# Weighting for External Validity*

Isaiah Andrews
MIT and NBER

Emily Oster
Brown University and NBER

March 8, 2017

**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 covariates, 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 an empirically tractable framework for evaluating external validity in such cases. Our approach relies on the fact that under mild conditions 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. Using intuitive bounds for selection on the portion of the treatment effect explained by covariates relative to the residual, we derive bounds on extrapolation bias. We then suggest natural benchmarks for assessing external validity, discuss implementation, and apply our results to data from several recent experiments.

*Sophie Sun provided excellent research assistance. We thank participants in seminars at at Brown University, Harvard University and University of Connecticut for helpful comments.

# 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 et al (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 closing their offices and having all their employees 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 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 of 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 obtain population-appropriate estimates (as in e.g. Hellerstein and Imbens, 1999). In this paper, we consider cases where the sample may also be selected on unobservable dimensions;

[1] In medicine, for example, the efficacy of drugs is tested on study particpants who may differ systematically from the general population of possible users. See Stuart et al (2011) for discussion.

this is a common concern in practice, especially in cases where there may be private information on treatment effects.[2] We provide an empirically tractable framework for evaluating 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 observed. Nonetheless, assumptions about these weights provide intuitive restrictions for the form of selection, and we explore robustness of estimates to such assumptions.

We begin in Section 2 by describing a simulated example, based on a real dataset, which we use to illustrate the biases arising from selection of the experimental population. We return to this example throughout the paper to build intuition and illustrate our approach.

Section 3 introduces 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 will show that a suitable choice of target function yields the average treatment effect 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, then the necessary weights are unknown.

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 average treatment effect estimation, this highlights that the 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.

Without further restrictions, the bias from extrapolating trial population moments to the

[2]One recent example of this is Alcott (2015), who uses the example of the energy company OPower to demonstrate that early study sites show much higher treatment effects than later ones. He finds that this difference is partially but not completely explained by observable differences between early and late sites, suggesting that unobservable differences must play an important role.

target population may be arbitrarily large. To make progress, we first assume that some features of the target population, for example demographic characteristics, are known. We then seek further intuitive restrictions which allow us to derive bounds.

We consider two possible further restrictions. First, we observe that known features of the target population allow us to derive a lower bound on the standard deviation of the weights. If one is willing to assume an upper bound on the weights - perhaps informed by this lower bound - we show that this suffices to bound extrapolation bias. Second, we adopt bounds on the ratio of the correlations between (1) the weights and the target function and (2) the weights and some benchmark function of the covariates whose mean in the target population is known. This ratio can be interpreted as measuring the intensity of selection on the target function relative to selection on the benchmark. We find the latter approach - bounding the correlation ratio - more interpretable in applications, and so focus on these results for the remainder of this paper. In Appendix B, however, we work out the details of the approach which bounds the standard deviation of the weights directly.

Section 4 further develops our bounding approach. The key question in deriving bounds based on correlation ratios is what benchmark function to use, and we suggest considering the predicted value for the target function given the covariates. For this choice, we define a transformation of the correlation ratio which measures the intensity of selection on residuals relative to predicted values. If we assume no selection on residuals, our approach recovers existing corrections for selection on observables (e.g. Hellerstein and Imbens, 1999; Hotz et al, 2005). Our approach considerably extends existing results, however, since we can both derive bounds under nonzero selection on residuals and ask how much selection on residuals would be required to overturn a given result.

Our bounds on extrapolation bias depend on two features of the data. The first is the adjustment for selection on observables. All else equal, a larger correction for selection on observables leads to wider bounds. The second is the explanatory power of the observables. All else equal, higher explanatory power for the observables leads to tighter bounds.

In Section 5, we specialize our results to the case of extrapolating average treatment effects. Section 6 discusses implementation, and we argue that a natural baseline is to consider results under the assumption that selection on the residual is in the same direction as, and no more intense than, the selection on the fitted values. We also suggest calculating the degree of

selection necessary to produce the smallest economically interesting treatment effect, which we take to be zero in our applications. Larger values indicate a more robust result, in the sense that more intense selection is required to overturn the result in the target population. We provide some intuition for these bounds, and discuss when they are likely to perform well, by which we mean both containing the true target population effect and being small. Finally, we briefly discuss inference, and how to account for sampling uncertainty.

In Section 7 we turn to applications, making use of our suggested approach to check robustness of estimates to assumptions about weights. We apply our results to data drawn from Bloom, Liang, Roberts, and Ying (2015), Dupas and Robinson (2013), and Olken, Onishi, and Wong (2014). As noted above, Bloom et al (2015) study the effect of working from home on productivity in a Chinese firm. In their setting, we find that correcting for selection on observables makes estimated treatment effects larger, and that selection on unobservables would have to operate in the opposite direction from selection on observable to overturn their results. Dupas and Robinson (2013) study the effect of savings technologies on investments in preventative healthcare and vulnerability to health shocks. They consider a variety of treatments and outcomes, and we find that some of the resulting relationships are more robust to concerns about external validity than others. Finally, Olken et al (2014) study the effect of performance incentives for Indonesian villages on maternal health and child education. They do not find significant results in their baseline analysis, but correcting for selection on observables increases the estimated treatment effects, and our baseline bounds on unobservable selection suggest the possibility of still-larger treatment effects in the overall population.

This paper contributes to the literature on external validity of treatment effects. Although we focus on cases where the trial and target population may differ on unobservables, we relate closely to the literature studying selection on observables (Hellerstein and Imbens, 1999; Hotz et al, 2005; Cole and Stuart, 2010; Stuart et al, 2011; Imai and Ratkovic, 2014; Dehijia et al, 2015; Hartman et al, 2015), as well as to the large literature on propensity score reweighting (e.g. Hahn 1998 and Hirano et al 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 a similar problem to ours and suggests bounds which result from assumptions on the level of dependence between the individual outcomes in the treated and

untreated states.

We also relate, if more distantly, to the recent literature on external validity in instrumental variables (Feller et al, 2016; Kowalski, 2016; Kline and Walters, *forthcoming;* Brinch et al*, forthcoming*) and regression discontinuity (Bertanha and Imbens, 2014; Angrist and Rokkanen, 2015; Rokkanen, 2015). These connections are discussed in Section 8 below.

## 2 Illustrative Example and Data Description

To develop intuition for the effect of sample selection on external validity, we begin by describing a constructed example. This is based on real data, and we will use it as a running example throughout the paper. Using a specific example helps fix ideas about sample selection, while using a constructed dataset ensures that we know the true form of selection. This allows us to illustrate our theoretical results by calculating quantities that are normally unknown.

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 Andra 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 of the Muralidharan and Sundararaman (2011) data. We then define the distribution of the trial population by taking a weighted sample from the target population. This allows us to observe the true sampling weights, as well as the full distribution of the data in both trial and target populations. The sampling weights depend both on the treatment effect as predicted by covariates *and* on the treatment effect as predicted by geographic regions, which we treat as unobserved in the rest of our analysis. Thus, in this example we have selection on both observables and unobservables.

Specifically, we define $\widehat{OTE_i}$ to be the expectation of the treatment effect conditional on a vector of covariates. Correspondingly, $\widehat{UTE_i}$ is the expectation of the treatment effect conditional on dummies for the mandal (geographic location) in which the school is located. We treat these mandal dummies, and thus $\widehat{UTE_i}$, as unobserved. We then define sampling

weights at the school level, setting

$$wt_1 = \begin{cases} a & \text{if } \widehat{OTE_i} < q_{66}\left(\widehat{OTE_i}\right) \\ 1 & \text{if } \widehat{OTE_i} \geq q_{66}\left(\widehat{OTE_i}\right) \end{cases}$$

$$wt_2 = \begin{cases} b & \text{if } \widehat{UTE_i} < q_{66}\left(\widehat{UTE_i}\right) \\ 1 & \text{if } \widehat{UTE_i} \geq q_{66}\left(\widehat{UTE_i}\right) \end{cases}$$

$$weight = \frac{1}{2}\left(wt_1 + wt_2\right)$$

where $q_{66}\left(X_i\right)$ indicates the 66th percentile of $X_i$ and $a, b < 1$. For our example we use $a = 0.1$ and $b = 0.4$. These weights favor schools with higher treatment effects, as predicted by both (observed) covariates and (unobserved) location dummies. We construct the trial population by sampling schools with replacement, and draw a sample much larger than the original population to eliminate sampling noise.

While our primary focus in this paper is on external validity of average treatment effects, our results in fact apply to the mean of any function of the data. For ease of exposition, our initial examples in the next section focus on cases where the target moments are the means of particular variables (i.e. teacher and school characteristics, baseline test scores) in the target population. As with average treatment effects, the means of these variables all differ between the trial and target populations due to sample selection.

## 3 Sample Selection and Reweighting

To develop our reweighting framework, assume 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 6 below.

**Illustration:** To illustrate, we return to the example based on Muralidharan and Sundararaman (2011) introduced in the last section. In this example, the trial population consists of students at schools reached by the sampling scheme, and $P_S$ denotes the distribution in this population. The target population, by contrast, is the population of all students, and the distribution in this population is given by $P$. The choice of target function $t(X_i)$ will reflect our quantity of interest. If we are interested in the average baseline test score across all schools, for example, we can take $t(X_i) = Score_i$. $\triangle$

To make progress, 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 $A$, $Pr_{P_S}\{X_i \in A\} = 0$ implies $Pr_P\{X_i \in A\} = 0$, and thus that all events which have zero probability in the trial population likewise have zero probability in the target population. This is a strong assumption, but it is closely analogous to probabilistic assignment as assumed in the literature on treatment effect estimation (see, for example, Definition 3.5 in Imbens and Rubin 2015). In particular, probabilistic assignment requires that for any value of the covariates and potential outcomes we may observe individuals in both the treated and untreated states. This in turn implies that the distributions of potential outcomes and covariates in the treated and untreated states are mutually absolutely continuous.

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 $S_i$.[3] We record this result in the following lemma.

**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 in this case. Thus, external validity issues in this setting arise directly from $X$-dependent selection into the trial population.

### 3.1 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 version of a classic result by Horvitz and Thompson (1952), and shows that we can recover expectations under $P$ by reweighting our observations from $P_S$ using the

[3] We define all densities with respect to a base fixed measure $\mu$. $\mu$ need not be Lebesgue measure, so our results allow the possibility that $X_i$ is not continuously distributed.

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 $X_i$. Thus, 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 both finite, then*

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

*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 (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 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 $f(\cdot)$,

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

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.

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 reflects the fact that $\sigma_{P_S}\left(W_i\right)$ measures the degree of sample selection along any dimension.

**Illustration (continued):** In our example based on Muralidharan and Sundararaman (2011) unlike in real data, we know the distributions in both trial and target populations. 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 teacher absence, teacher salary, teacher training, household income, and school infrastructure.

