Weighting for External Validity*

Isaiah Andrews
MIT and NBER

Emily Oster
Brown University and NBER

January 19, 2018

Abstract

External validity is a challenge in treatment effect estimation. Even in randomized trials, the experimental sample often differs from the population of interest. If participation decisions are explained by observed variables such differences can be overcome by reweighting. However, participation may depend on unobserved variables. Even in such cases, under a common support assumption there exist weights which, if known, would allow reweighting the sample to match the population. While these weights cannot in general be estimated, we develop approximations which relate them to the role of private information in participation decisions. These approximations suggest benchmarks for assessing external validity.

1 Introduction

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

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

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. For both reasons, volunteers may have systematically different treatment effects than non-volunteers. In this case, the average treatment effect estimated by the experiment will differ from that in the population of workers as a whole. This issue - that the experimental sample differs from the population of policy interest - is widespread in economics and other fields.[1]

In this paper we consider external validity of treatment effects estimated from a randomized trial in a non-representative sample. For average treatment effects estimated from a randomized trial to be externally valid for a target population, they should coincide with the average treatment effect in the target population. We focus on the case where participation in the experiment is possibly non-random; we refer to this as the "participation decision," although note that it may be either a decision by an individual to enroll in a trial, or a decision by a researcher to include a particular area or unit in their experimental set.

If participation is driven entirely by observable variables, this problem has a well-known solution; one can reweight the sample to obtain population-appropriate estimates (as in e.g. Stuart et al, 2011). However, when participation depends on unobservable factors, including directly on the treatment effect, adjusting for differences in observable characteristics may be insufficient.In the context of instrumental variables estimation with heterogenous treatment effects, as in Imbens and Angrist (1994), Heckman et al (2006) refer to this possibility as "essential heterogeneity."[2]

Like Nyugen et al (2017), we observe that even when participation is driven by unobservable variables, under a common support assumption there exist weights which deliver the

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

[2] We use "participation" for the decision to take part in the randomized trial, rather than "selection," to distinguish the decision to join the trial from the treatment decision. We thank Peter Hull for suggesting this terminology.

average treatment effect – or any other moment of the data – for the population as a whole. While these weights are in general unknown, they provide an natural lens through which to consider external validity. In particular, the bias in the experimental estimate of the average treatment effect, as an estimate for the average treatment effect in the population as a whole, is given by the covariance of the individual-level treatment effect with these weights.

While the covariance of the treatment effect with the weights is a natural statistical object, its magnitude is difficult to directly interpret. In our main result, we therefore model the participation decision. We derive an approximation which relates external validity bias to the role of unobservables in the participation decision. This approximation uses an estimate of bias under the assumption that the participation decision is due entirely to observable variables; we refer to this bias as the "participation on observables" bias. We show that the ratio of the true bias to this participation on observables bias is inversely proportional to the fraction of the covariance between the variables driving participation and the treatment effect which is explained by the observables.

In the special case where participation is driven directly by the treatment effect, this implies that the ratio of total bias to participation on observables bias is inversely proportional to the R^2 from regressing individual-level treatment effects on the observables. Thus, in this setting our results highlight that the degree of private information about the treatment effect determines the extent of external validity bias.

Our approximations yield expressions which involve unobservables, and so cannot be estimated from the data. Moreover, the plausible importance of unobservables relative to observables will vary across applications, not least depending on the available covariates. Thus the point of our analysis is not to deliver definitive estimates or universal bounds, but instead to reframe the question of external validity in terms of the relative importance of observable and unobservable variables in a given context, since these are objects about which researchers may have plausible intuition or priors.

To provide some intuition for the procedure we suggest, consider again the Bloom et al (2015) example discussed above. Through the experimental data collection the authors observe some demographic characteristics of the overall population of workers which can be compared to the population who volunteer for the experiment. The first step in our procedure is to formally adjust the treatment effect estimate to reflect differences in these observable

features; we discuss a straightforward regression-based approach for this. The second step is to calculate bounds on the population average treatment effect by combining the adjustment for observable differences with an assumption about the relative importance of the observable and unobservable features in driving the participation decision. In this case, the adjustment would be informed by our sense of how much private information people have about their relative productivity at home.

