x)$ then we can write the density in the trial population in terms of $p_X(x)$ and the distribution of S_i.[2]

Lemma 1 *Let P_X have density $p_X(x)$. If $E_P[S_i] > 0$ then $P_{X,S}$ is absolutely continuous with respect to P_X and the density of $P_{X,S}$ is*

$$p_{X,S}(x) = \frac{E_P[S_i | X_i = x]}{E_P[S_i]} p_X(x). \tag{1}$$

If we assume that S_i is independent of X_i, this result implies that $P_{X,S} = P_X$ and thus that the distributions in the trial and target populations are the same. Consequently, $E_{P_S}[t(X_i)] = E_P[t(X_i)]$, and there is no extrapolation problem. Thus, external validity issues in this setting arise directly from X-dependent selection into the trial population.

Illustration: To develop intuition, we begin by illustrating the selection problem in a dataset commonly used in economics: the National Longitudinal Survey of Youth 1979 (NLSY-79). The NLSY-79 is a longitudinal panel which began with youth aged 14 to 21 in 1979 and has continued to the present. At each round data is collected on education, labor market experiences, health, and other variables.

The NLSY oversampled certain groups (e.g. African-Americans). Due to this sampling scheme, moments of these data (for example, means of variables) will not be unbiased for those moments in the full population. We use the NLSY as illustration since in these data we observe the true sampling weights. These weights allow us to reweight the NLSY to obtain a representative sample from the US population.

In our terminology, we define the NLSY sample as our trial population, and the reweighted

[2] We define all densities with respect to a fixed base measure μ. μ need not be Lebesgue measure, so our results do not require that X_i be continuously distributed.

representative sample as our target population. The availability of the weights makes it possible to explicitly illustrate the reweighting calculations we develop below. In thinking about settings where we do not observe the true weights, we can use these data with the weights excluded to consider how to learn about the target population. In this example we take the target moments to be the means of variables in the data, and defer discussion of treatment effect estimation to Section 3. △

2.1 Reweighting Algebra

When the trial and target populations differ, Assumption 1 implies that we can reweight the trial population to match the target population.

Lemma 2 *Under Assumption 1, for $W_i = \frac{p_X(X_i)}{p_{X,S}(X_i)}$ and any function $f(\cdot)$,*

$$E_P[f(X_i)] = E_{P_S}[W_i f(X_i)]. \tag{2}$$

Lemma 2 is a well-known result (see e.g. Horvitz and Thompson, 1952), and shows that we can recover expectations under P by reweighting our observations from P_S using the weights $W_i = \frac{p_X(X_i)}{p_{X,S}(X_i)}$, which compare the densities of the trial and target populations at each X_i. If we knew these weights we could unbiasedly estimate the target moment $E_P[t(X_i)]$ by the sample mean of $W_i t(X_i)$. Since we have assumed that we know $P_{X,S}$, however, knowledge of the weights W_i is equivalent to knowledge of P_X. Absent perfect knowledge of the distribution of X_i in the target population, these weights are thus infeasible.

While unknown, the weights W_i provide a useful lens through which to consider sample selection. These weights are non-negative by construction, and taking $f(\cdot) = 1$ in Lemma 2 confirms that $E_{P_S}[W_i] = 1$. An immediate corollary of Lemma 2 allows us to characterize the bias of the sample mean of $f(X_i)$ as an estimator for $E_P[f(X_i)]$.

Corollary 1 *For any function $f(\cdot)$, under Assumption 1 we have*

$$E_{P_S}[f(X_i)] - E_P[f(X_i)] = -Cov_{P_S}(W_i, f(X_i)).$$

If we further assume that $E_{P_S}\left[f(X_i)^2\right]$ *and* $E_{P_S}\left[W_i^2\right]$ *are finite, then*

$$E_{P_S}\left[f(X_i)\right] - E_P\left[f(X_i)\right] = -\sigma_{P_S}(W_i)\,\rho_{P_S}(W_i, f(X_i))\,\sigma_{P_S}(f(X_i))\,, \tag{3}$$

for $\sigma_{P_S}(A_i)$ *and* $\rho_{P_S}(A_i, B_i)$ *the standard deviation of* A_i *and the correlation of* A_i *and* B_i *under* P_S*, respectively.*

The final term in equation (3), $\sigma_{P_S}(f(X_i))$, measures the standard deviation of $f(X_i)$ in the trial population and can be estimated from the data. The correlation $\rho_{P_S}(W_i, f(X_i))$ measures the strength of the relationship between the weights and $f(X_i)$, and can loosely be viewed as measuring the extent to which selection loads on $f(X_i)$. By the definition of the correlation coefficient this quantity is smaller than one in absolute value. Lastly, the standard deviation $\sigma_{P_S}(W_i)$ can be viewed as measuring the extent of selection on any dimension, since the bounds on $\rho_{P_S}(W_i, f(X_i))$ imply that for all functions $f(\cdot)$,

$$\left|E_P\left[f(X_i)\right] - E_{P_S}\left[f(X_i)\right]\right| \leq \sigma_{P_S}(W_i)\,\sigma_{P_S}(f(X_i))\,. \tag{4}$$

Thus, the mean of $f(X_i)$ in the target population can differ from its mean in the trial population by at most $\sigma_{P_S}(W_i)$ times its standard deviation.

One can make a loose analogy between the decomposition in equation (3) and the omitted variables bias formula in linear regression. In regression, the bias in the coefficients for the included variables is the product of the coefficient on the omitted variable with the regression coefficient of the omitted variable on the included variables. Thus, what matters for bias is not only the importance of the omitted variable but also the strength of its relationship with the included variables. Analogously, for the external validity bias we study it matters not only how strongly the trial population is selected (measured by $\sigma_{P_S}(W_i)$) but also how tightly selection is related to the variable of interest (measured by $\rho_{P_S}(W_i, f(X_i))$).

The same decomposition applies to any collection of moments. In particular, suppose we are interested in the mean of a vector of functions $f_1(X_i), f_2(X_i), ..., f_q(X_i)$ in the target population. Applying Corollary 1 to each element, we obtain

$$\begin{aligned}E_{P_S}\left[f_1\left(X_i\right)\right]-E_P\left[f_1\left(X_i\right)\right]&=-\sigma_{P_S}\left(W_i\right)\rho_{P_S}\left(W_i,f_1\left(X_i\right)\right)\sigma_{P_S}\left(f_1\left(X_i\right)\right)\\E_{P_S}\left[f_2\left(X_i\right)\right]-E_P\left[f_2\left(X_i\right)\right]&=-\sigma_{P_S}\left(W_i\right)\rho_{P_S}\left(W_i,f_2\left(X_i\right)\right)\sigma_{P_S}\left(f_2\left(X_i\right)\right)\\&\vdots\\E_{P_S}\left[f_q\left(X_i\right)\right]-E_P\left[f_q\left(X_i\right)\right]&=-\sigma_{P_S}\left(W_i\right)\rho_{P_S}\left(W_i,f_q\left(X_i\right)\right)\sigma_{P_S}\left(f_q\left(X_i\right)\right).\end{aligned}\tag{5}$$

A key feature of this decomposition is that the standard deviation of the weights, $\sigma_{P_S}(W_i)$ appears in all rows. This again reflects the fact that $\sigma_{P_S}\left(W_i\right)$ measures the degree of sample selection along any dimension.

Illustration (continued): In the NLSY we observe the weights W_i. Therefore, we can calculate all terms in the decomposition (5). In particular, we consider this decomposition when taking $f\left(X_i\right)$ to measure race (share white), high school completion, and gender (share male).

The first two columns of Table 1 report the trial and target population means for these variables in the NLSY. The final three columns show the elements of the bias decomposition in equation (3). The difference in means for each variable is the product of these three elements. The differences between trial and target population means reflect the sampling structure: the NLSY over-samples racial minority groups and individuals from lower socioeconomic status backgrounds. There are fewer whites and fewer high school graduates in the sample than in the overall US population. The bias is largest for race, which reflects the very high correlation between the sampling weights and race.

In contrast to race and education, there is little bias in the gender variable since the sample is not selected on gender. This lack of selection is reflected in the very small correlation between this variable and the weights. As noted above the standard deviation of weights is the same in all rows, since this is a measure of selection on *any* dimension. △

Even without further restrictions, the decomposition (3) provides a guide to intuition. In particular, the bias in the sample mean of a given function of the data is larger when (a) the sample is more heavily selected in general, (b) sample selection is more heavily weighted toward the function in question, and (c) there is more variability in the function.

2.2 Relative Selection

The central question of this paper is how we can use data from the trial population to draw conclusions about the target moment. That is, we assume that we observe $E_{P_S}[t(X_i)]$ and would like to know $E_P[t(X_i)]$. Corollary 1 shows that this is equivalent to knowing the covariance between the weights W_i and $t(X_i)$.

In most applications, we know some characteristics of the target population. In particular, suppose we know the target-population mean of some benchmark function, $E_P[b(X_i)]$. In such cases, it is natural to consider the relative selection ratio

$$\frac{E_P[t(X_i)] - E_{P_S}[t(X_i)]}{E_P[b(X_i)] - E_{P_S}[b(X_i)]} = \frac{Cov_{P_S}(W_i, t(X_i))}{Cov_{P_S}(W_i, b(X_i))} = \frac{\rho_{P_S}(W_i, t(X_i))\,\sigma_{P_S}(t(X_i))}{\rho_{P_S}(W_i, b(X_i))\,\sigma_{P_S}(b(X_i))} \qquad (6)$$

which compares the bias in the target moment to that in the benchmark moment, where we have used Corollary 1 to express this bias ratio as a ratio of covariances. When we take this ratio, the overall degree of selection $\sigma_{P_S}(W_i)$ drops out, so the only unknown term in the right-hand side of this expression is the correlation ratio $\rho_{P_S}(W_i, t(X_i))/\rho_{P_S}(W_i, b(X_i))$, which describes the intensity of selection on the target moment $t(X_i)$ relative to the benchmark moment $b(X_i)$. Thus, considering the relative selection ratio in equation (6) allows us to abstract from the overall degree of selection and instead focus on the intensity of selection on the target moment relative to the benchmark.

Illustration (continued): We next introduce three new variables - log wage, college completion and AFQT score (a measure of IQ available in the NLSY) - and explore the degree of selection on these variables relative to various benchmarks. The first two columns of Table 2 show summary information for these variables (their trial and target population means).

The remaining columns of Table 2 show the values of the relative selection ratio for varying benchmark functions; we consider means of the variables used in Table 1 as the benchmarks. For example, the degree of selection on high school completion is very similar to that on college completion - the entry in the second row, fifth column is 1.15. The selection on race is much larger than the selection on any of the additional variables, so all of the values in Column 4 are small. On the other hand, the selection on each of these variables is much greater than the selection on gender, so the values in Column 6 are very large. We can read these as saying that in order to correctly predict the target population mean of the additional variables, we would have to assume that selection on these variables is much more intense than selection on gender. △

This illustration highlights that a natural way to learn about the degree of selection is to use moments which are observed in both trial and target population. In particular, if we had a new moment which was known in the trial population and unknown in the target population, we could estimate its target population mean given a choice of benchmark function and a value for the relative selection ratio. The key questions are then what benchmark functions we should use and how to interpret the value of the relative selection ratio. The next section discusses these issues in the context of ATE extrapolation.

3 Application to Treatment Effects

We turn now to assessing external validity of the ATE. The results developed above continue to apply in this setting, except that individual-level treatment effects are not directly observed in the data (we observe each individual in only a single treatment state) so we need to construct a target function $t(X_i)$ with mean equal to the ATE. In the first subsection below we discuss a target function $t\left(X_i\right)$ that estimates the ATE under random assignment to treatment.

We then turn to the choice of benchmark function $b\left(X_i\right)$, and the interpretation of the relative selection ratio. In many settings it seems plausible that selection into the trial population is driven by expected treatment effects. In such settings, we show that if we take $b\left(X_i\right)$ to be the predicted treatment effect based on covariates, the relative selection ratio measures the degree of private information (that is, information not captured by covariates) about the treatment effect which is used in the selection process. We then extend this result to settings where selection is on some dimension other than the treatment effect.

In both cases, the procedure we suggest first adjusts for selection on observables, and then links the remaining degree of selection to private information in the selection process. Adjusting for selection on observables is sometimes already done formally in experimental papers (Chyn, 2016; Alcott, 2015) and is more commonly discussed intuitively (Dupas and Robinson, 2013; Muralidaran, Singh and Ganimian, 2016). Our framework directly links these adjustments to overall external validity bias.

3.1 Moments for Treatment Effect Estimation

Inference on the ATE is complicated by the fact that even in the trial population we observe each individual in only a single treatment state, and so never observe treatment effects at the individual level. By choosing $t(X_i)$ appropriately, our approach nonetheless allows us to draw inferences about the ATE in the target population.

To develop these results we adopt the usual potential outcomes framework (see e.g. Imbens and Rubin 2015). Formally, suppose we are interested in the effect of a binary treatment, with $D_i \in \{0,1\}$ a dummy equal to one when i is treated. We write the outcomes of individual i in the untreated and treated states as $Y_i(0)$, $Y_i(1)$, respectively. Assume that we observe a vector of covariates for each individual, C_i, which are unaffected by treatment. The observed outcome for i is

$$Y_i = Y_i(D_i) = (1-D_i)\,Y_i(0) + D_i Y_i(1),$$

and the observed data are $X_i = (Y_i, D_i, C_i)$. We are interested in inference on the ATE in the target population $E_P[TE_i]$, where the treatment effect $TE_i = Y_i(1) - Y_i(0)$ measures the effect of treatment on individual i.

We assume that treatment is randomly assigned in the trial population. In particular, we assume that D_i is independent of $(Y_i(0), Y_i(1), C_i)$ under P_S, with known mean $E_{P_S}[D_i] = d$. We can express the ATE in the trial population,

$$E_{P_S}[TE_i] = E_{P_S}[Y_i(1) - Y_i(0)]$$

as the difference between the mean outcome in the treated and untreated groups,

$$E_{P_S}[Y_i|D_i=1] - E_{P_S}[Y_i|D_i=0] = E_{P_S}\left[\frac{D_i}{d}Y_i - \frac{(1-D_i)}{1-d}Y_i\right].$$

Thus, under random assignment of D_i we can write the trial population ATE as $E_{P_S}[T_i]$ for

$$T_i = \frac{D_i}{d}Y_i - \frac{1-D_i}{1-d}Y_i.$$

While our analysis is motivated by the fact that we cannot randomly assign treatment in the target population, we define the distribution P as that which would arise were we able to randomly assign the target population to treatment, again with $E_P[D_i] = d$.[3] This allows us to write the target population ATE as $E_P[T_i]$. Hence we can cast estimation of ATEs in the target population into our general framework by taking $t(X_i) = T_i$.

3.2 Selection Models and the Choice of $b(X_i)$

To apply our approach to the ATE, in addition to setting $t(X_i) = T_i$ we must specify a benchmark function $b(X_i)$ and develop methods for interpreting the relative selection ratio. In this section, we show that for a large class of selection models, if we take $b(X_i)$ to be the predicted treatment effect given covariates, the relative selection ratio has an intuitive interpretation in terms of private information used in the selection process. We first consider the case where selection into the trial population is driven by expected treatment effects, and then discuss the case where selection is driven by other variables.