The first two columns of Table 1 report the trial and target population means for these variables in our constructed example. Note that the absence measure records the share of teachers *present* and that all the remaining variables other than salary are categorical. The final three columns show the elements of the bias decomposition in Corollary 1. The difference in means for each variable is the product of these three elements. The bias is smallest (in percentage terms) for household income, arising from the low correlation between this variable and the weights, which reflects that the sample is not much selected on this variable. By contrast, there is a large correlation between school infrastructure and the weights, indicating strong selection on this variable, which leads to correspondingly large bias. As noted above the standard deviation of weights is the same in all rows, since this is a measure of selection on *any* dimension. $\triangle$

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) the sample selection is more heavily weighted

toward the function in question, and (c) there is more variability in the function. We next consider additional assumptions which let us directly bound the target moment $E_P\left[t\left(X_i\right)\right]$.

### 3.2 Assessing Extrapolation Bias

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 $E_{P_S}\left[t\left(X_i\right)\right]$ and would like to know $E_P\left[t\left(X_i\right)\right]$. Corollary 1 shows that this is equivalent to knowing the covariance between the weights $W_i$ and $t\left(X_i\right)$. From equation (3) we see that this covariance in turn depends on two unknown objects, the standard deviation $\sigma_{P_S}\left(W_i\right)$ and the correlation $\rho\left(W_i, t\left(X_i\right)\right)$, along with the known standard deviation $\sigma_{P_S}\left(t\left(X_i\right)\right)$.

Since the weights $W_i$ are unknown, this decomposition does not by itself allow us to restrict the target moment $E_P\left[t\left(X_i\right)\right]$, and we must impose additional restrictions to make progress. Indeed, since the distribution of $X_i$ in the target population is unknown, any approach to external validity must impose some assumptions on the relationship between the trial and target populations. Given this necessity, our goal is to find assumptions which are interpretable, so researchers and audience members can assess for themselves whether an assumption is plausible in a given setting, and which also yield useful bounds on the target moment.

We first consider upper bounds on the standard deviation $\sigma_{P_S}\left(W_i\right)$ of the weights, which limit the overall degree of selection. We then turn to bounds on the correlation ratio $\frac{\rho_{P_S}(W_i,t(X_i))}{\rho_{P_S}(W_i,b(X_i))}$, for a benchmark function $b\left(X_i\right)$ with known mean in the target population. This second approach limits the degree of selection on the target function $t\left(X_i\right)$ relative to selection on the benchmark function $b\left(X_i\right)$.

**Absolute Selection Bounds** If we have an upper bound on the standard deviation of the weights, say $\sigma_{P_S}\left(W_i\right) \leq c$, we can seek the smallest and largest values for our target moment $E_P\left[t\left(X_i\right)\right]$ consistent with this restriction, which is straightforward using equation (4). The variance $\sigma^2_{P_S}\left(W_i\right)$ has a long history as a measure for the difference between two distributions (here $P_{X,S}$ and $P_X$): this is known as the $\chi^2$ or Pearson $\chi^2$ divergence and dates back at least to Pearson (1900). Thus, bounding $\sigma_{P_S}\left(W_i\right)$ above can be understood as bounding the overall degree of selection.

To sharpen the bounds from this approach, we can exploit known features of the target population. In particular, suppose we know the mean of some function of the data $r(X_i)$ in the target population

$$E_P[r(X_i)] = r_P, \tag{6}$$

where $r(X_i)$ may be a vector, $r(X_i) = (r_1(X_i), ..., r_k(X_i))'$. For example, we often know some descriptive statistics for the target population, and so we can use equation (6) to impose e.g. known means and variances for particular variables. We will call $r(X_i)$ the restriction functions, and $r_P$ the restricted moments.

The weights $W_i$ must rebalance not only the mean of the target function $t(X_i)$, but also the vector of restriction functions $r(X_i)$. Recall from equation (5) above that the same weights matter for each moment of the data. Knowledge of $r_P$ thus implies restrictions on the weights $W_i$, and in particular a lower bound on $\sigma_{P_S}(W_i)$, since the weights must be sufficiently large to match the restricted moments. When combined with an upper bound on $\sigma_{P_S}(W_i)$, including restricted moments yields tighter bounds on the target moment $E_P[t(X_i)]$, since it constrains the possible behavior for $W_i$. Solving for the resulting bounds on $E_P[t(X_i)]$ requires optimizing over the set of all possible weights and so may sound daunting, but turns out to be highly tractable. We derive the resulting bounds in Appendix B.

To select an upper bound, say $c$, on $\sigma_{P_S}(W_i)$, one natural approach is to take $c$ to equal the lower bound implied by the restricted moments. This corresponds to assuming a form of selection on observables, and in this case our bounds on $E_P[t(X_i)]$ collapse to a selection-on-observables-corrected point estimate. To allow selection on unobservables, one could instead take the upper bound $c$ to equal a multiple of the lower bound implied by the restricted moments. In applications, however, we found it difficult to argue for a particular choice of $c$, or to suggest natural benchmarks. To overcome these difficulties we next discuss an alternative approach which avoids specifying bounds on $\sigma_{P_S}(W_i)$ by instead comparing the degree of selection on different functions.

**Relative Selection Bounds** To avoid bounding the absolute level of selection $\sigma_{P_S}(W_i)$, we can instead limit the degree of selection on $t(X_i)$ relative to a benchmark function $b(X_i)$. In particular, let us suppose that $t(X_i)$ and $b(X_i)$ correspond to two rows in display (5),

$f_1(X_i) = t(X_i)$ and $f_2(X_i) = b(X_i)$ . Taking the ratio of these rows shows that

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

where the standard deviation $\sigma_{P_S}\left(W_i\right)$ of the weights drops out since it is the same across rows. Rearranging, we see that

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

where the only unknown term is the correlation ratio on the right hand side. Let us denote this ratio by

$$\Psi = \frac{\rho_{P_S}\left(W_i, t\left(X_i\right)\right)}{\rho_{P_S}\left(W_i, b\left(X_i\right)\right)}. \tag{7}$$

We thus see that if we can bound $\Psi$, this immediately implies bounds on $E_P\left[t\left(X_i\right)\right]$ without any restrictions on the magnitude of the difference between the trial and target populations as measured by the standard deviation $\sigma_{P_S}\left(W_i\right)$ of the weights.

The ratio $\Psi$ measures the extent to which selection loads on the target function $t\left(X_i\right)$ relative to the benchmark function $b\left(X_i\right)$. In particular, if we expect $b\left(X_i\right)$ to be at least as closely related to selection into the trial population as $t\left(X_i\right)$ then we can restrict $\Psi$ to lie in the interval $[-1, 1]$. More generally, if we restrict $\Psi$ to lie in the interval $[\Psi_L, \Psi_U]$ then we obtain the following bounds on the target moment:

**Proposition 1** *Suppose that*

$$\Psi = \frac{\rho_{P_S}\left(W_i, t\left(X_i\right)\right)}{\rho_{P_S}\left(W_i, b\left(X_i\right)\right)} \in \left[\Psi_L, \Psi_U\right].$$

*Suppose, further, that* $E_P\left[b\left(X_i\right)\right] - E_{P_S}\left[b\left(X_i\right)\right] > 0$. *Then* $E_P\left[t\left(X_i\right)\right] \in \left[t_P^L, t_P^U\right]$ *for*

$$t_P^L = E_{P_S}\left[t\left(X_i\right)\right] + \Psi_L \frac{\sigma_{P_S}\left(t\left(X_i\right)\right)}{\sigma_{P_S}\left(b\left(X_i\right)\right)} \left(E_P\left[b\left(X_i\right)\right] - E_{P_S}\left[b\left(X_i\right)\right]\right)$$

$$t_P^U = E_{P_S}\left[t\left(X_i\right)\right] + \Psi_U \frac{\sigma_{P_S}\left(t\left(X_i\right)\right)}{\sigma_{P_S}\left(b\left(X_i\right)\right)} \left(E_P\left[b\left(X_i\right)\right] - E_{P_S}\left[b\left(X_i\right)\right]\right).$$

Thus, given bounds $[\Psi_L, \Psi_U]$ on the correlation ratio it is straightforward to derive bounds on the target moment. Conversely, if we are interested in whether a given value $t_P^*$ for the

target moment is reasonable, we can calculate the implied correlation ratio

$$\Psi\left(t_P^*\right) = \frac{\left(t_P^* - E_{P_S}\left[t\left(X_i\right)\right]\right)/\sigma_{P_S}\left(t\left(X_i\right)\right)}{\left(E_P\left[b\left(X_i\right)\right] - E_{P_S}\left[b\left(X_i\right)\right]\right)/\sigma_{P_S}\left(b\left(X_i\right)\right)}.$$

If $\Psi\left(t_P^*\right)$ implies implausible selection patterns, this suggests that $E_P\left[t\left(X_i\right)\right] \neq t_P^*$.

The key input to this approach is a benchmark function $b\left(X_i\right)$ such that (a) $E_P\left[b\left(X_i\right)\right]$ is known and (b) we can bound the correlation ratio $\Psi$. The next section suggests a widely applicable approach for constructing such $b\left(X_i\right)$ based on predicted values for $t\left(X_i\right)$, where the prediction is based on variables whose behavior in the target population is known. Before introducing this approach, however, we explore the behavior of the correlation ratio $\Psi$ in our running example.

**Illustration (continued):** Consider introducing two new variables - average teacher education and baseline test score - which we observe in the trial population but treat as unknown in the target population. In this constructed example, of course, we do observe these variables in the target population. Thus we can compare the target popluation mean with the estimated bounds. The top of Panels A and B of Table 2 show summary information for these variables, including their trial and target population means and their correlations with the weights.

We show how we could use different choices of the benchmark moments - the same variables used in Table 1 - to calculate the mean of the unknown moments in the target population. For each benchmark moment, Column (1) in Table 2 shows the ratio $\Psi$ between the correlation of the target moment with the weights and the benchmark moment with the weights. Column (2) shows the estimate of the target population mean using this $\Psi$, which recovers the target population mean by construction. Column (3) then shows the bounds on the target population mean which result from assuming that $\Psi \in [-1, 1]$. The target population mean is contained in these bounds if and only if the true value of $\Psi$ indeed lies between -1 and 1.

As this table makes clear, the true value of $\Psi$ depends on the benchmark moment considered. For example, the correlation of average household income is small relative to that of teacher education. As a result, a large value of $\Psi$ is needed to match the target population mean in this case. On the other hand, the correlation between school infrastructure and the weights is large, implying a much smaller value $\Psi$. $\triangle$

In some settings there may be a particular benchmark function whose relationship to the target function is known. In such cases, this benchmark function can be used directly, as in the calculations just discussed. In many contexts, however, a suitable choice of benchmark is less clear. The key follow-on question is thus how to choose the benchmark function, which is the topic of the next section.

# 4 Residual Selection

In the last section, we showed that bounding selection on the target function $t(X_i)$ relative to selection on a benchmark function $b(X_i)$ yields bounds on $E_P[t(X_i)]$. A key step in implementing this approach is choosing the benchmark function $b(X_i)$. In this section we suggest a widely-applicable choice of $b(X_i)$ that builds on ideas from the literature studying selection on observables.