Our results rely on approximations which are developed by considering cases where the bias is small. In a simulation example based on Muralidharan and Sundararaman (2011), however, we find that these approximations remain reliable in an example with nontrivial participation on unobservables.

To further illustrate, we apply our framework to four experimental papers: Attanasio, Kugler and Meghir (2011), Bloom et al (2015), Dupas and Robinson (2013), and Olken, Onishi, and Wong (2014). In each case we discuss the settings and describe how we might identify data from the relevant target population. We then discuss the robustness of the results to concerns about external validity.

Our results relate to a number of previous literatures, including those studying selection on observables (e.g. Hellerstein and Imbens, 1999; Hotz et al, 2005; Cole and Stuart, 2010; Stuart et al, 2011; Imai and Ratkovic, 2014; Dehejia, Pop-Eleches and Samii, 2015; Hartman, Grieve, Ramsahai and Sekhon, 2015), and propensity score reweighting (e.g. Hahn 1998 and Hirano, Imbens and Ridder 2003). Olsen et al (2013) derive expressions for the bias arising from participation on unobservables, while Alcott (2015), Bell et al (2016), and Chyn (2016) document such biases in applications.

Like us, Gechter (2015) and Nyugen et al (2017) provide methods for sensitivity analysis when we are concerned about participation on unobservables. While our bounds assume limits on the role of private information, Gechter (2015) assumes limits of the level of dependence between the individual outcomes in the treated and untreated states, and Nyugen et al (2017) assume limits on the mean of the unobservables. Hence, both approaches are complementary to ours. Bell and Stuart (2016) emphasize the importance of considering external validity in practice and discuss a variety of methods which may be used to evaluate threats to external validity, while Olsen and Orr (2016) discuss strategies which may be used in the design of an experiment to improve external validity.

Our analysis builds on the literature on structural models for policy evaluation, reviewed in Heckman and Vytlacil (2007a) and Heckman and Vytlacil (2007b). We follow this literature in using a threshold crossing latent-index model to represent the participation decision, and similar to Vytlacil (2002) we show that this model is not restrictive. While the primary focus of this literature has been on instrumental variables models, the form of participation we consider is allowed by the general frameworks of Heckman and Navarro (2007) and Heckman and Vytlacil (2007a).

A particularly active recent strand of the structural policy evaluation literature considers external validity in instrumental variable settings using the marginal treatment effect (MTE) approach (see, for example, Brinch et al (2016), Kline and Walters (2016), Kowalski (2016), and Mogstad et al (2017)). In contrast to these papers we are interested in applications where there is perfect compliance with the assigned treatment, but there is an *ex ante* participation decision. As a result there are no "always takers" in the settings we consider, which complicates direct application of the MTE approach as in e.g. Brinch et al (2016). On the other hand, in standard IV settings MTE approaches take advantage of information in the data we do not use and so may deliver sharper conclusions.

Our approach is likewise related to the literature on sensitivity analysis, reviewed by Rosenbaum (2002), that supposes that there is an unobserved variable which influences selection into treatment. While we consider the decision to participate in the experiment, rather than selection into treatment, it seems likely that the methods studied in this literature could be extended to our setting. Related approaches include Altonji Elder and Taber (2005) and Oster (forthcoming). Finally we also relate, more distantly, to the recent literature on external validity in regression discontinuity settings (Bertanha and Imbens, 2014; Angrist and Rokkanen, 2015; Rokkanen, 2015). This link, along with the connection to methods for instrumental variables settings, is discussed further in Section 7.

In the next section, we briefly discuss the sorts of the external validity problems we aim to address. Following that, we introduce the setting we consider and give sufficient conditions for the existence of weights which recover the average treatment effect in the target population. Section 4 develops our main approximation result. Section 5 discusses implementation of our approach and illustrates in a constructed example based on data from Muralidharan and Sundararaman (2011). Section 6 details our applications, while Section 7 discusses extensions

of our results and Section 8 concludes.