Unobserved Variables Throughout this section we assume that in addition to $X_i = (Y_i, D_i, C_i)$, there are also variables U_i which are unobserved by the researcher but may play a role in the selection process. Further, we assume that the distribution of the covariates C_i in the target population is known (though we discuss in Section 4.1 below how we can proceed if we know only some aspects of this distribution). If there are variables which are observed in the trial population but whose distribution in the target population is entirely unknown, for the purposes of analysis we include these in U_i. For brevity of notation we will denote the conditional expectation in the trial population of a random variable B_i given (C_i, U_i) by $\widetilde{B}_i = E_{P_S}[B_i|C_i, U_i]$, and the conditional expectation given C_i alone by $\widehat{B}_i = E_{P_S}[B_i|C_i]$.

3.2.1 Selection Framework

Our results are based on a model for selection into the trial population. As in Lemma 1 we assume that the trial population is a subset of the target population, and define S_i to be an indicator for membership in the trial population (that is, observation i in the target

[3] While we focus on simple random assignment of D_i, if one instead considers random assignment conditional on covariates, with $D_i \perp (Y_i(1), Y_i(0))|C_i$ and $E_{P_S}[D_i|C_i] = d(C_i)$ for known $d(\cdot)$, we can instead take $T_i = \left(\frac{D_i}{d(C_i)} - \frac{1-D_i}{1-d(C_i)}\right) Y_i$ and our results below will go through provided we assume the same mechanism for assignment to treatment (conditional on covariates) in the target population. This follows from well-known results in the literature on propensity score reweighting- see Rosenbaum and Rubin (1983).

population also belongs to the trial population if and only if $S_i = 1$). Selection depends on the covariate C_i, the unobservable U_i and an independent idiosyncratic variable V_i. We assume that S_i is determined by a standard latent index model:

Assumption 2 *The selection dummy S_i satisfies*

$$S_i = 1\left\{g\left(C_i, U_i\right) \geq V_i\right\}, \tag{7}$$

where V_i is continuously distributed with density p_V independently of $\left(C_i, U_i, Y_i\left(0\right), Y_i\left(1\right)\right)$, and has support equal to $\mathbb{R}$. We further assume that $0 < E_P\left[S_i | C_i, U_i\right] < 1$ for all C_i, U_i.

This assumption nests a wide variety of parametric and nonparametric selection models. The restriction that $0 < E_P\left[S_i | C_i, U_i\right] < 1$ ensures that distributions in the trial and target populations are mutually absolutely continuous, so Assumption 1 holds.

Given this assumption we can show that the expected treatment effect given (C_i, U_i) is the same in the trial and target populations.

Lemma 3 *Under Assumption 2, the conditional expectation of the treatment effect given (C_i, U_i) is the same in the trial and target populations*

$$E_{P_S}\left[TE_i | C_i, U_i\right] = E_P\left[TE_i | C_i, U_i\right] = \widetilde{TE}_i.$$

3.2.2 Selection on Treatment Effects

We first consider the case where selection into the sample is driven by expected treatment effects $\widetilde{TE}_i$. Since we can take U_i to include TE_i, this also covers the case of direct selection on the treatment effect, with $\widetilde{TE}_i = TE_i$.

Assumption 3 *The function g in equation (7) is of the form*

$$g\left(C_i, U_i\right) = c \cdot \widetilde{TE}_i \tag{8}$$

for some constant $c \in \mathbb{R}$.

Under this assumption, we show that the weights to rebalance the joint distribution of the covariates and the expected treatment effects can be written in terms of $\widetilde{TE}_i$.

Lemma 4 *Under Assumptions 2 and 3, the weights to rebalance* $\left(C_i, \widetilde{TE}_i\right)$ *can be written as*

$$W_i = \frac{p\left(C_i, \widetilde{TE}_i\right)}{p_S\left(C_i, \widetilde{TE}_i\right)} = w\left(\widetilde{TE}_i\right)$$

for a continuously differentiable function w.[4]

This result implies that in order to calculate the mean of any function $f\left(C_i, \widetilde{TE}_i\right)$ in the target population it suffices to reweight based on $\widetilde{TE}_i$, $E_P\left[f\left(C_i, \widetilde{TE}_i\right)\right] = E_{P_S}\left[W_i f\left(C_i, \widetilde{TE}_i\right)\right]$. Since Lemma 3 implies that $E_P\left[TE_i\right] = E_P\left[\widetilde{TE}_i\right]$, this covers the ATE as a special case. This does not provide an implementable procedure, however, since $\widetilde{TE}_i$ is unobserved and the function $w\left(\cdot\right)$ is unknown and depends on the underlying selection process.

Approximate Bias: To overcome both of these difficulties, we consider an approximation to W_i. Specifically, we consider a Taylor approximation to $W_i = w\left(\widetilde{TE}_i\right)$ around the ATE in the trial population,[5]

$$W_i \approx W_i^* = w_0 + w_1 \widetilde{TE}_i.$$

Using W_i^* we obtain approximate bias expressions for any function $f\left(C_i, \widetilde{TE}_i\right)$, since by Corollary 1,

$$\begin{aligned} E_P\left[f\left(C_i, \widetilde{TE}_i\right)\right] - E_{P_S}\left[f\left(C_i, \widetilde{TE}_i\right)\right] &= Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i\right)\right) \\ \approx Cov_{P_S}\left(W_i^*, f\left(C_i, \widetilde{TE}_i\right)\right) &= w_1 Cov_{P_S}\left(\widetilde{TE}_i, f\left(C_i, \widetilde{TE}_i\right)\right). \end{aligned} \tag{9}$$

In the Appendix we show that under the assumption that $\widetilde{TE}_i$ is bounded and mild regularity conditions, the error in this approximation vanishes when selection is primarily driven by idiosyncratic factors, so the constant c in equation (8) is small (holding the distribution of V_i fixed).[6] Thus, equation (9) can be viewed as an approximation around the case where selection into the sample is purely random. To use this result to assess external validity of

[4] In fact, these weights suffice to rebalance $\left(C_i, U_i\right)$.

[5] First-order Taylor approximations yields $w_1 = w'(E_{P_S}[TE_i])$, $w_0 = w'(E_{P_S}[TE_i]) - w_1 E_{P_S}[TE_i]$.

[6] In particular, while $Cov\left(W_i^*, f\left(X_i\right)\right)$ and $Cov\left(W_i, f\left(X_i\right)\right)$ both tend to zero as $c \to 0$, we show that the approximation error is of lower order.

the ATE, the proof of Proposition 1 below shows that equation (9) implies

$$Cov_{P_S}\left(W_i, TE_i\right) = Cov_{P_S}\left(W_i, T_i\right) \approx Cov_{P_S}\left(W_i^*, T_i\right),$$

so we can use the weights W_i^* to obtain approximate expressions for the bias in the ATE.

Benchmark Function: To construct a relative selection ratio as in equation (6) we take as our benchmark moment the predicted treatment effect based on the covariates C_i

$$b\left(X_i\right) = \widehat{T}_i = E_{P_S}\left[T_i | C_i\right] = E_{P_S}\left[TE_i | C_i\right] = \widehat{TE}_i.$$

The target population mean of this moment,

$$E_P\left[\widehat{T}_i\right] = E_P\left[E_{P_S}\left[TE_i | C_i\right]\right]$$

corresponds to the estimate of the ATE corrected for selection on observables (as in e.g. Hotz et al (2005), Stuart et al (2011)). The difference

$$E_P\left[\widehat{T}_i\right] - E_{P_S}\left[\widehat{T}_i\right] = E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]$$

thus measures the adjustment for selection on observables, where $E_{P_S}\left[\widehat{T}_i\right] = E_{P_S}\left[T_i\right]$ by the law of iterated expectations.

For this choice of benchmark, we can relate the relative selection ratio to the degree of private information used in the selection process:

Proposition 1 *Under Assumptions 2 and 3, provided* $w_1 \neq 0$

$$\frac{E_P\left[TE_i\right] - E_{P_S}\left[TE_i\right]}{E_P\left[\widehat{TE}_i\right] - E_{P_S}\left[TE_i\right]} = \frac{E_P\left[T_i\right] - E_{P_S}\left[T_i\right]}{E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]} \approx \frac{Cov_{P_S}\left(W_i^*, T_i\right)}{Cov_{P_S}\left(W_i^*, \widehat{T}_i\right)} = \frac{Var_{P_S}\left(\widetilde{TE}_i\right)}{Var_{P_S}\left(\widehat{TE}_i\right)} = \Phi.$$

To interpret this result, recall that

$$\Phi = \frac{Var_{P_S}\left(\widetilde{TE}_i\right)}{Var_{P_S}\left(\widehat{TE}_i\right)} = \frac{Var_{P_S}\left(E_{P_S}\left[TE_i | C_i, U_i\right]\right)}{Var_{P_S}\left(E_{P_S}\left[TE_i | C_i\right]\right)}.$$

Thus Φ measures the variance of treatment effects predicted based on (C_i, U_i), relative to the variance of treatment effects predicted based on C_i alone. This can be interpreted as a measure for the degree of private information about the treatment effect used in the selection process. In the extreme case where selection is directly on the treatment effect and $\widetilde{TE}_i = TE_i$, Φ^{-1} measures the share of treatment effect heterogeneity captured by the covariates (specifically the R^2 from nonparametrically regressing TE_i on C_i).

By the law of total variance,

$$\Phi = 1 + \frac{E_{P_S}\left[Var_{P_S}\left(\widetilde{TE}_i|C_i\right)\right]}{Var_{P_S}\left(\widehat{TE}_i\right)},$$

so $\Phi \geq 1$. This is intuitive, and reflects the fact that since selection is based on (C_i, U_i) , the selection process always uses at least as much information about the treatment effects as is contained in the covariates.

Proposition 1 shows that under the assumption of selection on the treatment effect, the ratio of bias in the ATE to bias in the average of $\widehat{TE}_i = \widehat{T}_i$ depends on the amount of private information used in selection. When there is a large amount of private information the true bias in the ATE will be much larger than the bias in $\widehat{TE}_i$. By contrast, when there is little private information the bias in the ATE will be close to that in $\widehat{TE}_i$. In the extreme case where we assume no private information, $\Phi = 1$ and $E_P[TE_i] = E_P\left[\widehat{T}_i\right]$ so we can obtain the target population ATE by correcting for differences in the distribution of the covariates C_i between the trial and target populations. In this case our approach coincides with existing alternatives which correct for selection on observables.

Even when we do not assume selection on observables, one implication of the results above and the fact that $\Phi \geq 1$ is that (up to approximation error)

$$\frac{E_P[TE_i] - E_{P_S}[TE_i]}{E_P\left[\widehat{TE}_i\right] - E_{P_S}[TE_i]} = \frac{E_P[T_i] - E_{P_S}[T_i]}{E_P\left[\widehat{T}_i\right] - E_{P_S}[T_i]} \geq 1.$$

Thus, if correcting for selection on observables reduces the estimated ATE, assuming agents have private information will lead to still-larger reductions.

3.2.3 Selection on Other Variables

While the assumption of selection on the treatment effect (Assumption 3) is often plausible, it does not apply in all settings. In this section we generalize our results to allow selection on some variable $A_i \neq TE_i$. Since the results are quite similar to those in the case with selection on observables, we present them with minimal discussion except where they differ.

Assumption 4 *The function g in equation (7) is of the form*

$$g\left(C_i, U_i\right) = c \cdot \widetilde{A}_i$$

for $c \in \mathbb{R}$.

Lemma 3 immediately generalizes to this setting.

Lemma 5 *Under Assumptions 2 and 4, the conditional expectations of TE_i and A_i given (C_i, U_i) are the same in the trial and target populations*

$$E_{P_S}\left[TE_i | C_i, U_i\right] = E_P\left[TE_i | C_i, U_i\right] = \widetilde{TE}_i.$$

$$E_{P_S}\left[A_i | C_i, U_i\right] = E_P\left[A_i | C_i, U_i\right] = \widetilde{A}_i.$$

As before the weights to rebalance the joint distribution of the covariates and expected treatment effects can be written as a function of $\widetilde{A}_i$.

Lemma 6 *Under Assumptions 2 and 4, the weights to rebalance $\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)$ can be written as*

$$W_i = \frac{p\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)}{p_S\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)} = w\left(\widetilde{A}_i\right)$$

for a continuously differentiable function w.

Approximate Bias: We next consider a Taylor approximation to the weights $w\left(\widetilde{A}_i\right)$. Taylor expanding $w\left(\widetilde{A}_i\right)$ around $E_{P_S}\left[A_i\right]$ yields

$$W_i \approx W_i^* = w_0 + w_1 \widetilde{A}_i.$$

As before, these weights yield approximate bias expressions for any function $f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)$,

$$\begin{aligned} E_P\left[f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right] - E_{P_S}\left[f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right] &= Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) \\ \approx Cov_{P_S}\left(W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) &= w_1 Cov_{P_S}\left(\widetilde{A}_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right). \end{aligned} \tag{10}$$

Provided $\widetilde{A}_i$ is bounded, the error in this approximation again vanishes when the constant c in equation (7) is small.

Proposition 1 immediately extends to this case.

Proposition 2 *Under Assumptions 2 and 4, provided* $w_1 \neq 0$

$$\frac{E_P\left[TE_i\right] - E_{P_S}\left[TE_i\right]}{E_P\left[\widehat{TE}_i\right] - E_{P_S}\left[TE_i\right]} = \frac{E_P\left[T_i\right] - E_{P_S}\left[T_i\right]}{E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]} \approx \frac{Cov_{P_S}\left(W_i^*, T_i\right)}{Cov_{P_S}\left(W_i^*, \widehat{T}_i\right)} = \frac{Cov_{P_S}\left(\widetilde{A}_i, \widetilde{TE}_i\right)}{Cov_{P_S}\left(\widehat{A}_i, \widehat{TE}_i\right)} = \Phi_A.$$

To interpret Φ_A, note that

$$\Phi_A = \frac{Cov_{P_S}\left(\widetilde{A}_i, \widetilde{TE}_i\right)}{Cov_{P_S}\left(\widehat{A}_i, \widehat{TE}_i\right)} = \frac{Cov_{P_S}\left(E_{P_S}\left[A_i|C_i, U_i\right], E_{P_S}\left[TE_i|C_i, U_i\right]\right)}{Cov_{P_S}\left(E_{P_S}\left[A_i|C_i\right], E_{P_S}\left[TE_i|C_i\right]\right)}$$

measures the covariance between the predictions for A_i and TE_i based on (C_i, U_i), relative to the covariance of the predictions based on the covariates C_i alone. By the law of total covariance,

$$\Phi_A = 1 + \frac{E_{P_S}\left[Cov_{P_S}\left(\widetilde{A}_i, \widetilde{TE}_i|C_i\right)\right]}{Cov_{P_S}\left(\widehat{A}_i, \widehat{TE}_i\right)}. \tag{11}$$

The numerator in the second term measures the average covariance between $\widetilde{A}_i$ and $\widetilde{TE}_i$ after controlling for C_i, and can also be written as

$$E_{P_S}\left[Cov_{P_S}\left(\widetilde{A}_i, \widetilde{TE}_i|C_i\right)\right] = Cov_{P_S}\left(\widetilde{A}_i - \widehat{A}_i, \widetilde{TE}_i - \widehat{TE}_i\right).$$

The ratio in equation (11) therefore measures the covariance of the residuals from nonparametrically regressing $\widetilde{A}_i$ and $\widetilde{TE}_i$ on the covariates, divided by the covariance of the fitted values. Thus, if the covariance of $\widetilde{A}_i$ and $\widetilde{TE}_i$ is driven primarily by covariates this term will be small, while if the covariance is driven primarily by private information this term will be large.