**Selection on Observables** As noted above, in many applications we have some set of covariates, for example demographic characteristics, whose distribution is observed in both the trial and target populations. Let us denote these covariates by $C_i$. A common approach to external validity, considered by a range of papers including Hellerstein and Imbens (1999) and Hotz et al (2005), assumes that selection operates entirely through $C_i$, so that once we control for $C_i$ there is no further selection. When the trial population is a subset of the target population, for example, this assumption specifies that for $S_i$ again a dummy for membership in the trial population, $S_i \perp X_i | C_i$, so selection is independent of $X_i$ conditional on $C_i$. Under this assumption, one can show that

$$E_P[t(X_i)|C_i] = E_{P_S}[t(X_i)|C_i]. \tag{8}$$

Thus, since we know the distribution of $C_i$ in the target population we can recover $E_P[t(X_i)]$ by simply reweighting the distribution of covariates to match that in the target population. In particular, for $W_i^C = \frac{p_C(C_i)}{p_{C,S}(C_i)}$, the weights that match the distribution of covariates in the trial and target populations, Lemma 2 implies that

$$E_{P_S}\left[W_i^C t(X_i)\right] = E_{P_S}\left[W_i^C E_{P_S}[t(X_i)|C_i]\right] = E_P[E_{P_S}[t(X_i)|C_i]] = E_P[t(X_i)]. \tag{9}$$

Thus, we can estimate $E_P[t(X_i)]$ by the sample average of $W_i^C t(X_i)$ for known weights $W_i^C$. This approach is known as propensity score reweighting.

In this paper we are interested in selection on unobservables as well as on observables. Once we allow for selection on unobservables equation (8) typically no longer holds, however, and reweighting with $W_i^C$ no longer yields valid estimates. Nonetheless, the conditional expectation of $t(X_i)$ given the covariates $C_i$ provides a useful starting point for our analysis.

**Conditional Expectation Decomposition** To accommodate selection on unobservables, we take $b(X_i)$ to be the conditional expectation of the target function given $C_i$,

$$b(X_i) = E_{P_S}[t(X_i)|C_i].$$

Since this conditional expectation is a function of $C_i$, and we have assumed the distribution of $C_i$ in the target population is known, it follows that the distribution of $b(X_i)$ in the target population is known as well. To ensure that selection on $b(X_i)$ is positive, as assumed in Proposition 1, we can specify $b(X_i) = -E_{P_S}[t(X_i)|C_i]$ if needed. Thus, this choice of $b(X_i)$ satisfies the requirements of Proposition 1.

Since we allow selection on unobservables, we also need to account for the part of $t(X_i)$ not captured by the covariates. To do so, let us define $\varepsilon_i$ as the residual from $t(X_i)$ after taking out $b(X_i)$

$$\varepsilon_i = t(X_i) - b(X_i) = t(X_i) - E_{P_S}[t(X_i)|C_i].$$

Note that $E_{P_S}[\varepsilon_i] = 0$ by construction. Since $E_P[t(X_i)] = E_P[b(X_i)] + E_P[\varepsilon_i]$, and our assumptions imply that $E_P[b(X_i)]$ is known, the problem of inference on $E_P[t(X_i)]$ reduces to one of inference on $E_P[\varepsilon_i]$.

**Bounds** Selection on observables implies that the weights $W_i$ are orthogonal to the residual $\varepsilon_i$, $\rho_{P_S}(W_i, \varepsilon_i) = 0$, which is why $E_P[t(X_i)] = E_P[b(X_i)]$. Since we aim to relax this assumption, it is natural to allow selection to load on the residual to a limited extent, where we again measure the degree of selection on the residual relative to selection on $b(X_i)$.

Formally, we assume that

$$\Psi^{\varepsilon} = \frac{\rho_{P_S}(W_i, \varepsilon_i)}{\rho_{P_S}(W_i, b(X_i))} \in [\Psi_L^{\varepsilon}, \Psi_U^{\varepsilon}]. \tag{10}$$

For example, to require that selection on $\varepsilon_i$ and $b(X_i)$ operate in the same direction, we can set $\Psi_L^{\varepsilon} = 0$. To require that selection on $\varepsilon_i$ be no more intense than selection on $b(X_i)$ we can set $\Psi_U^{\varepsilon} = 1$. Conversely, to require that selection on $\varepsilon_i$ be more intense than selection on $b(X_i)$ we can set $\Psi_L^{\varepsilon} = 1$. Finally, to recover the selection-on-observables case we can set $\Psi_L^{\varepsilon} = \Psi_U^{\varepsilon} = 0$.

Bounds on $\Psi^{\varepsilon}$ imply bounds on $\Psi$ as defined in equation (7).

**Lemma 3** *For $\varepsilon_i$ and $b(X_i)$ as defined above,*

$$\Psi^{\varepsilon} = \frac{\rho_{P_S}(W_i, \varepsilon_i)}{\rho_{P_S}(W_i, b(X_i))} \in [\Psi_L^{\varepsilon}, \Psi_U^{\varepsilon}] \text{ if and only if } \Psi = \frac{\rho_{P_S}(W_i, t(X_i))}{\rho_{P_S}(W_i, b(X_i))} \in [\Psi_L, \Psi_U]$$

*for*

$$\begin{aligned}[\Psi_L, \Psi_U] &= \left[\frac{\sigma_{P_S}(b(X_i))}{\sigma_{P_S}(t(X_i))} + \Psi_L^{\varepsilon}\frac{\sigma_{P_S}(\varepsilon_i)}{\sigma_{P_S}(t(X_i))}, \frac{\sigma_{P_S}(b(X_i))}{\sigma_{P_S}(t(X_i))} + \Psi_U^{\varepsilon}\frac{\sigma_{P_S}(\varepsilon_i)}{\sigma_{P_S}(t(X_i))}\right] \\ &= \left[\sqrt{R^2} + \Psi_L^{\varepsilon}\sqrt{1-R^2}, \sqrt{R^2} + \Psi_U^{\varepsilon}\sqrt{1-R^2}\right]\end{aligned}$$

*where*

$$R^2 = \frac{\sigma_{P_S}^2(b(X_i))}{\sigma_{P_S}^2(t(X_i))} = \frac{\sigma_{P_S}^2(E_{P_S}[t(X_i)|C_i])}{\sigma_{P_S}^2(t(X_i))}.$$

The term $R^2$ measures the fraction of the variance of $t(X_i)$ explained by the covariates $C_i$, and is closely related to the $R^2$ in linear regression. Likewise,

$$1 - R^2 = \frac{\sigma_{P_S}^2(\varepsilon_i)}{\sigma_{P_S}^2(t(X_i))} = \frac{E_{P_S}\left[\sigma_{P_S}^2(t(X_i)|C_i)\right]}{\sigma_{P_S}^2(t(X_i))}$$

is the fraction of the variance of $t(X_i)$ unexplained by the covariates. The exact role of $R^2$ becomes clearer when we use Proposition 1 to calculate the implied bounds $t_P^L$ and $t_P^U$ on $E_P[t(X_i)]$.

**Corollary 2** *If*

$$\Psi^{\varepsilon} = \frac{\rho_{P_S}(W_i, \varepsilon_i)}{\rho_{P_S}(W_i, b(X_i))} \in [\Psi_L^{\varepsilon}, \Psi_U^{\varepsilon}]$$

*and* $E_P[b(X_i)] - E_{P_S}[b(X_i)] > 0$*, then* $E_P[t(X_i)] \in \left[t_P^L, t_P^U\right]$ *for*

$$t_P^L = E_P[b(X_i)] + \Psi_L^{\varepsilon}\sqrt{\frac{1-R^2}{R^2}}\left(E_P[b(X_i)] - E_{P_S}[b(X_i)]\right)$$

$$t_P^U = E_P[b(X_i)] + \Psi_U^{\varepsilon}\sqrt{\frac{1-R^2}{R^2}}\left(E_P[b(X_i)] - E_{P_S}[b(X_i)]\right).$$

In this result, the $R^2$ determines how our bounds for $E_P[t(X_i)]$ depend on our bounds for $\Psi^{\varepsilon}$. When the $R^2$ is close to one our bounds on $E_P[t(X_i)]$ are tight around the selection-on-observables-corrected value $E_P[b(X_i)] = E_P[E_{P_S}[t(X_i)|C_i]]$. By contrast, when the $R^2$ is small our bounds add and subtract large multiples of the selection-on-observables bias $E_P[b(X_i)] - E_{P_S}[b(X_i)]$, with the exact multiple determined by our bounds on the correlation ratio. This behavior is intuitive: if almost all of the variation of $t(X_i)$ is explained by the covariates, there is little scope for selection to introduce additional bias. By contrast, if little of the variability in $t(X_i)$ is explained by our covariates then even a small amount of selection on the residual could potentially result in large biases. Moreover, in all cases we measure selection on the residual relative to selection on $b(X_i)$, so more selection on $b(X_i)$ leads us to expect more selection on the residual as well.

**Linear Model** While our results above are derived in terms of the conditional expectation function $E_{P_S}[t(X_i)|C_i]$, we implement this approach using the linear regression model. To do so, we again take $r(X_i)$ to be a vector of restriction functions whose mean $r_P$ in the target population is known. We require that $r(X_i)$ includes a constant, and define

$$\gamma = E_{P_S}\left[r(X_i)\, r(X_i)'\right]^{-1} E_{P_S}\left[r(X_i)\, t(X_i)\right]$$

to be the coefficient from regressing $t(X_i)$ on $r(X_i)$. We can then define $b(X_i) = r(X_i)'\gamma$ to be the fitted value from this regression. Note that if we take $r(X_i) = h(C_i)$ to be a vector of transformations of the covariates $C_i$, then we can interpret $b(X_i)$ as an approximation to $E_{P_S}[t(X_i)|C_i]$. Moreover, standard approximation results show that by taking the functions $h(C_i)$ to be sufficiently rich we can approximate $E_{P_S}[t(X_i)|C_i]$ arbitrarily well.

A strength of this linear approach is that we can apply it even when we know relatively little about the target population. In particular, knowledge of $r_P = E_P[r(X_i)]$ allows us to

derive $E_P[b(X_i)]$, since

$$E_P[b(X_i)] = E_P\left[r(X_i)'\gamma\right] = E_P[r(X_i)]'\gamma.$$

If we consider correlation ratios based on this $b(X_i)$ and

$$\varepsilon_i = t(X_i) - r(X_i)'\gamma,$$

then Corollary 2 continues to apply, and yields bounds on the target moment. This ensures that our results continue to apply even in settings where $r(X_i)$ is insufficiently rich for us to be confident that $b(X_i) = r(X_i)'\gamma$ approximates $E_{P_S}[t(X_i)|C_i]$ well, for example when we know only a limited set of descriptive statistics $r_P$ for the target population.

**Illustration (continued):** We return to our running example and take teacher education and baseline test scores as the unknown moments. Above, we constructed bounds by using several different covariates as benchmark functions $b(X_i)$, while here we use the fitted values from a regression of each target function on the vector of restriction functions $r(X_i)$. Constructing our bounds in this case requires an assumption about the correlation between the fitted values from this regression and the weights, relative to the correlation between the residuals and the weights. We calculate the value of $\Psi^\varepsilon$ consistent with the true target population means for both teacher education and the baseline test score. We also calculate bounds under the assumption that $\Psi^\varepsilon \in [0, 1]$.