2 Scope of Problem

Before introducing our framework, it is useful to be explicit about the range of problems to which we hope to speak. To fix ideas, we first discuss two broad types of studies where we expect our approach to be useful. We then briefly discuss the kinds of external validity issues our approach is well-suited to address in the settings.

2.1 Settings

Experiments with a Participation Margin In many experimental settings there is an explicit choice by participants to take part in the study. Quite often the treatment offered is only available through the study. Examples include Attanasio, Kugler and Meghir (2011) and Gelber, Ibsen and Kessler (2016) on job training, Bloom et al (2015) on working from home, and Muraldiharan et al (2017) on computer-based tutoring. A key feature of these experiments is that a broad population is offered the chance to be in the experiment, and only those who volunteer are included in the randomization set. The estimates derived from such experiments are therefore valid for the sample who choose to take part, but may not be valid for the overall population. This is of particular concern if, for example, we think people are more likely to participate when they expect a large treatment effect.

Experiments with Selected Locations or Units A second group of experiments are those in which researchers select a set of areas or treatment units (schools, villages, etc) in which to run their experiment, and the locations are selected non-randomly. Examples include Muraldiharan and Sundararaman (2011) on teacher performance pay, Jensen (2012) on education and job opportunities, Olken et al (2014) on block grants, and Alcott (2015) on Opower. In contrast to the above, in these settings there is no individual participation margin, and locations typically do not select themselves into the study. Nevertheless, the selection of locations is often non-random in ways that may influence the results. As with individual participation decisions, this concern is particularly acute when we think researchers select units based in part on their predictions for the treatment effect.

2.2 Types of External Validity

Having run an experiment of the sort described above, there are many external validity questions one could ask. We may wonder about extrapolation to a random sample, or to the full population, or more broadly to other locations or time periods. We will briefly discuss the role of our approach in addressing each of these extrapolation problems.

Extrapolation to Random Sample Our approach is most directly applicable if we want to extrapolate from an experimental sample to a similarly sized random sample. This may be relevant if, for example, one planned a policy where a treatment would be offered to employees randomly rather than allowing them to select into it.

Extrapolation to Full Population In many settings our approach is also suited to considering extrapolation to a full population. This is a common type of external validity concern in practice. For example, one might have evaluated a policy in a subset of locations in a state or country and now want to extend to the whole area. Or one might have evaluated the policy using individuals who volunteered for an experiment, and now want to extend it to all individuals.

A complication is that treating the entire population could introduce important general equilibrium or spillover effects. Where such issues arise it may well be interesting to undertake the analysis we suggest, but to accurately predict the effect of treating the full population one will need to separately account for effects arising from the scale of treatment.

Extrapolation to Additional Locations, Circumstances Perhaps the most ambitious external validity goals relate to extrapolation to different time periods or locations - for example, to times with better labor market conditions or to different states. This case is beyond the scope of our approach, since we fundamentally rely on the assumption that the trial population is a subset of the target population. Bates and Glennerster (2017) provide a nuanced discussion of the extent to which one can port the results of randomized trials between locations within developing countries, while Gechter (2015) develops formal extrapolation bounds under assumptions on the relationship between treated and untreated outcomes.

As this discussion suggests, we view our approach as best suited to asking whether the

results in a particular experiment are likely to extend to an overall population. As we illustrate in our examples, this is often a question of policy interest. We turn now to developing our theoretical framework, after which we return to implementation and applied examples.

3 Participation Decisions and Reweighting

We assume that we observe a sample of observations i from a randomized trial in some population, and denote the distribution in this trial population by P_S. We are interested in the average treatment effect in a larger target population, whose distribution we denote by P. In this section, we show that under mild conditions we can reweight the trial population to match the target population. For simplicity we assume an infinite sample in developing our theoretical results, so the distribution of observables under P_S is known. Results on inference, which account for sampling uncertainty, are developed in Section 5.1 below.