Unlike for Φ there are not universal bounds on Φ_A. In particular Φ_A can be larger or smaller than one in absolute value, and can be either positive or negative. Together with the more complicated interpretation of Φ_A, this means that to assess external validity of the ATE when we think selection is on some variable A_i not equal to the treatment effect we must think carefully about the likely role of observable and unobservable variables in explaining the covariance between A_i and TE_i. In an important special case, however, the interpretation simplifies to that for selection on treatment effects above.

Selection on the Treatment Effect and Orthogonal Variables A useful special case arises when selection is driven by a combination of the treatment effect and unrelated variables. In particular, suppose that

$$A_i = \alpha_1 TE_i + \alpha_2 f\left(C_i\right) + \alpha_3 h\left(U_i\right),$$

where $\alpha_1 \neq 0$,

$$Cov_{P_S}\left(f\left(C_i\right), TE_i\right) = Cov_{P_S}\left(h\left(U_i\right), TE_i\right) = 0,$$

and

$$Cov_{P_S}\left(E_{P_S}\left[h\left(U_i\right)|C_i\right], TE_i\right) = 0.$$

This allows selection to depend on the covariates and unobservables through functions $f\left(C_i\right)$ and $h\left(U_i\right)$ other than $\widetilde{TE}_i$, but requires that these functions be uncorrelated with the treatment effect. The last restriction further requires that $h\left(U_i\right)$ remain uncorrelated with the treatment effect even after we take its conditional expectation given C_i. A simple sufficient condition is that $E_{P_S}\left[h\left(U_i\right)|C_i\right] = E_{P_S}\left[h\left(U_i\right)\right]$, so $h\left(U_i\right)$ is mean-independent of the covariates. Under these conditions,

$$\frac{Cov_{P_S}\left(\widetilde{A}_i, \widetilde{TE}_i\right)}{Cov_{P_S}\left(\widehat{A}_i, \widehat{TE}_i\right)} = \frac{Var_{P_S}\left(\widetilde{TE}_i\right)}{Var_{P_S}\left(\widehat{TE}_i\right)} = \Phi,$$

so we recover the relatively more straightforward interpretation for the selection ratio discussed in the selection on treatment effects case above.

4 Implementation and Example

This section discusses implementation of our approach and examines performance in a constructed example. It is important to note that our approach relies on the model and assumptions described in Section 3. To the extent that one wants to model the selection process differently, the objects we suggest here can still be calculated but will not have the same interpretation.

4.1 Implementation

Consider a case where we have an estimate of the ATE $E_{P_S}[TE_i]$ in a trial population. We are interested in the range of plausible values for the ATE in the target population. To account for the possibility of further selection on unobservables, we discuss two distinct approaches. First, we can consider a particular value for the ATE in the target population, t_P^*, and ask how much private information would have to be present in the selection process to obtain this value. Alternatively, we can impose bounds on the degree of private information and calculate the implied range of ATEs.

For both approaches, an important first step is correction for selection on observables, and specifically estimation of $\widehat{TE}_i = \widehat{T}_i = E_{P_S}[TE_i|C_i]$. In this section we discuss a simple approach based on linear regression of the treatment effect proxy T_i on functions of covariates, which can be applied even with limited knowledge of the features of the target population. In settings with richer data on the target population, one can also apply our general approach together with more sophisticated corrections for selection on observables.

Given the selection on observables adjustment, we can evaluate robustness by calculating the value of Φ or Φ_A sufficient to overturn our conclusions. To do this, we consider a target value of interest t_P^* (a natural value in many treatment effect settings is zero) and calculate the value Φ or Φ_A required to yield an ATE of t_P^* in the target population, $E_P[TE_i] = t_P^*$. Let us denote this value by $\Phi(t_P^*)$ (calculations based on Φ_A are identical). To calculate $\Phi(t_P^*)$, we simply compare the implied total adjustment to the selection on observables adjustment,

$$\Phi(t_P^*) = \frac{t_P^* - E_{P_S}[T_i]}{E_P\left[\widehat{T}_i\right] - E_{P_S}[T_i]}.$$

Alternatively, we can impose bounds $\Phi \in [\Phi_L, \Phi_U]$. Under these bounds we know that (assuming the selection on observables correction is positive),[7]

$$E_P[TE_i] \in E_{P_S}[T_i] + \left[\Phi_L\left(E_P\left[\widehat{T}_i\right] - E_{P_S}[T_i]\right), \Phi_U\left(E_P\left[\widehat{T}_i\right] - E_{P_S}[T_i]\right)\right]. \tag{12}$$

With the selection on observables correction and the trial population ATE, $E_{P_S}[TE_i]$, we can thus easily calculate the implied range of values for $E_P[TE_i]$.

Up to this point, the calculations are the same under either model of selection. Interpreting the results, however, requires taking a stand on the selection process.

Selection on Treatment Effects Consider first the case where we model the selection process by assuming that units are selected on the treatment effect. In this case the key object is Φ, which can be interpreted as a measure of the degree of private information about treatment effects used in the selection process. For example, suppose we find that a value of $\Phi = 2$ is necessary to eliminate a positive result. This indicates that the unobservables would have to be at least as informative about the treatment effect as the observables in order for the effect in the target population to be zero.[8]

Further intuition may be provided by thinking about the share of relevant variables missed by our observed covariates. In Appendix B.2, we describe a model where a large number of latent factors drive the treatment effect, and a random subset of these are measured by C_i while the rest are measured by U_i. In this setting Φ can be interpreted as the ratio of the total to the observed factors. In particular, $\Phi = 2$ reflects a case where the observed covariates capture 50% of the latent factors.[9]

Selection on Other Variables The second case is one in which we model the section process as occurring on another variable which is not observed in the target population. For

[7] If it is instead negative, then

$$E_P[TE_i] \in E_{P_S}[T_i] + \left[\Phi_U\left(E_P\left[\widehat{T}_i\right] - E_{P_S}[T_i]\right), \Phi_L\left(E_P\left[\widehat{T}_i\right] - E_{P_S}[T_i]\right)\right].$$

[8] In particular, if we regress the predicted treatment effects based on both unobservables and observables on the observables alone, the R^2 must be less than 0.5.

[9] The assumptions needed for this interpretation are considerably stronger than those required for the rest of our results. We thus regard the model yielding this result more as a way to build intuition than as a description of a plausible data generating process.

example, we might have a case where experimental locations were selected on accessibility, but we do not observe accessibility measures. In this case the ratio of the total to observable adjustment measures the relative importance of the observables and unobservables in explaining the covariance between this selection variable and the treatment effect, which we denote by Φ_A.

As above, suppose we find that a value of $\Phi_A = 2$ is necessary to eliminate a positive result. This means that unobservables would have to explain as much of the covariance of A_i and TE_i as the observables in order for the effect in the target population to be zero. Again, in the model described in Appendix B.2 one can relate Φ_A to the share of factors captured by the covariates, and a value of $\Phi_A = 2$ can be interpreted as a case where the observed covariates capture 50% of the latent factors.

4.1.1 Correction for Selection on Observables

To implement the approaches discussed above, we need an estimate of $E_P\left[\widehat{T_i}\right] - E_{P_S}\left[T_i\right]$. We can estimate $E_{P_S}\left[T_i\right]$ by the sample average of T_i, so the challenge is estimation of $E_P\left[\widehat{T_i}\right]$, the ATE corrected for selection on observables.

In our applications below we estimate $\widehat{T_i}$ by regressing T_i on a vector of functions of the covariates $r\left(C_i\right)$ whose mean $E_P\left[r\left(C_i\right)\right]$ in the target population is known,

$$T_i = r\left(C_i\right)' \delta + e_i,$$

where we assume $r\left(C_i\right)$ includes a constant. We then approximate $E_P\left[\widehat{T_i}\right]$ by $E_P\left[r\left(C_i\right)\right]' \delta$. If we assume a linear model for treatment effect heterogeneity,

$$E_{P_S}\left[T_i|C_i\right] = E_{P_S}\left[TE_i|C_i\right] = r\left(C_i\right)' \delta,$$

then this procedure exactly recovers $E_P\left[\widehat{T_i}\right] = E_P\left[\widehat{TE_i}\right]$. If on the other hand we consider the linear specification as an approximation, then this procedure delivers an approximation to $E_P\left[\widehat{T_i}\right]$, where the approximation error will vanish as we consider rich sets of functions $r\left(C_i\right)$.[10]

[10] Ideally we would include interactions and higher-order terms in $r\left(C_i\right)$, although this may be infeasible given data constraints. Nonetheless, whenever possible researchers should at a minimum include linear and

An advantage of the regression approach we use in this paper is that it can be implemented based on knowledge of $E_P\left[r\left(C_i\right)\right]$ alone, and so does not require us to know the full distribution of C_i in the target population. In settings where more is known about the distribution of C_i under P, however, one could also consider other methods, for example matching as in Hotz et al (2005), or propensity score reweighting as in Stuart et al (2011). Such approaches again yield estimates of the ATE corrected for selection on observables, $E_P\left[\widehat{T}_i\right]$, which can be plugged into our approach exactly as described above.

4.1.2 Inference

Thus far, we have conducted our analysis treating the distribution P_S in the trial population as known. In applications we observe only a finite sample from P_S, however, and need to account for sampling uncertainty. In discussing inference we focus on the case of simple random sampling, where treatment is assigned iid across units. For discussion of the complications arising from other randomization schemes see Bugni, Canay and Shaikh (2017). The development of inference results for our approach in such settings is an interesting question for future work.

Under the assumption of simple random assignment, we can conduct inference using the bootstrap.[11] Bootstrap standard errors for $\Phi\left(t_P^*\right)$ become unreliable when the correction for selection on observables is close to zero, however. In this case, the denominator in $\Phi\left(t_P^*\right)$ is almost zero, which results in problems very similar to those that arise from weak instruments.[12] In Appendix B.3 we discuss how to construct reliable confidence sets for $\Phi\left(t_P^*\right)$. These confidence sets are close to the usual ones when the selection on observables correction is large, but can be unbounded when it is small.

Confidence sets for the ATE $E_P\left[TE_i\right]$ are more straightforward. In particular, for $\left(\hat{\sigma}_L, \hat{\sigma}_U\right)$

squared terms in the covariates, since this will capture differences between the trial and target populations in the means and variances of these variables. In settings with richer data one should consider even more moments - interactions between the variables, higher moments of the distribution of each variable, etc.

[11] Note that when we estimate the distribution in the target population from a sample, we should bootstrap target population quantities as well in order to obtain accurate measures of uncertainty.

[12] The selection on observables correction $E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]$ plays the same role as the first-stage parameter in linear IV, so problems arise when this difference is close to zero relative to sampling variability.

bootstrap standard errors for our estimates $(\hat{\gamma}_L, \hat{\gamma}_U)$ of

$$(\gamma_L, \gamma_U) = \left(E_{P_S}\left[T_i\right] + \Phi_L\left(E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]\right), E_{P_S}\left[T_i\right] + \Phi_U\left(E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]\right)\right),$$

we can construct a (conservative) level $1-\alpha$ confidence interval for $E_P\left[TE_i\right]$ as

$$\left[\min\left\{\hat{\gamma}_L - \hat{\sigma}_L c_\alpha, \hat{\gamma}_U - \hat{\sigma}_U c_\alpha\right\}, \max\left\{\hat{\gamma}_L + \hat{\sigma}_L c_\alpha, \hat{\gamma}_U + \hat{\sigma}_U c_\alpha\right\}\right],$$

for c_α the two-sided level $1-\alpha$ normal critical value (e.g. 1.96 for a 95% confidence set).[13] Alternatively, one can report $(\hat{\gamma}_L, \hat{\gamma}_U, \hat{\sigma}_L, \hat{\sigma}_U)$, which allows readers to construct the confidence set of their choice.

4.2 Example

We illustrate our approach in an example. To ensure that we know the true form of selection while also having an empirically reasonable distribution for the data, we use a constructed example based on a real experiment.

4.2.1 Data and Empirical Approach

We base our example on data from Muralidharan and Sundararaman (2011), which is a randomized evaluation of a teacher performance pay scheme in India. The project includes student-level data from roughly 300 schools across the state of Andhra Pradesh. Teachers in "incentive" schools were paid more for better student test scores, while those in control schools were not. The primary outcome is student test scores. Muralidharan and Sundararaman (2011) find that student test scores increase as a result of incentive pay.

To construct our example, we define the distribution of the target population to be the empirical distribution in the Muralidharan and Sundararaman (2011) data. To abstract from issues of sampling variability, we collapse the data to the school level and sample from the data with replacement to create a large target population.

We predict treatment effects based on school-level characteristics: average teacher education, average teacher training, average teacher salary, average household income, a school

[13] In fact, one can typically use a critical value smaller than c_α, though more computation is required to derive the exact value. We do this in our applications. See Appendix B.3 for details.

infrastructure index, the share of the student population who is scheduled tribe or scheduled caste, and average teacher absence. We also include dummies for which mandal (a geographic area) the school is in. We can think of these controls as capturing differences across areas in how effective the program is.

From this target population, we extract a trial population selected either on the predicted treatment effect or on area-level characteristics. This selection process is described in more detail in each case below. Under both schemes the ATE in the trial population exceeds that in the target population. Our sample construction is such that if we observed all of the school-level characteristics in both the trial and target populations, we could recover the target population treatment effect using the selection on observables adjustment. Our approach, then, is to explore what happens as we treat increasingly large sets of the characteristics as unobserved.

4.2.2 Selection on Treatment Effect

We first model selection on the predicted treatment effect. We create a predicted treatment effect $\widetilde{TE}_i$ by defining T_i as in Section 3 above and regressing T_i on the full set of controls for school-level characteristics. We select schools into the trial population if $\widetilde{TE}_i \geq V_i$ where V_i is normally distributed with the same mean as $\widetilde{TE}_i$ and a standard deviation three times as large.[14]

The ATE in the target population is 0.074.[15] The ATE in the trial population is considerably larger, 0.15. If we assume that we observe all the characteristics used in the selection process, the adjustment for selection on observables delivers the correct value 0.074 for the target population ATE.

We next consider the case where we cannot observe some of the variables used in the selection process. We vary the size of the subset which is unobserved, considering what happens when we eliminate just one variable, then 10%, 20%, 30%, and 50% of the variables (chosen at random).[16] In each case we calculate the ATE correcting for selection on observables, where there is now also selection on unobservables. We consider all possible single-variable

[14] The mean of V_i ensures that roughly half of the population will be in treatment, and the larger standard deviation limits approximation error, since we approximate around the fully random case.

[15] This is slightly smaller than the effect in the original paper since we collapse to the school level.

[16] Performance in this example remains quite good even when we exclude 80% of the variables.

eliminations, while for the other cases we take 200 draws at random.

The first column of Panel A of Table 3 shows the average selection on observables adjusted ATE for each exclusion set. When only one covariate is treated as unobserved (the last row in Panel A) the estimate is extremely close to the target population ATE, since the unobservables are quite limited. As we treat larger sets as unobserved the estimate is further from the target population ATE and closer to the trial population estimate.

The second column of Panel A in this table reports the average value of Φ to match the target population treatment effect for each specification. This value is largest when the largest share of covariates is excluded. It is worth noting that the average values of Φ are quite close to the actual ratio of the number of total covariates to the observed covariates, reflecting the intuition described in cases with many covariates.