Panel A of Table 3 shows the results for teacher education, while Panel B shows results for the baseline test score. The first row of each column reports results including only the levels of salary and school infrastructure in the restriction function $r(X_i)$. For teacher education this adjustment moves the estimates towards the true value, although the change is small, and the $\Psi^\varepsilon$ which matches the target population mean is large. For the baseline test score the adjustment moves the estimates too far, and the $\Psi^\varepsilon$ which would match the target population means is negative. Column (5) shows the bounds which result from assuming that $\Psi^\varepsilon \in [0, 1]$. In neither case do these bounds include the true value, since the true $\Psi^\varepsilon$, reported in Column (4), is not in this range. The second row uses a richer specification for $r(X_i)$, which includes both linear and quadratic terms for all of the variables listed in Table 1. This improves

performance in two respects. First, the $\Psi^{\varepsilon}$ values are now both positive and between 0 and 1. In addition, the $R^2$ increases in both regressions.

The comparison between the two rows in each panel illustrates an important issue here. We build residual selection here on top of a model for selection on observables. It is, therefore, critical to consider the observable selection carefully. Using only a limited set of controls and assuming limited selection on residuals yields a misleading picture in this example, while using a richer control set yields better results. In empirical settings, we suggest including as rich a set of observables as possible, given the limits of the data. We discuss some practical details of this in the context of treatment effect estimation in Section 6. $\triangle$

# 5 Treatment Effects

While the results above are developed for an arbitrary target function $t(X_i)$, our primary interest is in the extrapolation of average treatment effects. Inference on treatment effects is complicated by the fact that even in the trial population we only observe a given individual in a single treatment state, and so we never observe treatment effects at the individual level. This means we cannot, for example, calculate the standard deviation of the treatment effect directly. We show in this section that our approach nonetheless allows us to draw inferences about average treatment effects in the target population.

To develop these results, we adopt the usual potential outcomes framework (see e.g. Imbens and Rubin 2015 for details). 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 (unobserved) full data for individual $i$ are $X_i^* = (Y_i(0), Y_i(1), D_i, C_i)$, while 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 average treatment effect (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 focus on contexts where 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 $E_{P_S}[D_i] = d$.[4] We can estimate the ATE in the trial population,

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

by 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 likewise think of the distribution $P$ as arising from hypothetical random assignment of the target population into treatment, again with $E_P[D_i] = d$. This allows us to write the target population ATE as $E_P[T_i]$. Hence we can cast estimation of average treatment effects in the target population into our general framework by taking $t(X_i) = T_i$.

Random assignment implies a number of additional restrictions which we can include as restricted moments. In particular, since we know $E_P[r(X_i)] = r_P$ in the target population, random assignment together with $E_P[D_i] = d$ implies that $E_P[r(X_i) D_i] = r_P \cdot d$. We thus recommend forming fitted values based on the interacted restriction function

$$\tilde{r}(X_i) = \left(r(X_i)', r(X_i)' D_i\right)'$$

in treatment-effect settings.

[4] While we focus on simple random assignment of $D_i$ for simplicity, 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.

**Illustration (continued):** In our running example, estimation of treatment effects proceeds exactly in parallel to estimation of means. In particular, to estimate the effect of the teacher performance pay treatment on test scores we set $t(X_i) = T_i$ for $T_i$ as described above. We again begin with the case in which we use a limited set of controls, and then introduce the full set. The results are shown in Panel C of Table 3.

In both rows, adjusting for selection on observables decreases the estimated treatment effect. The size of the adjustment is similar in both cases, reflecting the fact that school infrastructure is the observable most closely related to the treatment effect, and is included in both the limited and full set of controls. In both cases, the value $\Psi^{\varepsilon}$ to match the target population effect is between 0 and 1, so the bounds in the final column include the true target-population treatment effect. However, the $R^2$ is substantially higher for the full set of covariates, and the bounds are much tighter in this case. This illustrates another reason to include a rich set of covariates whenever possible: it will often lead to tighter bounds. $\triangle$

# 6 Implementation and Intuition

Several choices must be made to implement our approach in practice. This section describes these choices, and further discusses the intuition behind our approach. In applications, we also need to account for sampling uncertainty, and at the end of this section we briefly discuss inference.

**Implementation** Our approach relies on assumed bounds on $\Psi^{\varepsilon}$, which measures the intensity of selection on the fitted versus residual components of the treatment effect. We suggest two approaches to these bounds. First, we suggest that a natural set of bounds in many applications is $\Psi^{\varepsilon} \in [0, 1]$. This represents the case where selection on the residuals is in the same direction as, and no more intense than, selection on the fitted values. The upper bound of 1 may be compelling if, for example, we think that researchers have successfully identified most of the important variables on which the sample is selected. Second, we suggest that in many settings it may be appropriate to calculate

$$\Psi^{*,\varepsilon}(0) = \sqrt{\frac{R^2}{1-R^2}} \frac{0 - E_{P_S}[T_i]}{E_P[b(X_i)] - E_{P_S}[b(X_i)]},$$

which is the value of $\Psi^{\varepsilon}$ that would produce a treatment effect of zero. This is useful in cases where the sign of an effect (positive or negative) is important to policy. Target values other than zero could likewise be considered when appropriate.

In both of these approaches it is crucial to carefully consider the variables used to correct for selection on observables. We would ideally like to include a very rich set of transformations of the covariates in the restricted moments, though this will often be infeasible given data constraints. Nonetheless, whenever possible researchers should at a minimum include linear and squared terms in the covariates in $r(X_i)$. 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. The same point of course applies in general to adjustment for selection on observables.

**Intuition** There are two desirable features for a set of bounds, which are typically in tension. On the one hand, to be reliable a set of bounds should include the true target population treatment effect in most applications. On the other hand, bounds are more informative if they are smaller. The bounds we produce by assuming $\Psi^{\varepsilon} \in [0,1]$ include the target population effect if and only if the true $\Psi^{\varepsilon}$ is in this range. This is the fundamental assumption required for this baseline approach, and is ultimately untestable unless we know the average treatment effect in the target population, in which case there is no external validity problem to solve. Nonetheless, we should be more comfortable with this assumption in some settings than others.

The upper bound $\Psi^{\varepsilon} = 1$ is more likely to be conservative if the restricted moments - the observables - capture more of the selection. Indeed, if the observables $r(X_i)$ captured all of the selection, then under the linear specification $E[t(X_i)|C_i] = r(X_i)'\gamma$ the true value is $\Psi^{\varepsilon} = 0$ and there is no selection on residuals at all. When we allow selection on residuals, our approach relies on the belief that selection on fitted values is informative about selection on residuals and that, therefore, it is possible to draw insights about one from the other. If this is not the case then while there will still mechanically be some $\Psi^{\varepsilon}$ which produces the target population treatment effect, it is impossible to know what values of $\Psi^{\varepsilon}$ are plausible, since the selection on residuals could be enormous, even when selection on fitted values is small and there is no way to learn this from the data.

Given that we are addressing a problem of selection on *unobservables,* any approach will require some untestable assumptions. The value of the particular assumptions we consider, in our view, is their intuitive interpretation.

Taking as given that $\Psi^{\varepsilon} \in [0,1]$, the width of our bounds depend on two objects. First, the size of the selection on observables adjustment. If the treatment effect estimate is altered more by addressing selection on observables, the bounds will be wider. Second, the degree of unexplained variance $1 - R^2$. If the observables explain a greater share of $T_i$, the bounds will be tighter. Both of these are intuitive. If the observables make a great deal of difference in the treatment effect estimate, then under our assumption of similar intensity of selection on the residual we expect more selection on the residual as well. Similarly, if the residuals are small then there is little scope for additional bias from selection on unobservables.

Both the selection on observables adjustment and the $R^2$ are observed in the data, and imply that the same $\Psi^{\varepsilon} \in [0,1]$ assumption will produce larger bounds in some settings than in others. In principle these bounds can be tightened in empirical settings by expanding the set of observables used - that is, the richness of the restriction function $r(X_i)$ - to increase the share of the variability in $T_i$ explained. This will tighten the bounds as long as it does not increase the size of the selection on observables adjustment, but should be weighed against the additional variability from estimating a more flexible model.

**Inference** In our discussion thus far we have assumed that the distribution $P_S$ in the trial population is known. In practice, however, we typically have only a sample from the trial population, so the empirical distribution estimates $P_S$ with error. Likewise, while we have thus far treated $r_P = E_P[r(C_i)]$ as known, in some contexts we may also have non-negligible sampling error in our target population moments. In such settings it is necessary to account for sampling uncertainty in $t_P^L$, $t_P^U$, and $\Psi^{*,\varepsilon}(0)$. To differentiate the estimates for these quantities from their true values, we denote the estimates by $\hat{t}_P^L$, $\hat{t}_P^U$, and $\hat{\Psi}^{*,\varepsilon}(0)$.

As usual, $\hat{t}_P^L$, $\hat{t}_P^U$, and $\hat{\Psi}^{*,\varepsilon}(0)$ will be approximately normally distributed in large samples under mild conditions, and we suggest using the bootstrap to compute standard errors. Note that in settings where $r_P$ is estimated with error, we should simultaneously bootstrap samples from both the trial and target populations. We can then report $\hat{t}_P^L$, $\hat{t}_P^U$, and $\hat{\Psi}^{*,\varepsilon}(0)$ along with their associated confidence sets. To construct a confidence set for the target moment

with coverage at least 95% we can use

$$\left[\hat{t}_P^L - 1.96 \cdot \hat{\sigma}\left(\hat{t}_P^L\right), \hat{t}_P^U + 1.96 \cdot \hat{\sigma}\left(\hat{t}_P^U\right)\right],$$

for $\hat{\sigma}\left(\hat{t}_P^L\right)$ and $\hat{\sigma}\left(\hat{t}_P^U\right)$ the standard errors for $\hat{t}_P^L$ and $\hat{t}_P^U$, respectively. As noted by Imbens and Manski (2004), however, this confidence interval covers each point in the interval $\left[t_P^L, t_P^U\right]$ with probability strictly higher than 95%, and to obtain shorter intervals one can instead adopt the approaches proposed by Stoye (2009) or Elliott Mueller and Watson (2015).[5]

# 7 Applications

This section applies our approach to three representative applications.

## 7.1 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, and the question of interest is whether working from home results in productivity losses.

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 that there are substantial productivity gains - about 0.2 standard deviations on a combined productivity measure - from working from home.

Bloom et al (2015) is representative of a broader class of papers in which participants volunteer for a study and treatment is then randomized among volunteers. Examples include Gelber, Isen and Kessler (2016) on job training, and Alcott (2015) on site selection in OPower (see footnote 2 above). The concern in these papers, and in ours, is that while average treatment effect estimates from such studies are valid for the population which volunteers to

[5]These results are in progress.

participate, optimal policy may depend on the average treatment effect for the population as a whole.