We consider a binary treatment, with $D_i \in \{0, 1\}$ a dummy equal to one when i is treated. We write the outcomes of i in the untreated and treated states as $Y_i(0)$, $Y_i(1)$, respectively. We assume that we also observe a vector of covariates C_i for each individual which are unaffected by treatment. Let $X_i = (Y_i(0), Y_i(1), C_i)$ collect the potential outcomes and covariates. We observe each unit in only a single treatment state, so 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 (Y_i, D_i, C_i). See Imbens and Rubin (2015) for further discussion of the potential outcomes framework.

We assume that treatment is randomly assigned in the trial population.

Assumption 1 *Under P_S, $D_i \perp (Y_i(0), Y_i(1), C_i)$ and $E[D_i] = d$ for known d.*

Under Assumption 1, we can express the average treatment effect (ATE) in the trial population as

$$E_{P_S}[TE_i] = E_{P_S}[Y_i(1) - Y_i(0)] = E_{P_S}[T_i], \; T_i = \frac{D_i}{d} Y_i - \frac{1 - D_i}{1 - d} Y_i. \tag{1}$$

Since T_i can be calculated from the data, this confirms that we can estimate the trial-

population ATE from our experimental sample, as is well-understood.[3]

Our ultimate interest is not in the trial population ATE $E_{P_S}[TE_i]$, but instead in the target population ATE $E_P[TE_i]$.[4] We assume that the trial population is a subset of the target population, and that there is a variable S_i in the target population that 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 of X_i under P, P_X , has density $p_X(x)$ then we can write the density in the trial population in terms of $p_X(x)$ and S_i.[5]

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*

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

If we assume that S_i is independent of X_i, this result implies that $P_{X,S} = P_X$ and thus that the distribution of X_i is the same in the trial and target populations. If, on the other hand, S_i is not independent of X_i, then we will have $E_{P_S}[f(X_i)] \neq E_P[f(X_i)]$ for some functions $f(\cdot)$. In particular, if participation is related to the treatment effect $Y_i(1) - Y_i(0)$, in general $E_{P_S}[TE_i] \neq E_P[TE_i]$ so the ATE in the trial population is biased as an estimate of the ATE in the target population.

To draw conclusions about the target population based on data in the trial population, it is important that all values of X which are present in the target population also have some chance of appearing in the trial population.

[3]While we focus on 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. This follows from well-known results in the literature on propensity score reweighting- see Rosenbaum and Rubin (1983).

[4]Our analysis extends directly to other common target parameters in treatment effect settings. For example, to calculate average treatment effects for subgroups defined based on covariates we can limit the trial and target populations to these groups, while to calculate average treatment effects on the distribution function of an outcome variable $\tilde{Y}$ at some y we can define $Y_i(0) = 1\left\{\tilde{Y}_i(0) \leq y\right\}$ and $Y_i(1) = 1\left\{\tilde{Y}_i(1) \leq y\right\}$.

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

Assumption 2 $0 < E_P[S_i|X_i]$ *for all* X_i.

This assumption ensures that the distributions in the trial and target populations are mutually absolutely continuous. If this assumption fails, so there are some values of X_i in the target population which are never observed in the trial population, 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.

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

The NLSY oversampled certain groups (e.g. African-Americans). Due to this sampling scheme, moments of these data (for example, means of variables) will not be unbiased for those moments in the full population. We use the NLSY as illustration since in these data we observe sampling weights. These weights are intended to allow researchers to reweight the NLSY to obtain a representative sample from the US population.[6] Since the NLSY is not a randomized trial, we take the moments of interest to be the means of demographic variables in the data. We use the 1984 survey data for this illustration.

In our terminology, we define the NLSY sample as our trial population, and the reweighted representative sample as our target population. The availability of the weights makes it possible to explicitly illustrate the reweighting calculations we develop below. $\triangle$

3.1 Reweighting Algebra

When the trial and target populations differ, Assumption 2 implies that (in principle) we can reweight the trial population to match the target population.

[6]The NLSY provides a number of different weights; we use the overall sampling weights, not the cross sectional weights. One could raise concerns that the NLSY weights may not fully accurately match the overall population. Given that our interest here is solely in illustration, however, we abstract away from these concerns.