We can visualize the range of values Φ which generate the target population ATE, given each set of unobservables. This is done in Figure 1. As we exclude a larger set of variables, the range of Φ goes up, consistent with the presence of more private information in the selection process. These values of Φ correspond directly to the relative importance of the observed versus unobserved covariates in predicting the treatment effect. To see this more directly, we calculate the ratio of the R-squared from regressing the treatment effect proxy T_i on the observed covariates to the R-squared from regressing on all the school characteristics. We graph this against the value of Φ to match the true bias. Deviations from equality arise from approximation error. Figure 2 suggests such error here is limited.

4.2.3 Selection on Covariates: Results

We next model selection on features of the data other than the treatment effect. In particular, we imagine that we select areas based on mandal-level teacher training. We divide the sample into quartiles based on the mandal-level average of teacher training, and then calculate the average treatment effect within each quartile, which we use as our $\widetilde{A}_i$. In practice, this puts more weight on mandals with the highest teacher training values, and on areas in the second quartile of training. This approximates a case where experimental locations are selected on average teacher training, with a preference for teacher training levels predictive of high treatment effects.[17]

[17] The selection on teacher training here is non-linear, reflecting the actual patterns in the data.

Given this index $\widetilde{A}_i$, we select schools into the sample if $\widetilde{A}_i \geq V_i$ where V_i is normally distributed with the same mean as $\widetilde{A}_i$ and a standard deviation three times as large.

Although the structure of the sample selection is similar to the selection on treatment effects case discussed above, the difference in ATEs between the trial and target populations is less extreme. The target population ATE is again 0.074, while that in the trial population is 0.119.

We again consider the case where we cannot observe a subset of the variables used in the construction of the index $\widetilde{A}_i$. As above, we consider varying the size of the subset which is unobserved and calculate the same selection on observables quantities as above.

Panel B of Table 3 replicates Panel A for this selection procedure. When we treat larger sets as unobserved, the estimate is further from the target population ATE, and closer to the trial population estimate. The values of Φ_A are largest for the largest exclusion set, and reflect the share of covariates missing from the observable set.

Figure 3 plots the distribution of the values Φ_A that would generate the target population ATE as we consider different sets of observables. With small exclusion sets the values are relatively small, although with large sets of variables treated as unobserved the results are noisier, and sometimes imply very large values of Φ_A to match the true treatment effect. This is also reflected in Figure 4, which graphs the value of Φ_A to match the true bias against the ratio of the covariance of $\widetilde{A}_i, \widetilde{TE}_i$ to the covariance of $\widehat{A}_i, \widehat{TE}_i$. There is a strong relationship here, but it is not as tight as in the case of treatment effects. It is worth noting that as we increase the share of missing covariates, the behavior of this ratio is more erratic.

5 Applications

This section discusses a number of specific examples applying our framework to papers in the literature. Before moving to these examples, however, we briefly discuss in what sorts of applications we expect our approach to be useful.

5.1 Scope of Application

The problem of external validity is quite broad, and encompasses a wide variety of different questions. Many of the examples we discuss below focus on extrapolating from a sample of people to the broader population from which they are drawn, but one might also be interested in extrapolating from one location to another, or from one time period to another. A further sort of external validity concern relates to the general equilibrium consequences of treating an entire population as opposed to a small sample, regardless of how the sample was selected. Our approach is better suited to handling some of these problems than others.

To make the range of different external validity concerns concrete, consider a (hypothetical) experiment studying a job training program. Imagine this is a small program, run in a single city, at a time of high unemployment rates. Selection into the program is based on a lottery among individuals who express interest. There are at least four types of extrapolation we might be interested in: extrapolation to a similarly sized *random* sample of the full population, extrapolation to the full population, extrapolation to a time period with a lower unemployment rate, and extrapolation to other locations. We will briefly discuss the role of our approach in addressing each of these extrapolation problems.

Extrapolation to Random Sample Our approach is most directly applicable if we want to extrapolate to a similarly sized random sample. In the hypothetical job training example above, for example, we might want to extrapolate to the average treatment effect on a random sample from the city's population. Our main assumption (Assumption 1) is plausible in this setting, and there is a clear intuition for how to apply the models outlined in Section 3 to model selection (in our hypothetical example, the decision to volunteer).

Extrapolation to Full Population In many settings our approach is also potentially suited to considering extrapolation to a full population. In the job training example above, for instance, we might want to know what would happen if the program were expanded to cover everyone in the city. A complication, however, is that in some settings treating the entire population could introduce important general equilibrium or spillover effects. In settings where such issues arise it may well be interesting to undertake the analysis we suggest, but to accurately predict the effect of treating the full population one will need to separately account for additional effects arising from the scale of treatment.

An additional problem in extrapolating to the full population relates to our assumption of common support - Assumption 1 – which rules out the possibility that there are types in the target population that never arise in the trial population. This rules out extrapolation to people who could not be included in the experiment. If our hypothetical job training experiment is limited to high school dropouts, for instance, then our approach cannot speak to the impact on college graduates.

Extrapolation to Additional Locations, Circumstances Perhaps the most ambitious external validity goals relate to extrapolation to different time periods or locations - in our example, to times with better labor market conditions or to different cities. Assuming the researcher has data on some observable characteristics in the two locations or time periods it is in principal possible to use a reweighting-based approach. However, in these cases the models developed in Section 3 do not apply, since the trial population is not a subset of the target population. Moreover, if we run our job training experiment in a large city and then want to extrapolate to a rural area it may be possible to match the rural population on age or education, but it seems difficult to develop intuition about the relationship between the observable selection and the unobservable selection, where the unobservable contains all unmeasured differences between the two locations. Bates and Glennerster (2017) provide a nuanced discussion of the extent to which one can port the results of randomized trials between locations within developing countries.

Below, we develop four examples which fit in the first two extrapolation categories. Where relevant, we highlight possible general equilibrium issues. In each case, the key empirical input is an adjustment for selection on observables. Although it is common to informally discuss the relationship between the sample and the overall population of interest in experimental settings, formal adjustments for differences in observable characteristics are less frequently considered. Implementing such adjustments requires observing features of the target population which can be matched to the trial population.

The first two examples below consider cases where we model selection as occurring on the treatment effect. In the last two examples, it seems more plausible to model selection as occurring on other features of the data.

5.2 Attanasio, Kugler and Meghir (2011)

Setting Attanasio et al (2011) report results from an evaluation of a job training program in Colombia. The program provided vocational training to poor men and women in several cities. We focus here on the results for women since there were concerns about the validity of the program randomization for men. The results show large positive impacts on employment, hours and days worked, and salaries for women.

The experimental sample consists of individuals who applied to be in the program at a number of program centers. In many cases more people applied to be in the program than there was space in the center, and the evaluation is based on randomizing program enrollment among eligible individuals who chose to apply.

Attanasio et al (2011) is representative of a broader class of papers in which participants volunteer for a study and treatment is randomized among volunteers. Examples include Gelber, Isen and Kessler (2016), also on job training, and Muralidharan et al (2017) on computer-based tutoring in India.

In the particular case of Attanasio et al (2011), a question of interest for policy is whether it would be a good idea to extend the vocational training program to all individuals - perhaps making it part of a school curriculum.[18] If the ATE estimated in this experiment is valid for such an expansion, the answer is likely yes. Given the selection procedure, however, it seems unlikely that the ATE for the experimental sample is representative of that for the population as a whole. In particular, individuals who select into the sample may be those who expect vocational training to work for them. The in-sample ATE could then be biased upwards relative to the full population ATE.

Target Population Data A key step in implementing our approach is to identify the target population of interest and to find a data source for comparable information on that group. In this case, a natural target population is all eligible individuals in the cities in question. In the original paper, the authors note that there is a nationally representative survey, the National Household Survey, which can be used to provide target population estimates. The

[18] This is an example of a setting where one may also want to consider the possible general equilibrium effects of a broad expansion; those effects will not be captured by our adjusted estimate. By contrast, such concerns would be less pressing if one instead considered an expansion to a small, randomly selected subset of the population.

authors provide some general comparisons to this population, but do not formally adjust for differences in population characteristics.

The program studied in Attanasio et al (2011) is generally not open to people with degrees beyond high school.[19] We therefore exclude individuals with more than a high school education from the target population. We also exclude from the analysis the small number of people in the trial population who report having more than a high school degree, who should not have been eligible (this is only 1% of the sample and makes little difference to the trial population results).[20]

Appendix Table 1 reports summary statistics on the target population and the experimental group. As noted in the original paper, the target population is slightly less educated and less likely to be employed, but similarly likely to have a formal contract conditional on employment. The differences in education reflect that a much larger share of the trial population has completed high school. This might argue for using a dummy for high school completion in our correction for selection on observables, rather than the mean and variance of education. In fact, the results are very similar under both approaches.

Results Table 4 implements our calculations for each of the primary outcomes reported in Table 4A of Attanasio et al (2011) - that is, the main results for women on which the authors focus.[21]

Column 2 shows the baseline effects, which are mostly significant and show better labor market outcomes for the treatment group. Column 3 shows the estimate after correcting for selection on observables as described in Section 4.1 above. This correction substantially attenuates the estimates; in some cases the adjusted effect is zero or negative. The primary reason is that there is substantial treatment effect heterogeneity on education. While the magnitude of the differences in education may seem fairly small, when scaled by the large degree of heterogeneity on this dimension the implied treatment effect difference is substantial.

[19] See http://www.dps.gov.co/que/jov/Paginas/Requisitos.aspx

[20] These individuals may have been included in error, have special circumstances, or have reported their education incorrectly.

[21] We implement this as described above, constructing T_i and regressing it on the covariates. A complication is that there was a variation across cities and programs in the share of people randomized into the treatment group. As the authors note, in most cases the shares were close to 50% (which is the average). If we observed the exact share in treatment for each course we could use that in the construction of T_i. This was used in a robustness check discussed in the original paper but we were unable to get the data from the authors. We therefore use 50%, but note that it is an approximation.

The results on increased wage and salary earnings are the least affected.

Columns 4 and 5 show two measures of external validity. First, Column 4 reports bounds on the target population ATE under the assumption that $\Phi \in [1, 2]$. For the most part these bounds are much less encouraging about the effectiveness of the program than are the baseline estimates. The only exception is earnings, where the impacts seem somewhat robust. Second, Column 5 shows the value of Φ corresponding to a zero ATE. These figures are, in some cases, less than 1 - this implies that the unobservables would have to operate in the opposite direction of the observables to produce an effect of zero.

Confidence sets are reported in Columns 4 and 5. In Column 4 these are generally large, corresponding to the relatively large adjustments. The confidence sets in Column 5, which are mostly infinite, illustrate the fairly poor inference properties of $\Phi(0)$ in this setting. As we discuss above this is a known issue, and is related to the problem of weak instruments.[22]

5.3 Bloom, Liang, Roberts, and Ying (2015)

Setting Bloom et al (2015) report results from an experiment in a Chinese firm designed to evaluate the productivity consequences of working from home. The firm operates a call center, so it is possible for workers to perform their duties from home.

The design of the experiment is as follows. First, workers at the firm were informed of the possibility of working from home and given an opportunity to volunteer for the program. Approximately 50% of them did so. Treatment was then randomized among eligible volunteers. Eligibility was enforced only after volunteering, and was based on several criteria including whether the individual had a private bedroom. The results suggest substantial productivity gains - about 0.2 standard deviations on a combined productivity measure - from working from home.

In this case, a question of interest for the firm may be whether it would be sensible to have many or all eligible call center employees work from home.[23] If the ATE estimated in the experiment is valid for the entire workforce, then the answer is likely yes. In fact, given the expense of running an office, this might be a good policy even if the ATE on productivity

[22] The confidence set we use here is asymptotically optimal, so the poor performance seems to reflect fundamental difficulties in conducting inference on $\Phi(0)$, rather than a poor choice of confidence set.

[23] Again, however, there could be additional impacts of such a major expansion which would require additional attention.

were zero or slightly negative.

Given the selection procedure it seems plausible that the ATE for the experimental sample is not representative of that for the population as a whole. Individuals may be more likely to select into the sample if they expect working from home to work for them. The in-sample ATE could therefore be biased upwards relative to the full population ATE.

Target Population Data It is straightforward to identify the target population for this study: it is all workers at the firm with private bedrooms.[24] Bloom et al (2015) collect some basic characteristics for this overall population of workers. These can then be compared to the volunteers.

Appendix Table 2 reports summary statistics in the overall population and experimental group. There are some differences: the volunteer group has a longer commute, is more likely to be male, and more likely to have children. As suggested above, when we correct for selection on observables we use these variables and allow them to enter linearly and (for non-binary variables) squared.

Results Table 5 shows results. Column 2 shows the baseline effect for the primary outcome in the paper, which is the increase in overall performance. Column 3 shows estimates from the regression-based correction for selection on observables. This slightly decreases the effect, from 0.22 to about 0.20.[25]

Columns 4 and 5 again show the two measures of external validity. Column (4) illustrates the bounds on the effect if we assume $\Phi \in [1, 2]$. The lower bound is still well above zero, and the confidence interval indicates a significant effect. Column 5 shows the value of Φ which corresponds to an ATE of zero; this figure is a bit above 12, implying that the unobservables would have to be substantially more important than the observables in order to deliver an ATE of zero in the population.

[24] Note that the restriction to private bedrooms arises because eligibility for the program is limited to this group. It is therefore appropriate to consider the target population as all eligible workers, rather than all workers.

[25] We implement this adjustment as described above, by regressing the constructed T_i on the observables. An alternative approach is to regress the outcome on covariates for the treatment group and the control group separately and difference the predicted values. Assuming successful randomization, these will yield similar results. In this case there is some imbalance across treatment and control in commute time - specifically, the treatment group has longer commutes on average than the control group. As a result, these two approaches yield slightly different coefficients. In Appendix Table 3 we report these results using the alternative approach.

5.4 Dupas and Robinson (2013)

Setting Our third application uses data from Dupas and Robinson (2013), who analyze the impact of informal savings technologies on investments in preventative healthcare and vulnerability to health shocks. The experiment, run in Kenya, includes four treatment arms, each of which provided a different technology (a safe box for money, a locked box, and two health-specific savings technologies). The outcomes include investments in health and measures of whether people have trouble affording medical treatments.

The experiment finds significant results for some combinations of outcomes and treatments. We focus on the combinations of outcomes and treatments which the authors suggest should be significant based on their theory. The first two columns of Table 6 list these combinations. Most of these effects are significant at conventional levels (see Table 3 in Dupas and Robinson (2013)).

The experiment was run through Rotating Savings and Credit Associations (ROSCAs), and participants were required to be enrolled in a ROSCA at the start.[26] External validity concerns again arise here because of the sampling frame: ROSCA participants are likely to be a selected group. Most notably, ROSCAs are designed in part as a savings and investment mechanism, so participants may differ on characteristics related to their responsiveness to savings products.

From a policy standpoint, however, there is interest in how to increase savings behaviors broadly, not just among ROSCA participants. We would therefore like to evaluate the external validity of these results relative to the overall population.

To frame this in our language, our concern is that there is some feature - say, interest in saving - which influences selection into the sample and also co-varies with the treatment effect. We observe some correlates of this feature, but there is further private information among the participants. The question is how important this private information would have to be in order to produce ATEs equal to zero. We can use our approach to calculate sensitivity values Φ_A for each outcome-treatment pair in the data. These can be interpreted as measuring how much of the covariance between the selection variable and the treatment effect would need to

[26] ROSCAs are informal savings groups common in many developing countries. Although the setup varies, typically these groups come together on a regular basis and contribute to a common pot of money which is taken home by one member on a rotating basis.

be due to private information in order to eliminate the result. As above, higher values point to a more robust result.