In this particular case, a question of interest for the firm may be whether it would be sensible to have all call center employees work from home. If the treatment effect estimated in the experiment is valid for the entire workforce, then the answer is clearly yes. In fact, given the expense of running an office, this might be a good policy even if the average treatment effect on productivity were zero or slightly negative.

Given the selection procedure, however, it seems unlikely that the treatment effect for the experimental sample is representative of that for the population as a whole. In particular, those who select into the sample may be those who expect working from home to work for them. The in-sample treatment effect could then be biased upwards relative to the full population treatment effect. Similar concerns arise in the Gelber et al (2016) and Alcott (2015) examples discussed above.

**Results** Appendix Table 1 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 here we use these variables and allow them to enter linearly and (for non-binary variables) with a squared term, as well as interacted with treatment.

Table 4 implements our suggested calculations on external validity. Column 3 shows the baseline effect for the primary outcome in the paper, which is the increase in overall performance. Column 4 shows estimates from the regression-based covariate reweighting, which in fact increases the estimated effect. Columns 5 and 6 show two metrics of external validity. Column 5 shows bounds under the assumption that $\Psi^{\varepsilon} \in [0, 1]$. We find that the bounds on the impact do not include zero, and that under $\Psi^{\varepsilon} \in [0, 1]$ the impact could be as large as a 0.4 standard deviation increase in productivity. In Column 6 we show the other measure of selection. As implied by the fact that correcting for selection on observables *increases* the estimated treatment effect, residual selection would have to work in the opposite direction of the fitted selection to produce a treatment effect of zero.

This table also reports standard errors on all metrics. Reflecting the fairly small sample size, the estimates are somewhat noisy. Of particular note is that the standard errors around

$\Psi^{*,\varepsilon}(0)$ are extremely large. This is true throughout the applications and arises from the fact that this is a ratio and so is large in cases where the observable adjustment is small.[6]

### 7.2 Dupas and Robinson (2013)

**Setting** Our second 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 combination of outcomes and treatments which the authors suggest should be significant based on their theory. The first two columns of Table 5 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.[7] 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, making it plausible that participants differ on characteristics which relate 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.

**Results** Appendix Table 2 shows summary statistics for the population and sample in Columns 1 and 2, respectively. The population estimates are based on estimating the dif-

[6]Indeed, when the adjustment for observables is very close to zero, bootstrap standard errors for $\Psi^{*,\varepsilon}(0)$ can be unreliable. This problem is closely related to that of weak instruments, and to obtain reliable inferences on $\Psi^{*,\varepsilon}(0)$ we can adapt existing weak instrument-robust confidence sets. Empirical results reflecting these corrections are in progress. Note that this issue does not arise for $t_P^L$ and $t_P^U$ , however.

[7]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.

ference between ROSCA participants and non-participants using a second survey run in the same area.[8] The most notable differences are that the population is less female and less well educated. Because we know only the means of covariates in the target population, we can only control for observables linearly.

Table 5 shows the results, with the same format as Table 4. Most notable here is the substantial variation across outcome-treatment pairs in our sensitivity measures.

We can compare, for example, the impact of the "Safe Box" treatment (basically, a locked piggy-bank) on the three outcomes. The baseline impact is significant on all three outcomes. However, the conclusions from our bounding approach vary considerably. This derives both from differences in the selection on observables adjustment and from differences in the predictability of the treatment effect. This illustrates how one can use our approach to compare outcomes within a given paper. Even when many outcomes are significant, some are more robust than others to concerns about external validity, in the sense of surviving a wider range of values $\Psi^{\varepsilon}$.

### 7.3 Olken, Onishi, and Wong (2014)

**Setting** Olken, Onishi and Wong (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 possible participation, as were the 28% of districts which did not have access to the overall 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

[8] This is the data on which Appendix Table A8 in Dupas and Robinson (2013) is based.

in the experiment differ on some observable dimensions from the overall population.

To apply our approach in this setting, we need to identify a set of restricted moments. 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 in the restriction function. We generate these 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 values for the restricted moments are the averages for all of Indonesia.[9]

**Results** Appendix Table 3 shows summary statistics both for Indonesia overall and for the sub-districts in the study. Relative to the country overall, districts in the sample are more likely to have a dirt floor in the house and to receive cash transfers (consistent with having lower income on average) but also have higher rates of vaccination and contraception use.

Table 6 shows the results, with the same format as Tables 4 and 5. As noted, the baseline impact is insignificant for most of the outcomes. However, a notable feature of this setting is that in all cases reweighting on the observables increases the estimated size of the effects. Consequently, most of the sensitivity measures in the final column 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. For all of the outcomes, the bounds are substantially more encouraging about the impact of the experiment than the baseline effects. 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 would change behavior the most.

# 8 Discussion

While our primary focus in this paper is on external validity of average treatment effects 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

[9]The set of covariates we use for the restricted moments 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 3 the differences between the means of the covariates in the sample and target population are of the same order as the variability within the the sample.

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 (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 $P_S$. The target population is the population on the full range of $R_i$. We can then treat this problem as in the experimental case above.[10] Note, however, that relative to approaches like those proposed by Angrist and Rokkanen (2015) and Rokkanen (2015), our approach does not exploit additional structure from the regression discontinuity setting.

**Instrumental Variables** The central component of the LATE critique is that instrumental variables approaches identify the average treatment effect 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 the trial population $P_S$ as the population of compliers and the target population $P$ as the overall population, including compliers, never takers and always takers. It is then possible to proceed in the same way as above, using demographic characteristics or other variables in the data to form the restricted moments. Unlike recent work on external validity in instrumental variables models by Kowalski (2016) and Brinch et al *(forthcoming)*, however, our approach again does not exploit the additional structure imposed by the instrumental variables setting.

**Non-Treatment Effect Moments** We focus on cases where the unknown moment of interest in the target population is an average treatment effect. However, as should be clear from the development of the theory in Sections 3 and 4, our approach is not limited to estimating average treatment effects. Of particular interest may be cases where the object of interest is the mean of some variable in the target population.

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

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).

# 9 Conclusion

This paper addresses 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 residual.

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


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

**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.

**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.

**Borwein, J. M. and A. S. Lewis**, "On The Convergence of Moment Problems," *Transactions of the American Mathematical Society*, 1991.

___ **and** ___ , "Partially-Finite Programming in L1 and the Existence of Maximum Entropy Estimates," *Siam Journal of Optimization*, 1993.

**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*, Forthcoming.

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

**Elliott, Graham, Ulich K. Mueller, and Mark Watson**, "Nearly Optimal Tests when a Nuisance Parameter is Present Under the Null Hypothesis," *Econometrica*, 2015, *83*, 771–811.

**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.

**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.

**Hansen, Lars Peter and Ravi Jagannathan**, "Assessing Specification Errors in Stochastic Discount Factor Models," *Journal of Finance*, 1997, *52* (2), 557–590.

**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 Charles F. Manski**, "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 2004, *72* (6), 1845–1857.

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

**Kitamura, Yuichi**, "Empirical Likelihood Methods in Econometrics: Theory and Practice," in Whitney Newey Richard Blundell and Torsten Persson, eds., *Advances in Economics and Econometrics*, 1 ed., Vol. 3, Cambridge: Cambridge University Press, 2007, pp. 174–237.

**Kline, Patrick and Christopher Walters**, "Evaluating Public Programs with Close Substitutes: The Case of Head Start," *Quarterly Journal of Economics*, forthcoming.

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

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

**Newey, Whitney K and Richard J Smith**, "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 2004, *72* (1), 219–255.

**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.

**Owen, Art**, "Empirical likelihood for linear models," *The Annals of Statistics*, 1991, pp. 1725–1747.

**Pearson, Karl**, "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1900, *50* (302), 157–175.

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

**Rubin, Donald B**, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, 1974, *66* (5), 688.

___ , "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational and Behavioral statistics*, 1977, *2* (1), 1–26.

**Stoye, Jorg**, "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 2009, *77* (4), 1299–1315.

**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.

Table 1: **Bias Decomposition**

| Variable | Trial Pop. Mean | Target Pop. Mean | $\sigma_{P_S}(f_j(X_i))$ | $\rho_{P_S}(W_i, f_j(X_i))$ | $\sigma_{P_S}(W_i)$ |
|---|---|---|---|---|---|
| Share Teachers Present | 0.815 | 0.818 | 0.107 | 0.090 | 0.288 |
| Teacher Salary | 7886.9 | 8033.5 | 2364.7 | 0.215 | 0.288 |
| Teacher Training | 2.64 | 2.61 | 0.451 | -0.208 | 0.288 |
| Avg. HH Income | 3.64 | 3.63 | 0.522 | -0.084 | 0.288 |
| School Infrastructure | 3.27 | 2.96 | 1.31 | -0.815 | 0.288 |

*Notes*: This table illustrates the bias decomposition in our running example. The data is sampled with a known scheme so we know the weights. The moments of interest are the means of these variables. $\rho_{P_S}(W_i, f_j(X_i))$ is the correlation between the weights and the moments. $\sigma_{P_S}(W_i)$ is the standard deviation of the weights and $\sigma_{P_S}(f_j(X_i))$ is the standard deviation of the moments. The data are constructed based on Muralidharanand Sundararman (2011).

Table 2: **Extrapolation to Unknown Moments**

| **Panel A: Teacher Education** | | | |
|---|---|---|---|
| *Samp. Mean = 3.062, Pop Mean = 3.028, Corr w/ True Weights = -0.252* | | | |
| | (1) | (2) | (3) |
| *Restricted Moment* | $\Psi$ to Match | Adjusted Mean Educ | Bounds, $\Psi \in [-1, 1]$ |
| Share Teachers Present | -2.83 | 3.028 | [3.050,3.074] |
| Teacher Salary | -1.19 | 3.028 | [3.033, 3.090] |
| Teacher Training | 1.23 | 3.028 | [3.034,3.089] |
| Avg. HH Income | 3.06 | 3.028 | [3.050,3.073] |
| School Infrastructure | 0.31 | 3.028 | [2.953,3.170] |
| **Panel B: Baseline Test Score** | | | |
| *Samp Mean = 0.019, Pop Mean = 0.023, Corr w/ True Weights = 0.049* | | | |
| | (1) | (2) | (3) |
| *Restricted Moment* | $\Psi$ to Match | Adjusted Mean Score | Bounds, $\Psi \in [-1, 1]$ |
| Share Teachers Present | 0.56 | 0.023 | [0.010,0.027] |
| Teacher Salary | 0.23 | 0.023 | [-0.001,0.038] |
| Teacher Training | -0.24 | 0.023 | [-0.0002,0.037] |
| Avg. HH Income | -0.60 | 0.023 | [0.011,0.026] |
| School Infrastructure | -0.062 | 0.023 | [-0.056,0.094] |

*Notes*: This table illustrates the extrapolation approach detailed in Section 4.1. The data are based on Muralidharan and Sundararman (2011).