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

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

Lemma 2 is a well-known result (see e.g. Horvitz and Thompson, 1952), and shows that we can recover expectations under P_X by reweighting our observations from $P_{X,S}$ using the weights $W_i = \frac{p_X(X_i)}{p_{X,S}(X_i)}$, which compare the densities of the trial and target populations at each X_i. Since Lemma 2 implies that $E_P[TE_i] = E_{P_S}[W_i TE_i]$, if we observed the weights W_i we could estimate the ATE in the target population as $E_{P_S}[W_i T_i]$. The weights depend on both the distribution in the target population and the potential outcomes, however, and so cannot be calculated in practice.

While unknown, the weights W_i provide a useful lens through which to consider external validity. 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 $E_P[f(X_i)]$. A closely related result was previously derived in Olsen et al (2013).

Corollary 1 *For any function* $f(\cdot)$, *under Assumption 2 we have*

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

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

$$E_{P_S}[f(X_i)] - E_P[f(X_i)] = -\sigma_{P_S}(W_i)\,\rho_{P_S}(W_i, f(X_i))\,\sigma_{P_S}(f(X_i)), \tag{2}$$

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

The final term in equation (2), $\sigma_{P_S}(f(X_i))$, measures the standard deviation of $f(X_i)$ in the trial population. 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 the participation decision loads on $f(X_i)$. By definition this quantity is smaller than one in absolute value. Lastly, the standard deviation $\sigma_{P_S}(W_i)$ can be viewed as measuring how

much the trial and target populations differ along any dimension of X_i, since the bounds on $\rho_{P_S}(W_i, f(X_i))$ imply that for all functions $f(\cdot)$,

$$|E_P[f(X_i)] - E_{P_S}[f(X_i)]| \leq \sigma_{P_S}(W_i)\,\sigma_{P_S}(f(X_i)). \tag{3}$$

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

$$\begin{aligned}
E_{P_S}[f_1(X_i)] - E_P[f_1(X_i)] &= -\sigma_{P_S}(W_i)\,\rho_{P_S}(W_i, f_1(X_i))\,\sigma_{P_S}(f_1(X_i)) \\
E_{P_S}[f_2(X_i)] - E_P[f_2(X_i)] &= -\sigma_{P_S}(W_i)\,\rho_{P_S}(W_i, f_2(X_i))\,\sigma_{P_S}(f_2(X_i)) \\
&\vdots \\
E_{P_S}[f_q(X_i)] - E_P[f_q(X_i)] &= -\sigma_{P_S}(W_i)\,\rho_{P_S}(W_i, f_q(X_i))\,\sigma_{P_S}(f_q(X_i)).
\end{aligned} \tag{4}$$

A key feature of this decomposition is that the standard deviation of the weights, $\sigma_{P_S}(W_i)$ appears in all rows. This again reflects the fact that $\sigma_{P_S}(W_i)$ measures the degree to which the trial and target populations differ along any dimension of X_i.

Illustration (continued): In the NLSY we observe the weights W_i. Therefore, we can calculate all terms in the decomposition (4). In particular, we consider this decomposition when taking $f(X_i)$ to measure race (share white), high school completion, and gender (share male).

The first two columns of Table 1 report the trial and target population means for these variables in the NLSY. The final three columns show the elements of the bias decomposition in equation (2). The difference in means for each variable is the product of these three elements. The differences between trial and target population means reflect the sampling structure: the NLSY over-samples racial minority groups and individuals from lower socioeconomic status backgrounds. There are fewer whites and fewer high school graduates in the sample than in the overall US population. The bias is largest for race, which reflects the very high correlation between the sampling weights and race.

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

While the decomposition (2) does not deliver precise conclusions about the degree of bias without further restrictions, it provides a useful guide to intuition. In particular, the bias in the trial population ATE is larger when (a) the trial and target populations differ more in general, (b) participation decisions are more correlated with the treatment effect, and (c) there is more variability in the treatment effect. In the next section we build on these intuitions, introducing unobserved variables that drive participation decisions and developing approximations which allow us to translate beliefs about the role of private information in participation decisions to bounds on the plausible degree of external validity bias.