Target Population Data Column 1 of Appendix Table 4 shows summary statistics for the sample in Dupas and Robinson (2013). To perform an adjustment for selection on observables, it is necessary to observe the same variables for people who do not participate in ROSCAs. In an appendix to Dupas and Robinson (2013), the authors provide evidence on differences between ROSCA participants and non-participants using a second survey run in the same area. These differences can be used to construct population-level values for the covariates. These are shown in Column 2 of Appendix Table 4. We use these to adjust for selection on observables, where we allow the variables listed to enter linearly.

Results Table 6 shows the results. For most of the analyses adjustment for selection on observables moves the coefficient towards zero, suggesting the patterns of selection are such that those with larger treatment effects are more likely to be in the sample. However, there is substantial variation across the outcome-treatment pairs in the degree of sensitivity. For example, the relationship between the treatments and the variable measuring whether people have trouble affording treatment is fairly robust. The selection on observables adjustment is extremely small and in one case goes in the opposite direction, suggesting that adjustments for selection on observables actually increase the ATE. By contrast, the results for investment in health show larger adjustments.

These differences are reflected in the metrics of external validity in Columns 5 and 6. The bounds in Column 5 for the trouble affording treatment outcome generally remain close to the baseline effect. In contrast, the bounds for investments in health suggest less robust impacts.

5.5 Olken, Onishi, and Wong (2014)

Setting Olken et al (2014) report results from an experiment in Indonesia which provides block grants to villages to improve maternal health and child education. A subset of the grants include performance incentives, and the paper reports data on a wide variety of outcomes. The primary conclusion of the paper is that these grants have little or no effect on outcomes. The estimates are fairly small and mostly insignificant.

To implement the experiment, the government approached provinces, giving them the opportunity to take part. Five provinces volunteered to participate. Within these provinces, the richest 20% of districts were excluded from participation, as were the 28% of districts which did not have access to the rural infrastructure project through which the program was administered. Among the remaining districts, 20 were randomly selected, and sub-districts within these were eligible for the program if they were less than 67% urban. There were 300 eligible sub-districts and these were randomized into one of two treatment groups - with or without incentives - or the control group. The experimental sample is clearly not a simple random sample, and as the authors note the sub-districts eligible for inclusion in the experiment differ on some observable dimensions from the overall population.

Target Population Data To apply our approach, we need to identify a set of characteristics from the target population. The concern is that the sub-districts in the experiment are not representative of all of Indonesia. We therefore focus on sub-district-level characteristics. The data collected in the experiment did not include comparable information about the target population. However, we can extract these data from a nationally representative survey of Indonesia (SUSENAS) which we merge at the level of the sub-district with the data used in Olken et al (2014). The target population corresponds to all of Indonesia.[27] This is an example of how our approach might be used in a setting like this, where an experiment includes a subset of locations within a country or region, and external data is available for the entire region.

Results Appendix Table 5 shows summary statistics both for Indonesia overall and for the sub-districts in the study. Relative to the country overall, households in districts in the sample are more likely to have a dirt floor and to receive cash transfers (consistent with having lower income on average) but also have higher rates of vaccination and contraceptive use.

Table 7 shows the results. As noted, the baseline impact is insignificant for most outcomes. However, a notable feature of this setting is that in all cases but one correction for selection on observables increases the estimated size of the effect. Consequently, most of the sensitivity

[27] The set of covariates we use do not include those on which the sample is constructed, so the common support assumption remains plausible here. For example, as shown in Appendix Table 5 the differences between the means of the covariates in the sample and target population are of the same order as the variability within the sample.

measures in Column 5 are negative. Under our baseline assumptions, these results suggest that the effects in the trial population may actually *understate* the overall effects in the target population in many cases.

This is made most concrete by Column 4 of Table 7, which shows bounds under the assumption that $\Phi_A \in [1, 2]$. For all of the outcomes, the bounds are substantially more encouraging about the impact of the experiment than are the baseline effects. Based on the confidence intervals, many of these adjusted effects are significantly different from zero. In this case, our analysis casts doubt on the conclusion that this intervention does not change outcomes. It may simply be that the population used for the trial is not the one for which this intervention was most effective.

6 Discussion

While our primary focus in this paper is on external validity of ATEs estimated from randomized trials, one could potentially apply analogous approaches in regression discontinuity and instrumental variables settings. In this section we briefly discuss these possibilities, as well as application of our results to estimate non treatment-effect moments in the target population.

Regression Discontinuity Regression discontinuity estimates are identified from behavior at the discontinuity; this leads to concern that treatment effects may differ for individuals distant from the discontinuity (Bertanha and Imbens, 2014; Angrist and Rokkanen, 2015; Rokkanen, 2015). Consider a sharp RD design with running variable R_i for individual i, where $D_i = 1\{R_i \geq r^*\}$ is an indicator for R_i exceeding some threshold r^*. The regression discontinuity approach estimates the treatment effect by a regression of Y_i on D_i in a small neighborhood of R_i around r^*. We can define the observations in an infinitesimal neighborhood of r^* as the trial population. The target population is the population for the full range of $R_{i.}$ We can then treat this problem as in the experimental case above.[28] Note, however, that relative to approaches proposed in the literature, our approach does not exploit additional structure from the regression discontinuity setting and so may yield less precise results.

[28] For our absolute continuity assumption (Assumption 1) to hold, X_i must not include R_i.

Instrumental Variables The central component of the LATE critique is that instrumental variables approaches identify the ATE and other quantities only in the population of compliers, which may differ from the population of interest. In the language of this paper, we can define P_S as the distribution in the population of compliers and P as the distribution in the overall population, including compliers, never takers and always takers. It is then possible to proceed in the same way as above. Unlike recent work on external validity in instrumental variables models by Kowalski (2016) and Brinch et al *(*2017), however, our approach again does not exploit the additional structure imposed by the instrumental variables setting, and so again may yield less precise results.

Non-Treatment Effect Moments We focus on cases where the unknown moment of interest in the target population is an ATE. However, as should be clear from the development of the theory in Section 2, our approach is not limited to estimating ATEs. Of particular interest may be cases where the object of interest is the mean of some variable in the target population.

An example of this sort is polling data: surveys collect voting intentions in a trial population and the object of interest is the voting intentions in a target population. It is common to reweight polling data to match observable demographics in the target population. Our approach could be used in concert with such reweighting to think systematically about possible selection on unobservables (for example: people who respond to polling calls may be more passionate about the election, or have a lower value of time).

7 Conclusion

This paper considers the problem of external validity when the trial population for a study differs from the target population of interest. We focus on the case where the trial population is selected, at least in part, on characteristics which are unobserved by the researcher. We analyze this problem through the lens of reweighting. We show that this framework can be used to bound the target population moments under assumptions about the intensity of selection on the portion of the treatment effect explained by observables relative to the unobservable.

Our approach is straightforward to implement. The only added data requirement above what would be used in the main analysis in a paper is knowledge of some characteristics of the target population. In many cases we could use, for example, demographic variables, where the moments in the target population are available from standard public datasets. In designing experiments going forward the range of application for this technique might be improved by either collecting some minimal data on a target population or by structuring data collection in the trial population to ensure comparability with known features of the target population.

References

Altonji, Joseph G., Timothy Conley, Todd E. Elder, and Christopher R. Taber, "Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables," 2010. Unpublished Manuscript.

____ **, Todd E. Elder, and Christopher R. Taber**, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

Anderson, T. W. and Herman Rubin, "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 1949, *20* (1), 46–63.

Angrist, Joshua D. and Miikka Rokkanen, "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff," *Journal of the American Statistical Association*, 2015, *110* (512), 1331–1344.

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.

Bates, Mary Ann and Rachel Glennester, "The Generalizability Puzzle," 2017.

Bertanha, Marinho and Guido W. Imbens, "External Validity in Fuzzy Regression Discontinuity Designs," Working Paper 20773, National Bureau of Economic Research December 2014.

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.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall, "Beyond LATE with a discrete instrument. Heterogeneity in the quantity-quality interaction of children," *Journal of Political Economy*, 2017, *125* (4), 985–1039.

Bugni, Federico, Ivan Canay, and Azeem Shaikh, "Inference under Covariate-Adaptive Randomization," 2017. Working Paper.

Chyn, Eric, "Moved to Opportunity: The Long-Run Effect of Public Housing Demolition on Labor Market Outcomes of Children," 2016.

Cole, Stephen R. and Elizabeth A. Stuart, "Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial," *American Journal of Epidemiology*, 2010, *172* (1), 107–115.

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.

Dupas, Pascaline and Jonathan Robinson, "Why don't the poor save more? Evidence from health savings experiments," *The American Economic Review*, 2013, *103* (4), 1138–1171.

Feller, Avi, Todd Grindal, Luke W. Miratrix, and Lindsay C. Page, "Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Settings," *Annals of Applied Statistics*, 2016.

Fieller, E. C., "Some problems in interval estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1954, *16* (2), 175–185.

Gechter, Michael, "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India," *manuscript, Pennsylvania State University*, 2015.

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.

Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 1998, *66*, 315–331.

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S. Sekhon, "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2015, *178* (3), 757–778.

Hellerstein, Judith K and Guido W Imbens, "Imposing moment restrictions from auxiliary data by weighting," *Review of Economics and Statistics*, 1999, *81* (1), 1–14.

Hirano, Keisuke, Guido Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, *71*, 1161–1189.

Horvitz, Daniel G and Donovan J Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association*, 1952, *47* (260), 663–685.

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.

Imai, Kosuke and Marc Ratkovic, "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014, *76* (1), 243–263.

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

Imbens, Guido W and Joshua D Angrist, "Identification and estimation of local average treatment effects," *Econometrica*, 1994, *62* (2), 467–475.

Kline, Patrick and Christopher Walters, “Evaluating Public Programs with Close Substitutes: The Case of Head Start,” *Quarterly Journal of Economics*, 2016, *131* (4), 1795–1848.

Kowalski, Amanda E, “Doing More When You’re Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments,” 2016.

Muraldiharan, Karthik, Abhijeet Singh, and Alejandro Ganimian, “Teaching All Students, and Not Just the Top of the Class: Experimental Evidence on Technology-led Education in India,” 2016.

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

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.

Oster, Emily, “Unobservable Selection and Coefficient Stability: Theory and Validation,” *Journal of Business Economics and Statistics*, Forthcoming.

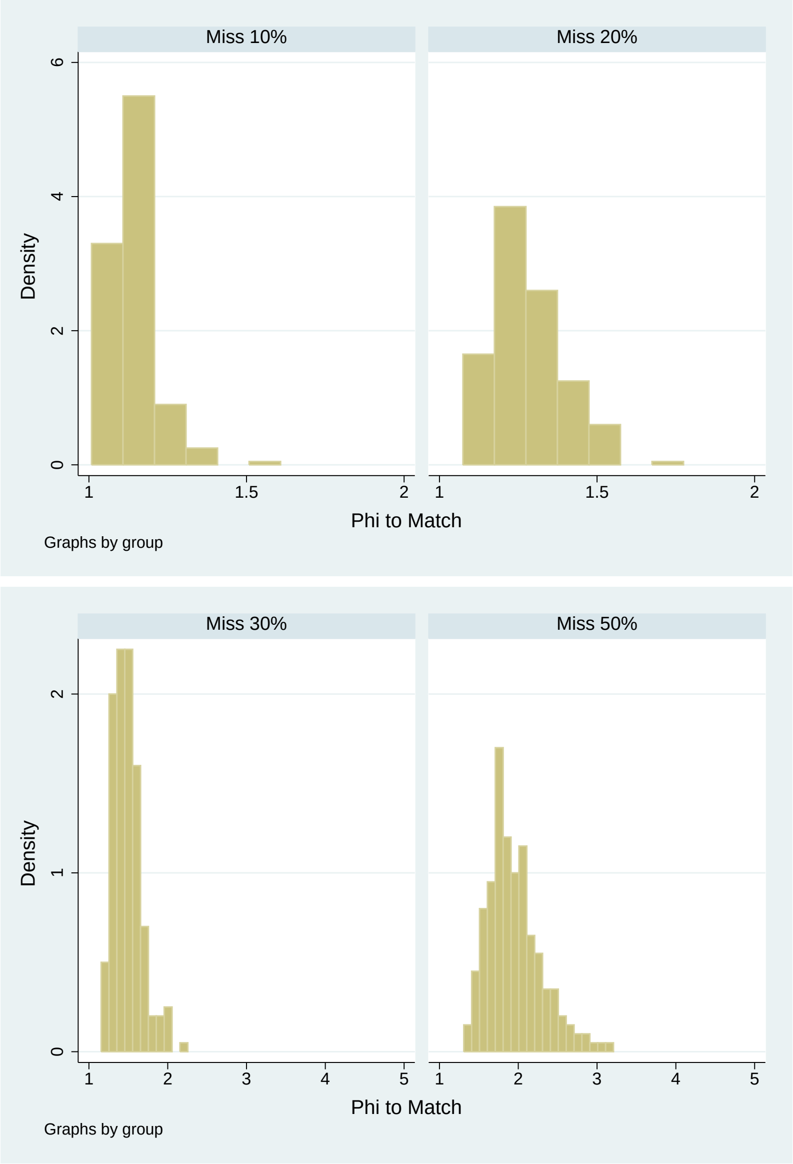
Rokkanen, Miikka, “Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design,” 2015. Working Paper.

Rosenbaum, Paul R, *Observational Studies*, Springer, 2002.

Rosenbaum, Paul R. and Donald B. Rubin, “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 1983, *70* (1), 41–55.

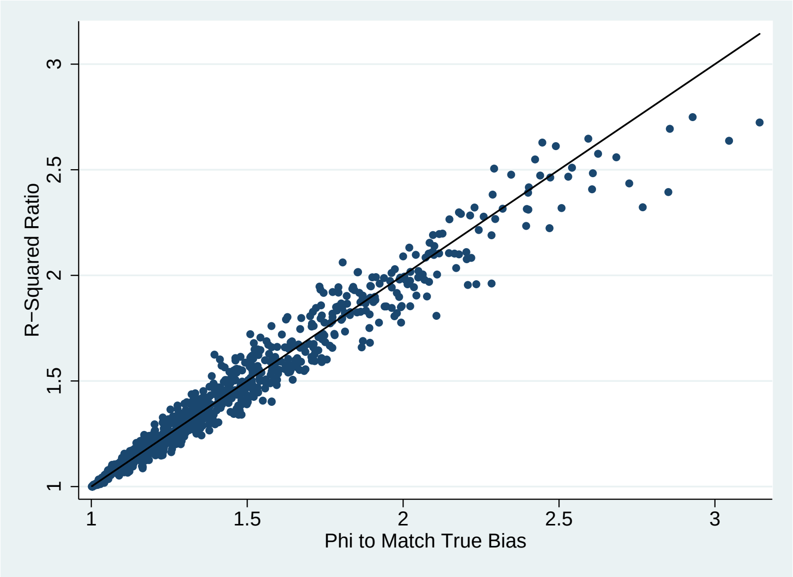
Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf, “The use of propensity scores to assess the generalizability of results from randomized trials,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2011, *174* (2), 369–386.