Table 3: **Extrapolation Using Residual Selection**

| | | | | | |
|---|---|---|---|---|---|
| **Panel A: Teacher Education** | | | | | |
| *Target Population Mean: 3.028* | | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| | Baseline Sample Mean | Adj. for Sel on Obs. | $R^2$ of Obs. | $\Psi$, Match Target | Bounds, $\Psi \in [0,1]$ |
| Fewer Restricted Moments | 3.062 | 3.061 | 0.052 | 5.28 | [3.054,3.061] |
| More Restricted Moments | 3.062 | 3.043 | 0.378 | 0.614 | [3.018,3.043] |
| **Panel B: Baseline Test Score** | | | | | |
| *Target Population Mean: 0.023* | | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| | Baseline Sample Mean | Adj. for Sel on Obs. | $R^2$ of Obs. | $\Psi$, Match Target | Bounds, $\Psi \in [0,1]$ |
| Fewer Restricted Moments | 0.018 | 0.027 | 0.011 | -0.047 | [0.027,0.102] |
| More Restricted Moments | 0.018 | 0.023 | 0.102 | 0.005 | [0.023,0.037] |
| **Panel C: Treatment Effect on Test Score** | | | | | |
| *Target Population Effect: 0.114* | | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| | Baseline Sample Mean | Adj. for Sel on Obs. | $R^2$ of Obs. | $\Psi$, Match Target | Bounds, $\Psi \in [0,1]$ |
| Fewer Restricted Moments | 0.223 | 0.196 | 0.007 | 0.28 | [-0.100,0.196] |
| More Restricted Moments | 0.223 | 0.186 | 0.095 | 0.64 | [0.073,0.186] |

*Notes*: This table shows the extrapolation approach detailed in Sections 4 and 5, using the fitted values as the restricted moment for comparison. The first row in each panel uses a smaller set of moments (linear controls for school infrastructure and average teacher salary). The second row uses all of the variables in Table 1 and includes both linear and squared terms. In Panel C we also include controls for treatment and treatment interacted with the other moments. The data are constructed based on Muralidharan and Sundararman (2011).

Table 4: **Application: Bloom et al (2015)**

| *Outcome* | *Treatment* | Baseline Effect | Observable Adjusted | Bounds, $\Psi \in [0,1]$ | $\Psi^{*,\varepsilon}(0)$ |
|---|---|---|---|---|---|
| Overall Performance | Treatment | 0.206 | 0.258 | [0.259;0.431] | -1.49 |
| | | (0.165, 0.246) | (0.198,0.319) | (0.198, 0.589) | (-9.97, 6.97) |

*Notes*: This table shows the application of our sensitivity procedure to Bloom et al (2015). The restricted moments comes from the study. Standard errors are bootstrapped.

Table 5: **Application: Dupas and Robinson (2013)**

| *Outcome* | *Treatment* | Baseline Effect | Observable Adjusted | Bounds, $\Psi \in [0,1]$ | $\Psi^{*,\varepsilon}(0)$ |
|---|---|---|---|---|---|
| Investment in Health | Safe Box | 165.9 | 103.0 | [25.5, 103.0] | 1.32 |
| | | (26.9, 304.9) | (-11.4,217.5) | (-138.4, 217.5) | (-77.1, 79.7) |
| Investment in Health | Locked Box | 48.33 | 17.18 | [-18.10, 17.18] | 0.486 |
| | | (-56.1, 152.8) | (-74.4,108.8) | (-146.3,108.8) | (-33.8, 34.8) |
| Investment in Health | Health Pot | 287.8 | 211.1 | [112.4, 211.1] | 2.13 |
| | | (121.8, 453.8) | (62.4, 360.3) | (-147.4, 360.3) | (-79.2, 83.4) |
| Trouble Affording Treatment | Safe Box | -0.111 | -0.101 | [-0.101, -0.087] | 7.23 |
| | | (-0.250, 0.028) | (-0.236, 0.033) | (-0.236, 0.092) | (-4.42, 18.9) |
| Trouble Affording Treatment | Health Savings | -0.134 | -0.130 | [-0.130,-0.124] | 22.62 |
| | | (-0.268, 0.0001) | (-0.252, -0.007) | (-0.252, 0.046) | (-23.8, 69.1) |
| Reached Health Goal | Safe Box | 0.155 | 0.113 | [0.066, 0.113] | 2.40 |
| | | (0.002, 0.309) | (-0.022, 0.284) | (-0.157, 0.284) | (-72.4, 77.3) |
| Reached Health Goal | Locked Box | -0.020 | -0.052 | [-0.094, -0.052] | -1.24 |
| | | (-0.159, 0.118) | (-0.170, 0.066) | (-0.282, 0.066) | (-11.0, 8.51) |
| Reached Health Goal | Health Pot | 0.120 | 0.134 | [0.134, 0.151] | -8.10 |
| | | (-0.034, 0.275) | (-0.003, 0.272) | (-0.003, 0.357) | (-68.6, 52.5) |
| Reached Health Goal | Health Savings | 0.056 | 0.017 | [-0.030,0.017] | 0.363 |
| | | (-0.097, 0.209) | (-0.125, 0.160) | (-0.199, 0.160) | (-137.7, 139.5) |

*Notes*: This table shows the application of our sensitivity procedure to Dupas and Robinson (2013). The restricted moments are generated using evidence from an auxiliary survey measuring differences between participants and non-participants. Standard errors are bootstrapped.

Table 6: **Application: Olken et al (2014)**

| *Outcome* | *Treatment* | Baseline Effect | Observable Adjusted | Bounds, $\Psi \in [0,1]$ | $\Psi^{*,\varepsilon}$ (0) |
|---|---|---|---|---|---|
| Prenatal Visits | Incentive Treatment | 0.198 | 0.521 | [0.521,0.694] | -3.00 |
| | | (-0.505, 0.902) | (0.116, 0.925) | (0.116, 1.28) | (-48.4, 42.4) |
| Assisted Delivery | Incentive Treatment | 0.008 | 0.038 | [0.038,0.056] | -2.24 |
| | | (-0.074, 0.089) | (-0.008, 0.086) | (-0.008, 0.135) | (-27.0, 22.5) |
| Postnatal Visits | Incentive Treatment | -0.197 | -0.014 | [-0.014,0.228] | 0.057 |
| | | (-0.44, 0.048) | (-0.32, 0.28) | (-0.32, 0.77) | (-6.7, 6.8) |
| Iron Pills | Incentive Treatment | 0.045 | 0.089 | [0.089, 0.117] | -3.13 |
| | | (-0.137, 0.229) | (-0.052, 0.231) | (-0.061, 0.34) | (-190.8, 184.5) |
| Immunization | Incentive Treatment | 0.004 | 0.011 | [0.011,0.015] | -3.27 |
| | | (-0.054, 0.062) | (-0.02, 0.043) | (-0.02, 0.062) | (-14.6, 8.1) |
| No. Weight Checks | Incentive Treatment | 0.147 | 0.210 | [0.210,0.239] | -7.28 |
| | | (-0.009, 0.304) | (0.106, 0.315) | (0.106, 0.387) | (-88.6, 74.0) |
| Vitamin A Supplements | Incentive Treatment | 0.015 | 0.049 | [0.049,0.070] | -2.33 |
| | | (-0.148, 0.179) | (-0.056, 0.154) | (-0.056, 0.235) | (-25.0, 20.4) |
| Malnourished | Incentive Treatment | 0.002 | 0.005 | [0.005,0.011] | -0.838 |
| | | (-0.026, 0.030) | (-0.030, 0.040) | [(-0.030, 0.077) | (-18.1, 16.4) |

*Notes*: This table shows the application of our sensitivity procedure to Olken et al (2014). Restricted moment values are generated using on location-level variables from a nationally representative survey. Standard errors are bootstrapped.

## Appendix A: Proofs

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

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

while by Bayes Theorem we can take

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

Thus,

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

**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,

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

As noted in the text, however, $E_{P_S}[W_i] = 1$ by Lemma 2, so the result follows by another application of Lemma 2.

**Proof of Proposition 1** We prove the lower bound: the upper bound follows by an analogous argument. Using equation (3), note that if we consider the ratio of the bias in $t(X_i)$ to the bias in $b(X_i)$, normalizing each by their standard deviation, we recover the correlation ratio:

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

Thus,

$$\frac{(E_P[t(X_i)] - E_{P_S}[t(X_i)])/\sigma_{P_S}(t(X_i))}{(E_P[b(X_i)] - E_{P_S}[b(X_i)])/\sigma_{P_S}(b(X_i))} \geq \Psi_L.$$

Rearranging and using the fact that the denominator on the left hand side is assumed to be strictly positive yields the result.

**Proof of Lemma 3** Note that

$$\frac{\rho_{P_S}(W_i, t(X_i))}{\rho_{P_S}(W_i, b(X_i))} = \frac{Cov_{P_S}(W_i, t(X_i))/\sigma_{P_S}(t(X_i))}{Cov_{P_S}(W_i, b(X_i))/\sigma_{P_S}(b(X_i))}$$

$$= \frac{(Cov_{P_S}(W_i, b(X_i)) + Cov_{P_S}(W_i, \varepsilon_i))/\sigma_{P_S}(t(X_i))}{Cov_{P_S}(W_i, b(X_i))/\sigma_{P_S}(b(X_i))}$$

$$= \frac{\sigma_{P_S}(b(X_i))}{\sigma_{P_S}(t(X_i))} + \frac{\rho_{P_S}(W_i, \varepsilon_i)}{\rho_{P_S}(W_i, b(X_i))} \cdot \frac{\sigma_{P_S}(\varepsilon_i)}{\sigma_{P_S}(t(X_i))},$$

from which the result follows.

**Proof of Corollary 2** We again prove the result for the lower bound, while the result for the upper bound follows by an analogous argument. Recall from Proposition 1 that

$$t_P^L = E_{P_S}[t(X_i)] + \Psi_L \frac{\sigma_{P_S}(t(X_i))}{\sigma_{P_S}(b(X_i))} \left(E_P[b(X_i)] - E_{P_S}[b(X_i)]\right).$$

Substituting in the expression for $\Psi_L$ from Lemma 3 and using the fact that $E_{P_S}[b(X_i)] = E_{P_S}[t(X_i)]$, we see that

$$t_P^L = E_P[b(X_i)] + \frac{\sigma_{P_S}(\varepsilon_i)}{\sigma_{P_S}(b(X_i))} \left(E_P[b(X_i)] - E_{P_S}[b(X_i)]\right),$$

from which the result follows.

# Appendix B: Derivations under Bounds on $\sigma_{P_S}(W_i)$

Here, we discuss how we can bound the bias of the trial population moments using bounds on the standard deviation of the weights.

## B.1 Bounding Bias using Reweighting

We note in Section 3.2 that it is possible to bound the bias by bounding the standard deviation of the weights. Recall that display (5) from the main text states that

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

The true weights $W_i$ must match the true value of our target moment, along with any additional moments we know from the target population. If we have some upper bound on the standard deviation of the weights, say $\sigma_{P_S}(W_i) \leq c$, we can thus seek the smallest and largest values for our target moment $E_P[t(X_i)]$ consistent with this restriction, along with any auxiliary restrictions on other moments whose value in the target population is known. This requires optimizing over the set of all possible weights, and so may sound computationally daunting, but turns out to be highly computationally tractable.