4 Bias Approximations

While the the results of the last section are helpful for developing intuition, to use these results to draw quantitative conclusions about the bias of the sample average of T_i as an estimator for $E_P[TE_i]$ (i.e $E_{P_S}[TE_i] - E_P[TE_i]$) we need to know certain properties of the weights W_i. These weights, in turn, depend on distribution of X_i in the target population and so are unknown in applications. In this section we develop approximate expressions for the weights which allow us to express the bias $E_{P_S}[TE_i] - E_P[TE_i]$ in the trial population ATE in terms of (a) the bias due to participation on observables and (b) the role of private information in participation decisions. Bias from participation on unobservables can be estimated, so if researchers have a view about the degree of private information in a given setting, they can use these results to estimate the plausible degree of bias. Our approximations become exact when the participation decision is nearly random, in a sense we make precise below, but we find in the next section that they perform well in simulated examples where participation decisions depend non-trivially on unobservables.

4.1 Participation Model

Throughout this section we assume that in addition to $X_i = (Y_i(0), Y_i(1), C_i)$, there are also variables U_i which are unobserved by the researcher but may play a role in the participation decision. Further, we assume that the distribution of the covariates C_i in the target population is known (though we discuss in Section 5.1 below how we can proceed if we know only some aspects of this distribution). If there are variables which are observed in the trial population but whose distribution in the target population is entirely unknown, for the purposes of analysis we include these in U_i. See Nyugen et al (2017) for alternative approaches to using variables observed in the trial population but not the target population.

Since we can take U_i to include $(Y_i(0), Y_i(1))$, it is without loss of generality to assume that participation depends only on (C_i, U_i), so S_i is conditionally independent of the potential outcomes

$$S_i \perp (Y_i(0), Y_i(1)) \,|\, (C_i, U_i). \tag{5}$$

Analogous to Assumption 2, we further assume that every value of (C_i, U_i) may appear in the trial population, where we now strengthen this assumption by bounding the conditional probability of participation away from zero and one.

Assumption 3 *Equation (5) holds, and there exists $\nu > 0$ such that $\nu < E_P[S_i|C_i = c, U_i = u] < 1 - \nu$ for all (c, u).*

Assumption 3 implies Assumption 2. To develop our approximation results, it will be helpful to model participation using a threshold-crossing latent index model. Under Assumption 3, this is without loss of generality.

Lemma 3 *Under Assumption 3 we can write*

$$S_i = 1\{g(C_i, U_i) \geq V_i\}, \tag{6}$$

where V_i is continuously distributed with Lipschitz density p_V independently of $(C_i, U_i, Y_i(0), Y_i(1))$ and has support equal to $\mathbb{R}$, and $g(C_i, U_i)$ is bounded.

This result shows that since we do not restrict the functional form of $g(C_i, U_i)$, it is

without loss of generality to assume a latent index model.[7] Thus, similar to Vytlacil (2002), a sufficiently flexible latent index model imposes no restrictions beyond those implied by Assumption 3. See also Vytlacil (2006). The independence condition (5) then implies that there exist weights W_i that depend only on $g(C_i, U_i)$ and allow us to calculate the target-population mean of any function $f(X_i)$.

Lemma 4 *Under Assumption 3, there exist weights*

$$W_i = \frac{p(C_i, U_i)}{p_S(C_i, U_i)} = w(g(C_i, U_i))$$

for a continuously differentiable function w, such that for any function $f(\cdot)$

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

Lemma 4 builds on Lemma 2 by showing that under Assumption 3, we can take the weights W_i to depend only on the function $g(C_i, U_i)$ in the latent index model. This result directly captures the bias, but the function $w(\cdot)$ is unknown and depends on the data generating process, so this result is still not directly useable in practice.