Figure 1: **Histogram of Values of Φ to Match Population Effect**



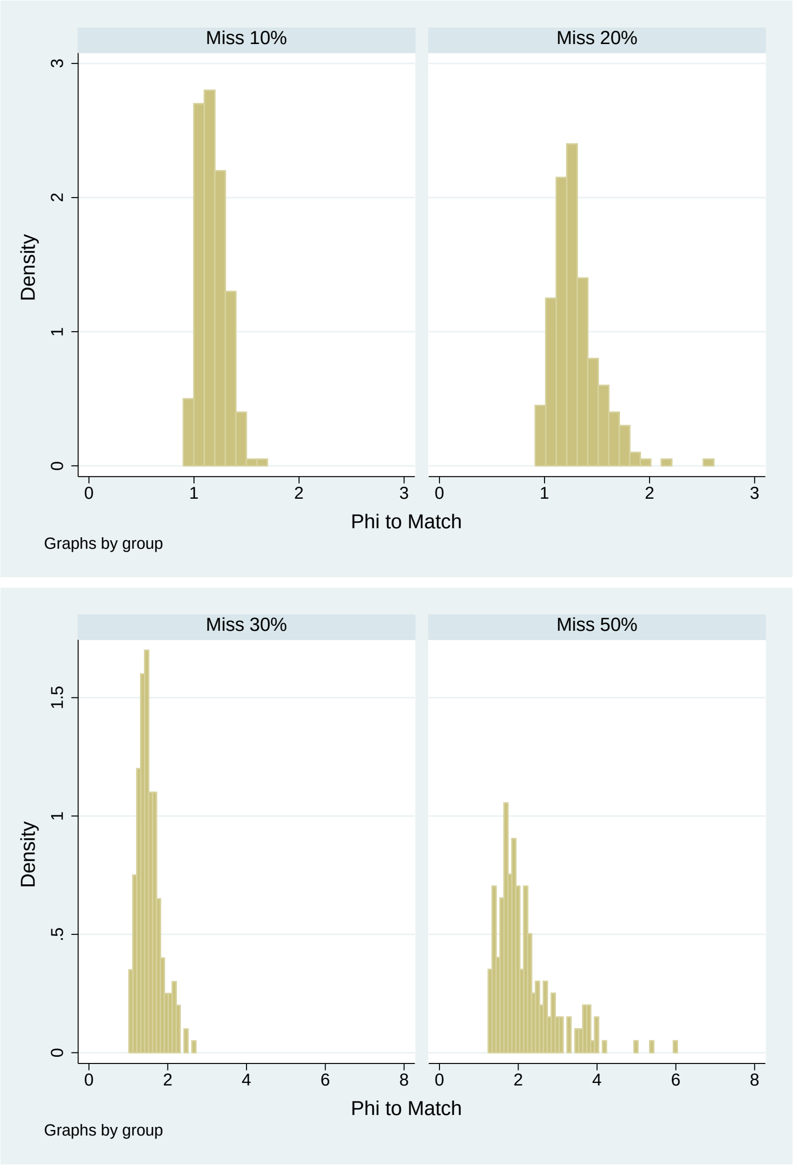
Notes: This figure shows the values of Φ which would match the population effect in the example based on Muraldiharan and Sundararaman (2011) with varying sets of covariates treated as unobserved. In this example the data are selected on the predicted treatment effect, where the prediction is constructed using observables and unobservables.

Figure 2: **Relationship between Φ and R-Squared**



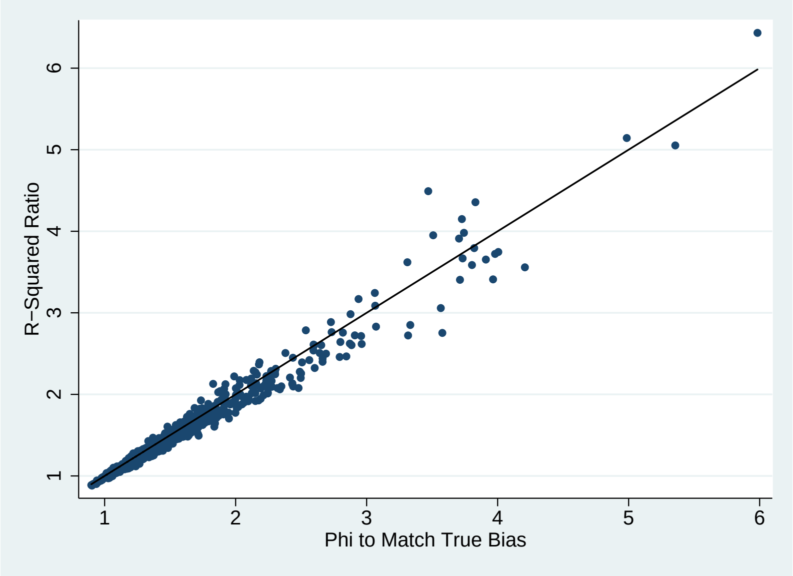
Notes: This figure shows the relationship between the values of Φ to match the population effect and the relative R-squared in a regression of the treatment effect on all variables in the example based on Muraldiharan and Sundararaman (2011) with varying sets of covariates treated as unobserved. In this example the data are selected on the predicted treatment effect, where the prediction is constructed using observables and unobservables. The 45 degree line is plotted in black.

Figure 3: **Histogram of Values of Φ_A to Match Population Effect**



Notes: This figure shows the values of Φ_A which would match the population effect in the example based on Muraldiharan and Sundararaman (2011) with varying sets of covariates treated as unobserved. In this example the data are selected on mandal-level average teacher training.

Figure 4: **Relationship between Φ_A and Covariance Ratio**



Notes: This figure shows the relationship between the values of Φ_A to match the population effect and the relative R-squared in a regression of the treatment effect on all variables in the example based on Muraldiharan and Sundararaman (2011) with varying sets of covariates treated as unobserved. In this example the data are selected on mandal-level average teacher training. The 45 degree line is plotted in black.

Table 1: **Bias Decomposition**

Variable	Trial Pop. Mean	Target Pop. Mean	$\sigma_{P_S}(f_j(X_i))$	$\rho_{P_S}(W_i, f_j(X_i))$	$\sigma_{P_S}(W_i)$
White (0/1)	0.59	0.80	0.492	0.519	0.815
HS Completion (0/1)	0.84	0.89	0.363	0.163	0.815
Male (0/1)	0.50	0.51	0.50	0.017	0.815

Notes: This table illustrates the bias decomposition in the NLSY. $\sigma_{P_S}(f_j(X_i))$ is the standard deviation of the moments, $\rho_{P_S}(W_i, f_j(X_i))$ is the correlation between the weights and the moments and $\sigma_{P_S}(W_i)$ is the standard deviation of the weights.

Table 2: **Relative Selection on Additional Moments**

			Relative Selection Ratio for Comparison With:		
Additional Variable	Trial Pop. Mean	Target Pop. Mean	White	HS Completion	Male
Log Hourly Wage	1.60	1.64	0.15	0.48	4.72
College Completion (0/1)	0.23	0.29	0.36	1.15	11.18
AFQT Score	41.0	48.1	0.59	1.87	18.15

Notes: This table illustrates the difference between trial and target population on three additional variables in the NLSY. The relative selection ratio (defined in equation (6)) is the ratio of the standardized bias on each additional variable relative to that on the initial benchmark variables.

Table 3: **Auxiliary Evidence, Selection Models with Varying Exclusion Sets**

Panel A: Select on Treatment Effect		
	Average Selection-on-Obs. Effect	Average Φ
Exclude 50%	0.111	2.05
Exclude 30%	0.098	1.49
Exclude 20%	0.090	1.27
Exclude 10%	0.083	1.14
Exclude only one covariate	0.076	1.02
Panel B: Select on Mandal Teacher Training		
	Average Selection-on-Obs. Effect	Average Φ_A
Exclude 50%	0.092	2.09
Exclude 30%	0.087	1.56
Exclude 20%	0.082	1.31
Exclude 10%	0.079	1.16
Exclude only one covariate	0.075	1.04

Notes: This table illustrates the evidence from the constructed example in Section 4. The sample is selected based either on the predicted treatment effect (Panel A) or the Mandal-level average of teacher training (Panel B). We then calculate the average value for Φ or Φ_A which would match the target population treatment effect, treating different sets of the covariates as unobserved.

Table 4: **Application: Attanasio et al(2011)**

Outcome	Baseline Effect	Observable Adjusted	Bounds, $\Phi \in [1, 2]$	$\Phi(0)$
(1)	(2)	(3)	(4)	(5)
Employment	0.062	-0.007	[-0.076, -0.007]	0.89
	(0.02,0.11)	(-0.15, 0.14)	(-0.31, 0.16)	(-∞,∞)
Paid Employment	0.056	-0.007	[-0.071, -0.007]	0.88
	(0.01, 0.10)	(-0.14, 0.13)	(-0.30, 0.16)	(-∞,∞)
Days Worked in Last Month	1.53	0.13	[-1.26, 0.13]	1.09
	(0.39, 2.68)	(-3.06, 3.33)	(-6.47, 3.94)	(-∞,∞)
Hours/Week	3.46	0.51	[-2.45, 0.51]	1.17
	(0.82,6.10)	(-7.29, 8.31)	(-14.7, 9.8)	(-∞,∞)
Job Tenure	-1.30	-0.75	[-0.75,-0.20]	2.37
	(-2.48,-0.17)	(-3.49, 1.98)	(-4.56, 4.15)	(-∞,∞)
Wage and Salary Earnings	31,116	24,336	[17,555, 24,336]	4.58
	(14,104, 48,129)	(-4677, 53,350)	(-27,566, 62,678)	$(-\infty, -1.6] \cup [0.9, \infty)$
Self-Employment Earnings	5213	-2194	[-9603, -2194]	0.70
	(-9982, 20,410)	(-33,603, 29,214)	(-59,518, 40,311)	(-∞,∞)

Notes: This table shows the application of our sensitivity procedure to Attanasio et al (2011). The target population moments are generated using a nationally representative survey of the same areas in which the study was run. Analytic and bootstrap confidence intervals are reported in Columns (2) and (3), respectively, while the confidence sets in Columns (4) and (5) are computed as described in Appendix B.3, with simulation-based critical values c_α^* used in Column (4).

Table 5: **Application: Bloom et al (2015)**

Outcome	Baseline Effect	Observable Adjusted	Bounds, $\Phi \in [1, 2]$	$\Phi(0)$
(1)	(2)	(3)	(4)	(5)
Job Performance	0.222	0.204	[0.185, 0.204]	12.08
	(0.172, 0.272)	(0.149, 0.258)	(0.125, 0.252)	$(-\infty, -39.2] \cup [5.3, \infty)$

Notes: This table shows the application of our sensitivity procedure to Bloom et al (2015). The target population moments comes from the study. Analytic and bootstrap confidence intervals are reported in Columns (2) and (3), respectively, while the confidence sets in Columns (4) and (5) are computed as described in Appendix B.3, with simulation-based critical values c_α^* used in Column (4).

Table 6: **Application: Dupas and Robinson (2013)**

Outcome	*Treatment*	Baseline Effect	Observable Adjusted	Bounds, $\Phi_A \in [1, 2]$	$\Phi_A(0)$
(1)	(2)	(3)	(4)	(5)	(6)
Investment in Health	Safe Box	165.9	85.17	[4.44, 85.17]	2.05
		(12.1, 319.6)	(-65.8,236.1)	(-151.4, 217.2)	(0.52, 15.5)
Investment in Health	Locked Box	48.33	15.05	[-18.2, 15.1]	1.45
		(-56.1, 152.8)	(-93,123.9)	(-130.9, 111.3)	(-∞,∞)
Investment in Health	Health Pot	287.8	150.1	[12.4, 150.1]	2.09
		(121.8, 453.8)	(-33.7, 334.0)	(-186.6, 300.3)	(1.15, 7.65)
Trouble Affording Treat.	Safe Box	-0.111	-0.142	[-0.172, -0.141]	-3.58
		(-0.250, 0.028)	(-0.285, -0.009)	(-0.339, -0.006)	(-∞,∞)
Trouble Affording Treat.	Health Savings	-0.134	-0.134	[-0.134, -0.134]	685.8
		(-0.268, 0.0001)	(-0.266, -0.001)	(-0.272, 0.005)	(-∞,∞)
Reached Health Goal	Safe Box	0.155	0.113	[0.070, 0.112]	3.64
		(0.002, 0.309)	(-0.067, 0.284)	(-0.116, 0.266)	(-∞,∞)
Reached Health Goal	Locked Box	-0.020	-0.029	[-0.038, -0.029]	-2.18
		(-0.159, 0.118)	(-0.157, 0.098)	(-0.174, 0.096)	(-∞,∞)
Reached Health Goal	Health Pot	0.120	0.059	[-0.0005, 0.059]	1.99
		(-0.034, 0.275)	(-0.12, 0.24)	(-0.213, 0.223)	$(-\infty, -3.5] \cup [-0.6, \infty)$
Reached Health Goal	Health Savings	0.056	0.045	[0.034, 0.045]	5.34
		(-0.097, 0.209)	(-0.091, 0.182)	(-0.110, 0.184)	(-∞,∞)

Notes: This table shows the application of our sensitivity procedure to Dupas and Robinson (2013). The target population moments are generated using evidence from an auxiliary survey measuring differences between participants and non-participants. Analytic and bootstrap confidence intervals are reported in Columns (3) and (4), respectively, while the confidence sets in Columns (5) and (6) are computed as described in Appendix B.3, with simulation-based critical values c^*_α used in Column (5).

Table 7: **Application: Olken et al (2014)**

Outcome	Baseline Effect	Observable Adjusted	Bounds, $\Phi_A \in [1, 2]$	$\Phi_A(0)$
(1)	(2)	(3)	(4)	(5)
Prenatal Visits	0.198	1.51	[1.51,2.83]	-0.15
	(-0.505, 0.902)	(0.72, 2.30)	(0.67, 4.13)	(-0.64, 0.35)
Assisted Delivery	0.008	0.119	[0.11, 0.231]	-0.067
	(-0.074, 0.089)	(0.021, 0.217)	(0.027,0.372)	(-0.56, 0.43)
Postnatal Visits	-0.197	0.059	[0.059,0.316]	0.768
	(-0.44, 0.048)	(-0.29,0.41)	(-0.24, 0.78)	(0.25, 8.75)
Iron Pills	0.045	0.284	[0.284,0.524]	-0.191
	(-0.137, 0.229)	(0.031, 0.538)	(0.067, 0.857)	(-1.18, 0.32)
Immunization	0.004	0.102	[0.102,0.20]	-0.040
	(-0.054, 0.062)	(0.023, 0.181)	(0.031, 0.305)	(-0.55, 0.44)
No. Weight Checks	0.147	0.419	[0.419,0.692]	-0.54
	(-0.009, 0.304)	(0.199, 0.640)	(0.223,0.990)	(-1.53, -0.03)
Vitamin A Supplements	0.015	0.185	[0.185,0.335]	-0.089
	(-0.148, 0.179)	(-0.026, 0.397)	(-0.005,0.636)	(-1.58, 0.91)
Malnourished	0.002	0.016	[0.016,0.032]	-0.117
	(-0.026, 0.030)	(-0.019,0.053)	[(-0.019, 0.083)	$(-\infty,\infty)$

Notes: This table shows the application of our sensitivity procedure to Olken et al (2014). The target population moments are generated using location-level variables from a nationally representative survey (SUSENAS). Analytic and bootstrap confidence intervals are reported in Columns (2) and (3), respectively, while the confidence sets in Columns (4) and (5) are computed as described in Appendix B.3, with simulation-based critical values c^*_α used in Column (4).