### Minimum Reweighting

Recall our notation for a set of moments whose mean is known in *both* the trial and target populations. We can represent these moment restrictions in the target population as $E_P[r(X_i)] = r_P \in \mathbb{R}^k$ for a vector-valued moment function $r(X_i)$ and a known vector of constants $r_P$. In this section, we study the implications of these restrictions for the standard deviation of the weights $\sigma_{P_S}(W_i)$.

Equation (4) in the main text implies a different lower bound on the standard deviation of the weights for each moment restriction, since for each $j$

$$\sigma_{P_S}(W_i) \geq \left| \frac{E_{P_S}[r_j(X_i)] - r_{P,j}}{\sigma_{P_S}(r_j(X_i))} \right|. \tag{11}$$

These bounds are obtained by setting the absolute value of the correlation between $r_j(X_i)$ and the weights to one. Note, however, that unless the different $r_j(X_i)$ are perfectly correlated with each other it is impossible for the true weights to be perfectly correlated with all of these elements simultaneously. Hence this bound, by considering only one moment at a time, discards the information contained in the covariances of the moment functions.

To obtain bounds which use all of the moment restrictions simultaneously, we search over the space of all possible weights $\frac{\tilde{p}_X(X_i)}{p_{X,S}(X_i)} = \tilde{W}_i$ consistent with the restrictions that $E_{P_S}\left[\tilde{W}_i\right] = 1$ and $E_{P_S}\left[\tilde{W}_i r(X_i)\right] = r_P$ and attempt to find the weights yielding the smallest standard deviation. Our next result shows that if we relax the optimization problem by allowing $\tilde{W}_i$ to be negative, the resulting optimal $\tilde{W}_i$ and $\sigma_{P_S}\left(\tilde{W}_i\right)$ both have simple expressions.

**Proposition 2** *Let $\mathcal{P}_{X,\pm}(r, r_P)$ be the class of signed measures $\tilde{P}_X$ for $X$ which are absolutely continuous with respect to $P_{X,S}$ and satisfy $E_{P_S}\left[\tilde{W}_i\right] = 1$, $E_{P_S}\left[\tilde{W}_i r(X_i)\right] = r_P$ for $\tilde{W}_i = \tilde{p}_X(X_i)/p_{X,S}(X_i)$. Provided $Var_{P_S}(r(X_i))$ is non-singular, the minimum standard deviation $\sigma_{P_S}\left(\tilde{W}_i\right)$ consistent with these restrictions is*

$$\begin{aligned}\min\left\{\sigma_{P_S}\left(\tilde{W}_i\right) : \tilde{P}_X \in \mathcal{P}_{X,\pm}(r, r_P)\right\} &= d\left(P_{X.S}, \mathcal{P}_{X,\pm}(r, r_P)\right) \\ &= \sqrt{(r_{P_S} - r_P)' Var_{P_S}(r(X_i))^{-1}(r_{P_S} - r_P)},\end{aligned} \tag{12}$$

*for $r_{P_S} = E_{P_S}[r(X_i)]$. Further, this minimal standard deviation is attained by $\tilde{P}_X \in \mathcal{P}_{X,\pm}(r, r_P)$ with*

$$\tilde{W}_i = 1 + (r_P - r_{P_S})' Var_{P_S}(r(X_i))^{-1}(r(X_i) - r_{P_S}).$$

We can think of $d\left(P_{X,S}, \mathcal{P}_{X,\pm}(r, r_P)\right)$ as the distance between the distribution of $X$ in the trial population and the closest measure in the family $\mathcal{P}_{X,\pm}(r, r_P)$ which satisfies our moment restrictions. This minimal distance is equal to the square root of the variance-weighted squared distance between the moments in the trial and target populations.[11] This result is closely related to a number of previous findings including results in Owen (1991), Hansen and Jagannathan (1997), Newey and Smith (2004), and Kitamura (2007).

We have assumed that the distribution $P_X$ in the target population is absolutely continuous with respect to that in the trial population, which implies that $P_X \in \mathcal{P}_{X,\pm}(r, r_P)$ provided our moment restriction $E_{P_X}[r(X_i)] = r_P$ is correct. Consequently, Proposition 2 implies a lower bound on $\sigma_{P_S}(W_i)$.

**Corollary 3** *Under Assumption 1, if $E_{P_X}[r(X_i)] = r_P$ then*

$$\sigma_{P_S}(W_i) \geq \sqrt{(r_{P_S} - r_P)' Var_{P_S}(r(X_i))^{-1}(r_{P_S} - r_P)}.$$

Thus, if we know a vector of moments $r_P$ in the target population we obtain a simple lower bound on the standard deviation of the true weights.

While intuitive, the bound in Corollary 3 is often slack, since the true weights $W_i$ are non-negative but we do not enforce this restriction in Proposition 2. To obtain a sharper bound, our next result re-imposes this sign restriction. This yields a sharp lower bound on $\sigma(W_i)$, but the form of this bound is less transparent than the bound obtained above. This result is also implied by the results of Hansen and Jagannathan (1997).

**Proposition 3** *Let $\mathcal{P}_X(r, r_P)$ be the class of distributions for $X$ which are absolutely continuous with respect to $P_{X,S}$ and satisfy $E_{\tilde{P}_X}[r(X_i)] = r_P$. If $P_X \in \mathcal{P}_X(r, r_P)$ and $P_X$, $P_{X,S}$ are mutually absolutely continuous then*

$$\begin{aligned}\min\left\{\sigma_{P_S}\left(\tilde{W}_i\right) : \tilde{P}_X \in \mathcal{P}_X\right\} &= d\left(P_{X,S}, \mathcal{P}_X(r, r_P)\right) \\ &= \sqrt{\max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^k}\left(\lambda - E_{P_S}\left[\tfrac{1}{4}\left(\lambda + \gamma'(r(X_i) - r_P)\right)_+^2 + 1\right]\right)},\end{aligned} \tag{13}$$

[11] This is in turn equal to the Mahalanobis distance from the vector $r_P$ to the distribution of $r(X_i)$ under $P_S$.

*where* $(A)_+ = \max\{A, 0\}$. *Further, this minimal standard deviation is attained by* $\tilde{P}_X$ *with*

$$\tilde{W}_i = \frac{1}{2}\max\left\{\bar{\lambda} + \bar{\gamma}'\left(r\left(X_i\right) - r_P\right), 0\right\},$$

*where* $\left(\bar{\lambda}, \bar{\gamma}\right)$ *are the optimal values in equation (13).*

**Corollary 4** *If* $E_P\left[r\left(X_i\right)\right] = r_P$ *and* $P_X$, $P_{X,S}$ *are mutually absolutely continuous, then*

$$\sigma_{P_S}\left(W_i\right) \geq \sqrt{\max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^k}\left(\lambda - E_{P_S}\left[\frac{1}{4}\left(\lambda + \gamma'\left(r\left(X_i\right) - r_P\right)\right)_+^2 + 1\right]\right)},$$

*and this bound is sharp, in the sense that it is the largest possible lower bound.*

Relative to Corollary 3, Corollary 4 derives a tighter bound on the standard deviation of the weights, though the two bounds can coincide if the weights $\tilde{W}_i$ in Proposition 2 are always weakly positive. At the same time, however, the bound in Corollary 4 is harder to interpret and is not generally available in closed form.[12]

### Bounding $E_P\left[t\left(X_i\right)\right]$ Under Bounds on $\sigma_{P_S}\left(W_i\right)$

So far we have discussed using Propositions 2 and 3 and moment restrictions $E_P\left[r\left(X_i\right)\right] = r_P$ to learn about the standard deviation $\sigma_{P_S}\left(W_i\right)$ of the weights. If we know something about the standard deviation of the weights $\sigma_{P_S}\left(W_i\right)$, however, then we can also use these results to derive bounds on the target moment $E_P\left[t\left(X_i\right)\right]$. We take as given a bound $\sigma_{P_S}\left(W_i\right) \leq c$ for some constant $c$ and explore the implications for the possible values of $t_P = E_P\left[t\left(X_i\right)\right]$ where we now allow $t$ to be multi-dimensional, $t\left(X_i\right) \in \mathbb{R}^q$.

In particular, note that under our assumed bound on the standard deviation of the weights, since the true weights satisfy

$$E_{P_S}\left[W_i\begin{pmatrix} r\left(X_i\right) \\ t\left(X_i\right)\end{pmatrix}\right] = \begin{pmatrix} r_P \\ t_P\end{pmatrix},$$

for any collection of restricted moments $E_P\left[r\left(X_i\right)\right] = r_P$, it follows that

$$d\left(P_{X,S}, \mathcal{P}_X\left(\left(r, t\right), \left(r_P, t_P\right)\right)\right) \leq c.$$

Thus, we obtain the following result:

[12] In addition, Proposition 3 adds the assumption that $P_{X,S}$ is absolutely continuous with respect to $P_X$. Without this assumption it remains the case that

$$d\left(P_{X,S}, \mathcal{P}_X\left(r, r_P\right)\right) \geq \sqrt{\sup_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^k}\left(\lambda - E_{P_S}\left[\frac{1}{4}\left(\lambda + \gamma'\left(r\left(X_i\right) - r_P\right)\right)_+^2 + 1\right]\right)},$$

and that if the optimum in the right hand side is attained at some point $\left(\bar{\lambda}, \bar{\gamma}\right)$ then the optimal weights are of the form stated in Proposition 3. However, without absolute continuity the above inequality may be strict in some contexts, and the optimum in the right hand side need not be attained. Moreover, if our moment restriction is incorrect and $P_X \notin \mathcal{P}_X\left(r, r_P\right)$, the lower bound on the distance is infinite, which reflects the fact that there are no weights which match the desired moments.

**Corollary 5** *The set of values $t_P$ consistent with $\sigma_{P_S}(W_i) \leq c$ and $E_P[r(X_i)] = r_P$ is*

$$\{t_P : d(P_{X,S}, \mathcal{P}_X((r,t),(r_P,t_P))) \leq c\}.$$

*In particular, under the conditions of Proposition 3 this set is*

$$\left\{t_P : \sqrt{\max_{\lambda\in\mathbb{R},\delta\in\mathbb{R}^q,\gamma\in\mathbb{R}^k}\left(\lambda - E_{P_S}\left[\frac{1}{4}\left(\lambda + \gamma'(r(X_i) - r_P) + \delta'(t(X_i) - t_P)\right)_+^2 + 1\right]\right)} \leq c\right\}. \tag{14}$$

Thus, using Proposition 3 we can easily calculate the set of values for $t_P$ consistent with both our moment restrictions and a given value of $c$. This result gives us a tractable way to translate bounds on the distance $\sigma_{P_S}(W_i)$ between the trial and target populations, along with any additional moment restrictions, to an implied set of values for $t_P$ in the target population.