Appendix A: Proofs

Proof of Lemma 1 This result is immediate from Bayes Theorem. Note, in particular, that for any measurable set $\mathcal{A}$,

$$Pr_{P_S}\left\{X_i \in \mathcal{A}\right\} = Pr_P\left\{X_i \in \mathcal{A}|S_i = 1\right\} = \int_{\mathcal{A}} p_X\left(x|S_i = 1\right)d\mu$$

while by Bayes Theorem we can take

$$p_X\left(x|S_i = 1\right) = \frac{E_P\left[S_i|X_i = x\right]}{E_P\left[S_i\right]}p_X\left(x\right).$$

Thus,

$$Pr_{P_S}\left\{X_i \in \mathcal{A}\right\} = \int_{\mathcal{A}} \frac{E_P\left[S_i|X_i = x\right]}{E_P\left[S_i\right]}p_X\left(x\right)d\mu.$$

□

Proof of Lemma 2 We have assumed that P_X is absolutely continuous with respect to $P_{X,S}$, and the density of P_X with respect to $P_{X,S}$ is given by $\frac{p_X}{p_{X,S}}$. The result follows immediately. □

Proof of Corollary 1 By the definition of the covariance,

$$\begin{gathered} E_P\left[f\left(X_i\right)\right] = E_{P_S}\left[W_i f\left(X_i\right)\right] \\ = Cov_{P_S}\left(f\left(X_i\right), W_i\right) + E_{P_S}\left[f\left(X_i\right)\right]E_{P_S}\left[W_i\right]. \end{gathered}$$

As noted in the text, however, $E_{P_S}\left[W_i\right] = 1$ by Lemma 2, so the result follows. □

Proof of Lemma 3 Applying Lemma 1 conditional on (C_i, U_i), we know that the weights to rebalance the conditional distribution of TE_i are

$$W_i = \frac{E_P\left[S_i|C_i, U_i\right]}{E_P\left[S_i|C_i, U_i, TE_i\right]} = \frac{E_P\left[S_i|C_i, U_i\right]}{E_P\left[S_i|C_i, U_i\right]} = 1,$$

where in the second equality we have used the fact that S_i is independent of TE_i given (C_i, U_i). Thus, since $E_P\left[TE_i|C_i, U_i\right] = E_{P_S}\left[W_i TE_i|C_i, U_i\right]$ by Lemma 2 applied conditional on (C_i, U_i), the result follows immediately. □

Proof of Lemma 4 By Lemma 1, we know that the weights to rebalance $\left(C_i, \widetilde{TE}_i\right)$ are

$$W_i = \frac{E_P\left[S_i\right]}{E_P\left[S_i|C_i, \widetilde{TE}_i\right]} =$$

$$\frac{1 - Pr_P\left\{V_i \leq c \cdot \widetilde{TE}_i\right\}}{1 - Pr_P\left\{V_i \leq c \cdot \widetilde{TE}_i|C_i, \widetilde{TE}_i\right\}} = \frac{1 - Pr_P\left\{V_i \leq c \cdot \widetilde{TE}_i\right\}}{1 - Pr_P\left\{V_i \leq c \cdot \widetilde{TE}_i|\widetilde{TE}_i\right\}},$$

where in the second equality we have used the assumption of selection on the treatment effect. Since we have assumed that V_i is continuously distributed it follows from Assumption 2 that this is a continuously differentiable function of $\widetilde{TE}_i$. □

Proof of Proposition 1 Note that

$$\frac{E_P\left[TE_i\right] - E_{P_S}\left[TE_i\right]}{E_P\left[\widehat{TE}_i\right] - E_{P_S}\left[TE_i\right]} = \frac{E_P\left[\widetilde{TE}_i\right] - E_{P_S}\left[\widetilde{TE}_i\right]}{E_P\left[\widehat{TE}_i\right] - E_{P_S}\left[\widehat{TE}_i\right]} = \frac{Cov_{P_S}\left(W_i, \widetilde{TE}_i\right)}{Cov_{P_S}\left(W_i, \widehat{TE}_i\right)},$$

where the first equality follows from the law of iterated expectations and Lemma 3. Applying approximation (9) in the main text, we see that

$$\frac{Cov_{P_S}\left(W_i, \widetilde{TE}_i\right)}{Cov_{P_S}\left(W_i, \widehat{TE}_i\right)} \approx \frac{Cov_{P_S}\left(W_i^*, \widetilde{TE}_i\right)}{Cov_{P_S}\left(W_i^*, \widehat{TE}_i\right)} = \frac{Cov_{P_S}\left(W_i^*, \widetilde{T}_i\right)}{Cov_{P_S}\left(W_i^*, \widehat{T}_i\right)},$$

where the second equality again follows from the law of iterated expectations. Note, however, that

$$\frac{Cov_{P_S}\left(W_i^*, \widetilde{TE}_i\right)}{Cov_{P_S}\left(W_i^*, \widehat{TE}_i\right)} = \frac{Cov_{P_S}\left(\widetilde{TE}_i, \widetilde{TE}_i\right)}{Cov_{P_S}\left(\widetilde{TE}_i, \widehat{TE}_i\right)} = \frac{Var_{P_S}\left(\widetilde{TE}_i\right)}{Var_{P_S}\left(\widehat{TE}_i\right)},$$

where the first equality follows from the definition of W_i^* while the second again follows from the law of iterated expectations. □

Proof of Lemma 5 Follows by the same argument as Lemma 3. □

Proof of Lemma 6 Follows by the same argument as Lemma 4. □

Proof of Proposition 2 Follows by the same argument as Proposition 1. □

Appendix B: Additional Results

This appendix details several results mentioned in the main text. We first provide formal justification for the approximate weights W_i^* used in equation (9) of the main text, and show that the error in this approximation vanishes when selection is close to random. We then discuss a model, mentioned in Section 4.1 of the main text, under which Φ and Φ_A and can be interpreted as the share of relevant factors captured by the observed covariates. Finally, we provide additional details of and justification for our inference procedures.

Appendix B.1 Justification of Approximate Weights W_i^*

In the main text we claim that the error from using the approximate weights vanishes when we consider small values of c, so selection is nearly random. In this section we formalize this claim under regularity conditions. Without loss of generality, we focus on the case where selection is on $\widetilde{A}_i$ (since we can recover the treatment effects case by setting $\widetilde{TE}_i = \widetilde{A}_i$). Formally, we assume:

Assumption 5 *1. The density f_V of V_i is Lipschitz with Lipschitz constant K, and $f_V(0) > 0$.*

2. The support of $\widetilde{A}_i$ is bounded.

3. $E_P\left[f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)^2\right]$ is finite.

Under this assumption, we obtain the following result:

Proposition 3 *Under Assumptions 2, 4, and 5, as $c \to 0$,*

$$Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) = Cov_{P_S}\left(W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) + O\left(c^2\right).$$

Proof of Proposition 3

To show this result, recall note the proof of Lemma 1, generalized to selection on $\widetilde{A}_i$, shows that

$$W_i = \frac{1 - Pr_P\left\{V_i \le c \cdot \widetilde{A}_i\right\}}{1 - Pr_P\left\{V_i \le c \cdot \widetilde{A}_i | \widetilde{A}_i\right\}}.$$

Letting F_V denote the cdf of V_i, we can re-write this as

$$W_i = \frac{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}{1 - F_V\left(c \cdot \widetilde{A}_i\right)},$$

where only the denominator depends on $\widetilde{A}_i$.

Mean-Value Expansion of W_i: Let us consider a mean-value expansion of W_i around $E_{P_S}[A_i]$:

$$W_i = \frac{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}{1 - F_V\left(c \cdot E_{P_S}[A_i]\right)} + \frac{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}{\left(1 - F_V\left(c \cdot A_i^*\right)\right)^2} c \cdot f_V\left(c \cdot A_i^*\right)\left(\widetilde{A}_i - E_{P_S}[A_i]\right),$$

for A_i^* a value between $E_{P_S}[A_i]$ and A_i. Note that W_i^* is of the same form, but substitutes $E_{P_S}[A_i]$ for A_i^*. Since V_i is continuously distributed, for any $\varepsilon > 0$ there exists c_ε such that for all $c \in [0, c_\varepsilon]$,

$$Pr_P\left\{F_V\left(c \cdot A_i\right) \in \left[F_V(0) - \varepsilon, F_V(0) + \varepsilon\right]\right\} = 1.$$

Thus, for such c we know that

$$\frac{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}{\left(1 - F_V\left(c \cdot A_i^*\right)\right)^2} \leq \frac{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}{\left(1 - F_V(0) - \varepsilon\right)^2}.$$

If we consider the difference

$$W_i - W_i^* =$$

$$\left(\frac{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}{\left(1 - F_V\left(c \cdot A_i^*\right)\right)^2} c \cdot f_V\left(c \cdot A_i^*\right) - \frac{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}{\left(1 - F_V\left(c \cdot E_{P_S}[A_i]\right)\right)^2} c \cdot f_V\left(c \cdot E_{P_S}[A_i]\right)\right)\left(\widetilde{A}_i - E_{P_S}[A_i]\right),$$

the fact that $f_V(v)$ is Lipschitz implies that for $c \in [0, c_\varepsilon]$ the difference is bounded in absolute value by

$$\frac{1}{\left(1 - F_V(0) - \varepsilon\right)^2} c^2 K \left(\widetilde{A}_i - E_{P_S}[A_i]\right)^2 = c^2 K^* \left(\widetilde{A}_i - E_{P_S}[A_i]\right)^2,$$

for a constant K^*. Thus, we see that

$$\left|W_i - W_i^*\right| \leq c^2 K^* \left(\widetilde{A}_i - E_{P_S}[A_i]\right)^2.$$

Next, for some function $f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)$, let us consider the approximation error

$$Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) - Cov_{P_S}\left(W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) = Cov_{P_S}\left(W_i - W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)$$

$$= E_{P_S}\left[\left(W_i - W_i^*\right) f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right] - E_{P_S}\left[W_i - W_i^*\right] E_{P_S}\left[f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right].$$

Using our bounds above, for $c \in [0, c_\varepsilon]$ the first term is bounded in absolute value by

$$c^2 K^* \cdot E_{P_S}\left[\left(\widetilde{A}_i - E_{P_S}[A_i]\right)^2 \left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right],$$

while the second is bounded by $c^2 K^* \cdot E_{P_S}\left[\left(\widetilde{A}_i - E_{P_S}[A_i]\right)^2\right] \left|E_{P_S}\left[f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right]\right|$.

This almost completes the argument, except that these terms we have used as bounds themselves depend on c, since they are calculated in the target population. Thus, we next show that these terms are well-behaved for small c.

Behavior of moments for small c Note that

$$E_{P_S}\left[\left(\widetilde{A}_i - E_{P_S}\left[A_i\right]\right)^2 \left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right] = E_P\left[W_i^{-1}\left(\widetilde{A}_i - E_{P_S}\left[A_i\right]\right)^2 \left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right]$$

$$= E_P\left[\left(\frac{1 - F_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}\right)\left(\widetilde{A}_i - E_{P_S}\left[A_i\right]\right)^2 \left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right].$$

Since

$$E_P\left[\left(\frac{1 - F_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]} - 1\right)^2\right] \to 0$$

as $c \to 0$, the Cauchy-Schwarz inequality implies that

$$E_P\left[\left(\frac{1 - F_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]} - 1\right)\left(\widetilde{A}_i - E_{P_S}\left[A_i\right]\right)^2 \left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right] \to 0,$$

and thus that

$$E_{P_S}\left[\left(\widetilde{A}_i - E_{P_S}\left[A_i\right]\right)^2 \left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right] \to E_P\left[\left(\widetilde{A}_i - E_P\left[A_i\right]\right)^2 \left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right]$$

as $c \to 0$. Under our assumptions, we can likewise show that

$$E_{P_S}\left[\left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right] \to E_P\left[\left|f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right|\right]$$

and

$$E_{P_S}\left[\left(\widetilde{A}_i - E_{P_S}\left[A_i\right]\right)^2\right] \to E_P\left[\left(\widetilde{A}_i - E_P\left[A_i\right]\right)^2\right]$$

as $c \to 0$.

Completing the argument: Combing these results, we see that under our assumptions above,

$$Cov_{P_S}\left(W_i - W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) = O\left(c^2\right)$$

as $c \to 0$. □

Using this result, we can show that the approximation error from using W_i^* instead of W_i is of lower order than the bias as $c \to 0$.

Corollary 2 *Under Assumptions 2, 4, and 5, if* $Cov_P\left(\widetilde{A}_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) \neq 0$ *then*

$$\frac{Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)}{Cov_{P_S}\left(W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)} \to 1$$

as $c \to 0$.

Proof of Corollary 2: By Proposition 3 we know that

$$Cov_{P_S}\left(W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) = Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) + O\left(c^2\right),$$

and thus that

$$\frac{Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)}{Cov_{P_S}\left(W_i^*, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)} = \frac{Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)}{Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) + O\left(c^2\right)}.$$

Next, note that by Corollary 1,

$$Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) = E_P\left[f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right] - E_{P_S}\left[f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right],$$

which as in the proof of Proposition 3 we can re-write as

$$E_P\left[\left(1 - \frac{1 - F_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]}\right) f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right].$$

Note that the assumption that $f_V(\cdot)$ is Lipschitz, together with the fact that it is positive and integrates to one, implies that it is bounded. The dominated convergence theorem thus implies that $\frac{\partial}{\partial c} E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right] = E_P\left[\widetilde{A}_i f_V\left(c \cdot \widetilde{A}_i\right)\right]$, and that

$$\frac{\partial}{\partial c}\frac{1 - F_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]} = -\frac{\widetilde{A}_i f_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]} + \frac{1 - F_V\left(c \cdot \widetilde{A}_i\right)}{\left(1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]\right)^2} E_P\left[\widetilde{A}_i f_V\left(c \cdot \widetilde{A}_i\right)\right].$$

Since this quantity is bounded for small c, the dominated convergence theorem implies that

$$\frac{\partial}{\partial c} Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right) =$$

$$E_P\left[\left(\frac{\widetilde{A}_i f_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]} - \frac{1 - F_V\left(c \cdot \widetilde{A}_i\right)}{1 - E_P\left[F_V\left(c \cdot \widetilde{A}_i\right)\right]} E_P\left[\widetilde{A}_i f_V\left(c \cdot \widetilde{A}_i\right)\right]\right) f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right],$$

and thus that $Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)$ is continuously differentiable in c on a neighborhood of zero. Evaluating this derivative at $c = 0$ yields

$$\frac{\partial}{\partial c} Cov_{P_S}\left(W_i, f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right)\Big|_{c=0} = E_P\left[\left(\frac{\widetilde{A}_i f_V(0)}{1 - E_P\left[F_V(0)\right]} - \frac{E_P\left[\widetilde{A}_i f_V(0)\right]}{1 - E_P\left[F_V(0)\right]}\right) f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)\right].$$

Thus, we see that this derivative is nonzero so long as $\widetilde{A}_i$ is correlated with $f\left(C_i, \widetilde{TE}_i, \widetilde{A}_i\right)$

and $f_V(0) \neq 0$, which we have already assumed. Provided this derivative is nonzero, the result follows immediately. □

Appendix B.2 Interpretation of Φ and Φ_A Under Random Selection of Observables

To build intuition for the behavior of Φ and Φ_A, we consider a model in which the observable covariates represent a random subset of a larger collection of latent factors. As above, we focus on the case with selection on $\widetilde{A}_i$, while results under selection on $\widetilde{TE}_i$ follow as a special case.

Similar to Altonji et al. (2010), let us suppose that both the covariates C_i and the unobservables U_i are driven by a set of J unobserved factors F_i, with $J = \dim(C_i) + \dim(U_i)$. Let us also suppose that the factors F_i are conditional mean independent, in the sense that

$$E_{P_S}\left[F_{i,j} | F_{i,1}, ..., F_{i,j-1}, F_{i,j+1}, ..., F_{i,J}\right] = 0$$

for all j, so knowing the values of the other factors doesn't help us predict the value of the jth factor.

Suppose that $F_{C,i}$ and $F_{U,i}$ collect non-overlapping subsets of the factors, of size J_C and $J - J_C$ respectively, and that C_i and U_i are then generated as

$$C_i = \mu_C + \Lambda_C F_{C,i}$$

$$U_i = \mu_U + \Lambda_U F_{U,i},$$

where Λ_C and Λ_U have full rank. Note that $E_{P_S}[U_i|C_i] = \mu_U$ and $E_{P_S}[C_i|U_i] = \mu_C$, so the observables and unobservables are conditional mean independent.

Finally, let us suppose that the conditional expectations of both A_i and TE_i are linear in the factors,

$$E_{P_S}[A_i|F_i] = \mu_A + \gamma_F' F_i$$

$$E_{P_S}[TE_i|F_i] = \mu_{TE} + \delta_F' F_i.$$

This implies that the conditional expectations of these variables are linear in C_i, U_i as well:

$$\widetilde{A}_i = \tilde{\mu}_A + \gamma_C' C_i + \gamma_U' U_i,$$

$$\widehat{A}_i = \hat{\mu}_A + \gamma_C' C_i,$$

$$\widetilde{TE}_i = \tilde{\mu}_{TE} + \delta_C' C_i + \delta_U' U_i,$$

and

$$\widehat{TE}_i = \hat{\mu}_{TE} + \gamma_C' C_i.$$

For S_C and S_U are the selection matrices corresponding to $F_{C,i}$ and $F_{U,i}$,

$$\left(F_{C,i}, F_{U,i}\right) = \left(S_C F_i, S_U F_i\right),$$

the coefficients above are defined as

$$(\gamma_C, \gamma_U, \delta_C, \delta_U) = \left(\Lambda_C^{-1} S_C \gamma_F, \Lambda_U^{-1} S_U \gamma_F, \Lambda_C^{-1} S_C \delta_F, \Lambda_U^{-1} S_U \delta_F\right)$$

and

$$\left(\tilde{\mu}_A, \hat{\mu}_A, \tilde{\mu}_{TE}, \hat{\mu}_{TE}\right) =$$

$$\left(\mu_A - \gamma_C' \mu_C - \gamma_U' \mu_U, \mu_A - \gamma_C' \mu_C, \mu_{TE} - \delta_C' \mu_C - \delta_U' \mu_U, \mu_{TE} - \delta_C' \mu_C\right).$$

Under these assumptions, the fact that U_i and C_i are orthogonal implies that

$$\Phi_A = \frac{\gamma_C' \Sigma_C \delta_C + \gamma_U' \Sigma_U \delta_U}{\gamma_C' \Sigma_C \delta_C}$$

for Σ_C and Σ_U the variance matrices of C_i and U_i.

Random Selection of Factors: Thus far, we have treated the mapping from factors to variables as fixed. To obtain restrictions on Φ_A, let us instead model the selection of observable factors as random. In particular, suppose that non-overlapping sets of factors of size J_C and $J - J_C$ are drawn uniformly at random. Again denote vectors containing these factors by $F_{C,i}$ and $F_{U,i}$, respectively. Suppose that C_i and U_i are then generated as

$$C_i = \Lambda_C F_{C,i}$$

$$U_i = \Lambda_U F_{U,i},$$

where Λ_C and Λ_U again have full rank but may be random conditional on the set of factors selected.

Denoting expectations over the variable construction step by E^F, note that

$$E^F\left[\gamma_C' \Sigma_C \delta_C\right] = \frac{J_C}{J} \gamma_F' \Sigma_F \delta_F$$

while

$$E^F\left[\gamma_C' \Sigma_C \delta_C + \gamma_U' \Sigma_U \delta_U\right] = \gamma_F' \Sigma_F \delta_F.$$

Therefore, we see that

$$\frac{E^F\left[\gamma_C' \Sigma_C \delta_C + \gamma_U' \Sigma_U \delta_U\right]}{E^F\left[\gamma_C' \Sigma_C \delta_C\right]} = \frac{J}{J_C},$$

which is simply the inverse of the fraction of factors captured by the covariates. Unfortunately, however,

$$E^F\left[\Phi_A\right] \neq \frac{E^F\left[\gamma_C' \Sigma_C \delta_C + \gamma_U' \Sigma_U \delta_U\right]}{E^F\left[\gamma_C' \Sigma_C \delta_C\right]} = \frac{J}{J_C},$$

since the expectation of a ratio is not generally equal to the ratio of expectations.

This difficulty resolves if we take the number of factors to be large. In particular, let σ_j^2 denote the variance of factor j, and γ_j, δ_j the coefficients on this factor. Suppose that $\left(\sigma_j^2, \gamma_j, \delta_j\right)$ are drawn iid from some distribution such that $0 < E^F\left[\sigma_j^2 \gamma_j^2 + \sigma_j^2 \delta_j^2\right] < \infty$. If we take $J \to \infty$ and assume that $J_C/J \to \kappa_C$, then by the weak law of large numbers and the continuous mapping theorem

$$\Phi_A \to_p \frac{1}{\kappa_C},$$

so Φ_A has a natural interpretation in terms of the fraction of the factors captured by the covariates relative to the unobservables.

Appendix B.3 Inference Details

Here we discuss inference on the quantities we propose, including confidence sets for $\Phi\left(t_P^*\right)$ which remain valid when $E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]$ is small, and the justification for the confidence set proposed for the case when we have bounds $\Phi \in \left[\Phi_L, \Phi_U\right]$.

Appendix B.3.1 Confidence Set for $\Phi\left(t_P^*\right)$

To construct a confidence set for $\Phi\left(t_P^*\right)$, let

$$\begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}$$

denote the bootstrap estimate for the variance-covariance matrix of consistent and asymptotically normal estimates $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ for

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} t_P^* - E_{P_S}\left[T_i\right] \\ E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right] \end{pmatrix}.$$

We can use a version of the confidence set proposed by Anderson and Rubin (1949) and Fieller (1954). In particular, define the AR statistic evaluated at ϕ as

$$AR\left(\phi\right) = \frac{\left(\hat{\beta}_1 - \hat{\beta}_2\phi\right)^2}{\widehat{Var}\left(\hat{\beta}_1 - \hat{\beta}_2\phi\right)} = \frac{\left(\hat{\beta}_1 - \hat{\beta}_2\phi\right)^2}{\hat{\sigma}_1^2 - 2\phi\hat{\sigma}_{12} + \hat{\sigma}_2^2\phi^2}.$$

Note that $\beta_1 - \beta_2\Phi\left(t_P^*\right) = 0$. To construct a level α confidence set for $\Phi\left(t_P^*\right)$ we can simply collect the set of values where $AR\left(\phi\right)$ is less than a level $1-\alpha$ χ_1^2 critical value:

$$CS = \left\{\phi : AR\left(\phi\right) \leq \chi_{1,1-\alpha}^2\right\}.$$

One can show that this confidence set has correct coverage in large samples even when β_2 is close to (or exactly) zero. Moreover, when β_2 is large this confidence set behaves like the usual one, and so does not sacrifice efficiency in this case.

Appendix B.3.1 Confidence Set for $E_P\left[TE_i\right]$ Under Bounds on Φ

We next justify the proposed confidence set for $E_P\left[TE_i\right]$ under the assumption $\Phi \in \left[\Phi_L, \Phi_U\right]$. For $\left(\hat{\sigma}_L, \hat{\sigma}_U\right)$ bootstrap standard errors for our estimates $\left(\hat{\gamma}_L, \hat{\gamma}_U\right)$ of

$$\left(\gamma_L, \gamma_U\right) = \left(E_{P_S}\left[T_i\right] + \Phi_L\left(E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]\right), E_{P_S}\left[T_i\right] + \Phi_U\left(E_P\left[\widehat{T}_i\right] - E_{P_S}\left[T_i\right]\right)\right),$$

we proposed constructing a level $1-\alpha$ confidence interval for $E_P\left[TE_i\right]$ as

$$\left[\min\left\{\hat{\gamma}_L-\hat{\sigma}_L c_\alpha,\hat{\gamma}_U-\hat{\sigma}_U c_\alpha\right\},\max\left\{\hat{\gamma}_L+\hat{\sigma}_L c_\alpha,\hat{\gamma}_U+\hat{\sigma}_U c_\alpha\right\}\right],$$

To understand this procedure, note that $E_P\left[TE_i\right]$ is contained in the bounds implied by $\left[\Phi_L,\Phi_U\right]$ if and only if

$$\min\left\{\gamma_L,\gamma_U\right\}\leq E_P\left[TE_i\right]\leq\max\left\{\gamma_L,\gamma_U\right\},$$

or equivalently, either

$$H_0^a:\max\left\{\left(\gamma_L-E_P\left[TE_i\right]\right),-\left(\gamma_U-E_P\left[TE_i\right]\right)\right\}\leq 0$$

or

$$H_0^b:\max\left\{-\left(\gamma_L-E_P\left[TE_i\right]\right),\left(\gamma_U-E_P\left[TE_i\right]\right)\right\}\leq 0$$

holds.

However, this is the union of two hypotheses of the sort commonly tested in the literature on moment inequalities. Standard arguments in that literature show that the test that rejects

$$H_0^a:\max\left\{\left(\gamma_L-E_P\left[TE_i\right]\right),-\left(\gamma_U-E_P\left[TE_i\right]\right)\right\}\leq 0$$

only if

$$\max\left\{\frac{\hat{\gamma}_L-E_P\left[TE_i\right]}{\hat{\sigma}_L},-\frac{\hat{\gamma}_U-E_P\left[TE_i\right]}{\hat{\sigma}_U}\right\}>c_\alpha^*$$

for c_α^* the $1-\alpha$ quantile of $\max\left\{\xi_1,\xi_2\right\}$ for

$$\xi\sim N\left(0,\left(\begin{array}{cc}1 & \frac{\hat{\sigma}_{LU}}{\hat{\sigma}_L\hat{\sigma}_U}\\ \frac{\hat{\sigma}_{LU}}{\hat{\sigma}_L\hat{\sigma}_U} & 1\end{array}\right)\right)$$

has size at most α in large samples (where $\hat{\sigma}_{LU}$ is the bootstrap estimate of the covariance between Φ_L and Φ_U). Since we are interested in testing $H_0^a\cup H_0^b$, we thus consider the test which rejects only if our tests for H_0^a and H_0^b both reject. For a given hypothesized value $E_P\left[TE_i\right]$, this test rejects if and only if

$$\min\left\{\max\left\{\frac{\hat{\gamma}_L-E_P\left[TE_i\right]}{\hat{\sigma}_L},-\frac{\hat{\gamma}_U-E_P\left[TE_i\right]}{\hat{\sigma}_U}\right\},\max\left\{-\frac{\hat{\gamma}_L-E_P\left[TE_i\right]}{\hat{\sigma}_L},\frac{\hat{\gamma}_U-E_P\left[TE_i\right]}{\hat{\sigma}_U}\right\}\right\}>c_\alpha^*.$$

To form a confidence set, we can collect the set of non-rejected values, which is exactly

$$\left[\min\left\{\hat{\gamma}_L-\hat{\sigma}_L c_\alpha,\hat{\gamma}_U-\hat{\sigma}_U c_\alpha\right\},\max\left\{\hat{\gamma}_L+\hat{\sigma}_L c_\alpha,\hat{\gamma}_U+\hat{\sigma}_U c_\alpha\right\}\right].$$

Thus, this gives us a (conservative) level $1-\alpha$ confidence interval for $E_P\left[TE_i\right]$.

The confidence interval stated in the text is obtained by further noting that for all c,

$$Pr\left\{\max\left\{\xi_1,\xi_2\right\}>c\right\}\leq Pr\left\{\xi_1>c\right\}+Pr\left\{\xi_2>c\right\},$$

which implies that $c_\alpha^*\leq c_\alpha$ for c_α the two-sided level α normal critical value. Thus, we can form our confidence intervals with conventional critical values, though we will obtain better power by instead using the alternative (more computationally intensive) critical value c_α^*.

Appendix C: Tables and Figures

Table 1: **Observable Sample Selection**, **Attanasio et al (2011)**

Variable	Population: Mean (SD)	Sample: Mean (SD)
Age	21.6 (2.26)	22.8 (2.04)
Education	8.5 (2.98)	10.2 (1.6)
Prior Employment	0.205 (0.404)	0.468 (0.449)
Prior Contract	0.034 (0.018)	0.068 (0.252)
Prior Formal Employment	0.026 (0.16)	0.066 (0.249)

Notes: This table illustrates the moments in the sample and population for the Attanasio et al (2011) paper.

Table 2: **Observable Sample Selection, 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 illustrates the moments in the sample and population for the Bloom et al (2015) paper.

Table 3: **Application: Bloom et al (2015), Alternative Covariate Approach**

Outcome	Baseline Effect	Observable Adjusted	Bounds, $\Phi \in [1,2]$	$\Phi(0)$
Job Performance	0.271	0.289	[0.289, 0.309]	-14.7
	(0.22, 0.32)	(0.23,0.34)	(0.241, 0.370)	(-∞,∞)

Notes: This table shows the application of our sensitivity procedure to Bloom et al (2015). The moments comes from the study. Standard errors are bootstrapped. This table shows an alternative approach to adjusting for covariates, by regressing the outcome on covariates separately for treatment and control and generating the difference in predicted values to estimate the average treatment effect.

Table 4: **Observable Sample Selection, Dupas and Robinson (2013)**

Variable	Population: Mean	Sample: Mean
Age	40.95	39.03
Female	0.681	0.737
Hyperbolic	0.152	0.159
Time Inconsistent	0.175	0.177
High Discount Rate	0.467	0.442
Education	5.67	6.31
Female X Married	0.495	0.555
Female X Hyperbolic	0.110	0.116
Female X Time Inconsistent	0.108	0.127
Female X High Discount	0.318	0.334

Notes: This table illustrates the moments in the sample and population for the Dupas and Robinson (2013) paper. The difference between ROSCAs and Non-ROSCAS is drawn from external data, helpfully provided by the authors. Note that since we are inferring the population mean from data on the difference we cannot match the trial and target populations on standard deviations.

Table 5: **Observable Sample Selection, Olken et al (2014)**

Variable	Population: Mean (SD)	Sample: Mean (SD)
Dirt Floor Share	0.174	0.226 (0.244)
Cash Transfer Share	0.347	0.360 (0.227)
Avg. # Vaccinations	7.40	8.14 (2.58)
Avg. Length Breastfeed	15.6	15.7 (4.34)
Literate Share	0.908	0.917 (0.070)
Contraceptive Share	0.215	0.233 (0.099)

Notes: This table illustrates the moments in the sample and population for the Olken et al (2014) paper. The restricted moment come from the SUSENAS data on Indonesia, which is merged with the Olken et al (2014) data at the subdistrict level.