**Upper and Lower Bounds** To bound the set of possible values the target moment, we combine upper and lower bounds on $\sigma_{P_S}(W_i)$. In particular, given an upper bound on the distance between the trial and target populations, and thus on $\sigma_{P_S}(W_i)$, we can use our minimum-reweighting results to bound $t_P$. For this purpose it is critical that Proposition 3 yields a *lower* bound on $\sigma_{P_S}(W_i)$. If it instead yielded an upper bound then we would need a lower bound on $\sigma_{P_S}(W_i)$ from some other source to bound $t_P$.

## B.2: Proofs for Additional Results

**Proof of Proposition 2** This result follows from Theorem 3.4 of Borwein and Lewis (1993). Note, in particular, that since

$$\sigma_{P_S}^2\left(\tilde{W}_i\right) = E_{P_S}\left[\tilde{W}_i^2\right] - 1,$$

we can re-write the optimization problem under consideration as

$$\min_{\tilde{W}_i} E_{P_S}\left[\tilde{W}_i^2 - 1\right]$$

$$\text{subject to}: E_{P_S}\left[\tilde{W}_i \tilde{r}(X_i)\right] = \tilde{r}_P = \begin{pmatrix} 1 \\ r_P - r_{P_S} \end{pmatrix}$$

where $\tilde{r}(X_i) = \left(1, (r(X_i) - r_{P_S})'\right)'$ for $r_{P_S} = E_{P_S}[r(X_i)]$. This a precisely the sort of problem considered in Borwein and Lewis (1993), where in their notation $\phi(u) = u^2 - 1$. Define $(p,q) = (-\infty, \infty)$, again in the notation of Borwein and Lewis (1993), and note that the convex conjugate of $\phi$ is

$$\phi^*(y) = \sup_u (yu - \phi(u)) = \frac{1}{4}y^2 + 1.$$

Note that $\phi^*$ is essentially smooth and that $\phi(y)$ is both essentially smooth and essentially strictly convex, and thus is of Legendre type (see Borwein and Lewis for definitions). Note,

further, that the primal and dual constraint qualifications discussed by Borwein and Lewis (1993) hold trivially in this setting.

Thus, by Theorem 3.4 of Borwein and Lewis (1993), the minimized value in (12) is equal to

$$\sup_{\psi}\left(\tilde{r}_P'\psi - E_{P_S}\left[\phi^*\left(\psi'\tilde{r}\left(X_i\right)\right)\right]\right) = \tilde{r}_P'\psi - E_{P_S}\left[\frac{1}{4}\left(\psi'\tilde{r}\left(X_i\right)\right)^2 + 1\right]. \tag{15}$$

Taking first order conditions yields that the optimal value $\bar{\psi}$ solves

$$\tilde{r}_P - \frac{1}{2}E_{P_S}\left[\tilde{r}\left(X_i\right)\tilde{r}\left(X_i\right)'\right]\bar{\psi} = 0$$

and thus that $\bar{\psi} = 2E_{P_S}\left[\tilde{r}\left(X_i\right)\tilde{r}\left(X_i\right)'\right]^{-1}\tilde{r}_P$. Plugging this back in to (15) yields

$$\sup_{\psi}\left(\tilde{r}_P'\psi - E_{P_S}\left[\phi^*\left(\psi'\tilde{r}\left(X_i\right)\right)\right]\right) = \tilde{r}_P'E_{P_S}\left[\tilde{r}\left(X_i\right)\tilde{r}\left(X_i\right)'\right]^{-1}\tilde{r}_P - 1.$$

Note, however, that

$$E_{P_S}\left[\tilde{r}\left(X_i\right)\tilde{r}\left(X_i\right)'\right] = \begin{bmatrix} 1 & 0 \\ 0 & Var_{P_S}\left(r\left(X_i\right)\right) \end{bmatrix},$$

and thus that

$$\tilde{r}_P'E_{P_S}\left[\tilde{r}\left(X_i\right)\tilde{r}\left(X_i\right)'\right]^{-1}\tilde{r}_P - 1 = \left(r_P - r_{P_S}\right)'Var_{P_S}\left(r\left(X_i\right)\right)^{-1}\left(r_P - r_{P_S}\right),$$

which immediately implies the first part of the proposition.

We prove the second part of the proposition directly. In particular, define

$$\tilde{W}_i = \frac{1}{2}\bar{\psi}'\tilde{r}\left(X_i\right) = 1 + \left(r_P - r_{P_S}\right)'Var_{P_S}\left(r\left(X_i\right)\right)^{-1}\left(r\left(X_i\right) - r_{P_S}\right)$$

as in the statement of the proposition, and note that $E_{P_S}\left[\tilde{W}_i\right] = 1$ while

$$E_{P_S}\left[\tilde{W}_i r\left(X_i\right)\right] = r_{P_S} + \left(r_P - r_{P_S}\right) = r_P.$$

Moreover, we can directly verify that

$$\sigma^2_{P_{X,S}}\left(\tilde{W}_i\right) = E_{P_S}\left[\tilde{W}_i^2\right] - 1 = \left(r_{P_S} - r_P\right)'Var_{P_S}\left(r\left(X_i\right)\right)^{-1}\left(r_{P_S} - r_P\right),$$

as desired.

**Proof of Proposition 3** Because it may sometimes be useful to impose an upper bound on the weights, we prove a slightly more general result

$$\begin{array}{c} \min\left\{\sigma_{P_S}\left(\tilde{W}_i\right) : \tilde{P}_X \in \mathcal{P}_X^c\left(r, r_P\right)\right\} = d\left(P_{S,X}, \mathcal{P}_X^c\left(r, r_P\right)\right) \\ = \sqrt{\max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q}\left(\lambda - E_{P_S}\left[\phi^*\left(\lambda + \gamma'\left(r\left(X_i\right) - r_P\right)\right)\right]\right)}, \end{array} \tag{16}$$

where $\mathcal{P}_X^c(r, r_P)$ is the class of distributions for $X$ which are absolutely continuous with respect to $P_{X,S}$, satisfy $E_{\tilde{P}_X}[r(X_i)] = r_P$, and have $\tilde{W}_i \leq c$ almost surely for some (possibly infinite) constant $c$, while

$$\phi^*(y) = \sup_u (yu - \phi(u)) = \begin{cases} 1 & \text{if } y < 0 \\ \frac{1}{4}y^2 + 1 & \text{if } y \in [0, 2c] \\ yc - c^2 + 1 & \text{if } y > 2c \end{cases}.$$

Correspondingly, we show that the optimal weights are given by

$$\tilde{W}_i = \min\left\{\max\left\{\bar{\lambda} + \bar{\gamma}'(r(X_i) - r_P), 0\right\}, c\right\}.$$

Setting $c = \infty$, these results then imply those of Proposition 3.

The first part of the result again follows from Theorem 3.4 of Borwein and Lewis (1993). In particular, define

$$\phi(u) = u^2 + \infty \cdot 1\{u \notin [0, c]\} - 1,$$

and note that we can re-write the optimization problem (16) as

$$\min_{\tilde{W}_i} E_{P_{X,S}}\left[\phi\left(\tilde{W}_i\right)\right]$$

$$\text{subject to : } E_{P_{X,S}}\left[\tilde{W}_i \tilde{r}(X_i)\right] = \tilde{r}_P = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

where we let $\tilde{r}(X_i) = \left(1, (r(X_i) - r_P)'\right)'$. Note, further, that $\phi^*$ as defined above is the convex conjugate of $\phi$,

$$\phi^*(y) = \sup_u (yu - \phi(u)).$$

The dual constraint qualification holds in this example by construction since $\tilde{r}(X_i)$ contains a constant. Moreover, we have assumed that $P_X$ and $P_{X,S}$ are mutually absolutely continuous, which implies that $W_i > 0$ almost surely under $P_{X,S}$. Since $P_X \in \mathcal{P}(r, r_P)$, however, this implies that the primal constraint qualification of Borwein and Lewis (1993) holds. Thus, Theorem 3.4 of Borwein and Lewis (1993) implies that the minimized value in (16) is equal to

$$\sup_\psi \left(\tilde{r}_P'\psi - E_{P_S}\left[\phi^*\left(\psi'\tilde{r}(X_i)\right)\right]\right)$$

where $\psi = (\lambda, \gamma)$. Note however that $\tilde{r}_P'\psi = \lambda$, while $\psi'\tilde{r}(X_i) = \lambda + \gamma'(r(X_i) - r_P)$. Thus (16) holds.

Next, let us turn to the second part of the statement. Note that $\phi$ is essentially strictly convex, since $\phi^*$ is essentially smooth, and that the integrability condition of Borwein and Lewis (1993) holds trivially in this case (since $p = \infty$, $q = \infty$, and $\tilde{r}(X_i)$ is finite for all $X_i$). Theorem 4.1 of Borwein and Lewis implies that the optimum in (16) is uniquely attained, and that this optimal solution corresponds to weights

$$\tilde{W}_i = \frac{\partial}{\partial y}\phi^*\left(\bar{\lambda} + \bar{\gamma}'(r(X_i) - r_P)\right)$$

$$= \begin{cases} 0 & \text{if } \bar{\lambda} + \bar{\gamma}' \left( r \left( X_i \right) - r_P \right) < 0 \\ \frac{1}{2} \left( \bar{\lambda} + \bar{\gamma}' \left( r \left( X_i \right) - r_P \right) \right) & \text{if } \bar{\lambda} + \bar{\gamma}' \left( r \left( X_i \right) - r_P \right) \in [0, 2c] \\ c & \text{if } \bar{\lambda} + \bar{\gamma}' \left( r \left( X_i \right) - r_P \right) > 2c. \end{cases}$$

**Proof of Corollary 4** The fact that this bound is sharp follows from the fact that we can construct distributions $P_X^*$ which satisfy the assumed properties of $P_X$ but approximate the optimal solution $\tilde{P}_X$ in Proposition 3 arbitrarily closely.

**Proof of Corollary 5** This follows immediately from Proposition 3.

# Appendix C: Tables and Figures

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

| Variable | Population: Mean (SD) | Sample: Mean (SD) |
|---|---|---|
| Age | 23.2 (3.28) | 24.4 (3.54) |
| Gross Wage | 2.89 (2.82) | 2.97 (0.79) |
| Any Children | 0.130 (0.336) | 0.176 (0.381) |
| Married | 0.227 (0.419) | 0.274 (0.447) |
| Male | 0.400 (0.490) | 0.471 (0.500) |
| At Least Tertiary Educ | 0.422 (0.494) | 0.381 (.486) |
| Commute Time (Min) | 95.7 (60.25) | 112.9 (63.44) |

*Notes*: This table illustrates the restricted moments in the sample and population for the Bloom et al (2015) paper.

Table 2: **Observable Sample Selection, Dupas and Robinson (2016)**

| Variable | Population: Mean | Sample: Mean |
|---|---|---|
| Age | 40.95 | 39.03 |
| Female | 0.681 | 0.737 |
| Female X Married | 0.495 | 0.555 |
| Hyperbolic | 0.152 | 0.159 |
| Time Inconsistent | 0.175 | 0.177 |
| High Discount Rate | 0.467 | 0.442 |
| Education | 5.67 | 6.31 |

*Notes*: This table illustrates the restricted 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 3: **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 restricted 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.