4.2 Approximate Weights

To obtain more easily interpretable expressions, we consider Taylor approximations to W_i. In particular, since $w(\cdot)$ is a smooth function, Taylor expansion around $E_{P_S}[g(C_i, U_i)]$ suggests[8]

$$W_i \approx W_i^* = w_0 + w_1 g(C_i, U_i).$$

With this approximation, as in Corollary 1 we have

$$E_{P_S}[f(X_i)] - E_P[f(X_i)] \approx -Cov_{P_S}(W_i^*, f(X_i)) = -w_1 Cov_{P_S}(g(C_i, U_i), f(X_i)). \quad (7)$$

[7] In fact, the proof shows it is without loss to assume that the error V_i is normal as in e.g. Heckman (1979). Hence, unlike in the literature on empirical implications of the Roy model, e.g. Heckman and Honore (1990), normality of V_i imposes no restrictions in our setting beyond those implied by Assumption 3.

[8] First-order Taylor approximation yields $w_1 = w'(E_{P_S}[g(C_i, U_i)])$, $w_0 = w(E_{P_S}[g(C_i, U_i)]) - w_1 E_{P_S}[g(C_i, U_i)]$.

Thus, the bias in the trial-population mean of $f(X_i)$ is proportional to the covariance between the participation variable $g(C_i, U_i)$ and $f(X_i)$.

Our next result shows that this approximation becomes arbitrarily accurate when participation is nearly random. To formalize this idea, we make the following assumption.

Assumption 4 *We can write*

$$g(C_i, U_i) = \kappa \cdot h(C_i, U_i)$$

for a constant κ.

For a fixed data generating process, Assumption 4 is without loss of generality, since we can always take $\kappa = 1$ and $h(C_i, U_i) = g(C_i, U_i)$. To develop our approximations, however, we will hold the distribution of (X_i, U_i) in the target population, as well as the function h, fixed and take κ to zero, so participation becomes entirely random in the limit. Assumption 4 thus specifies precisely how the participation decision approaches the fully random case.

Proposition 1 *Under Assumptions 3 and 4, for any* $f(X_i)$ *such that* $E_P\left[f(X_i)^2\right]$ *is finite*

$$Cov_{P_S}(W_i, f(X_i)) = Cov_{P_S}(W_i^*, f(X_i)) + O\left(\kappa^2\right),$$

as $\kappa \to 0$. *In particular, if* $Cov_P(g(C_i, U_i), f(X_i)) \neq 0$ *then*

$$\frac{Cov_{P_S}(W_i, f(X_i))}{Cov_{P_S}(W_i^*, f(X_i))} \to 1.$$

This proposition shows that when participation is nearly random, the approximation error in (7) vanishes. In particular, while the overall degree of bias $Cov_{P_S}(W_i, f(X_i))$ tends to zero as $\kappa \to 0$, the approximation error vanishes even faster.

If we apply our bias approximation (7) with $f(X_i) = TE_i$ equal to the treatment effect,

$$E_{P_S}[TE_i] - E_P[TE_i] \approx -w_1 Cov_{P_S}(g(C_i, U_i), TE_i), \tag{8}$$

so the bias in the trial population ATE is approximately proportional to the covariance of the treatment effect with $g(C_i, U_i)$. Even if this covariance were known, however, this formula is

not usable due to the unknown constant of proportionality w_1. To eliminate this term, we compare the true bias to the bias from participation on observables.

Corrections for participation on observables, including Hotz et al (2005) and Stuart et al (2011), adjust the average treatment effect for differences between the trial and target populations in the distribution of the covariates C_i. Given enough data and a sufficiently flexible specification, these methods consistently estimate $E_P\left[E_{P_S}\left[TE_i|C_i\right]\right]$, which can be interpreted as the ATE corrected for participation on observables. In particular, this parameter begins with the average treatment effect conditional on each value of the covariates in the trial population, $E_{P_S}\left[TE_i|C_i\right]$, and then averages over these conditional effects using the distribution of covariates in the target population. Since we have assumed the distribution of covariates in the target population is known and that we observe data from a randomized experiment in the trial population, this is a quantity we can estimate.

If we apply our bias approximation (7) with $f\left(X_i\right) = E_{P_S}\left[TE_i|C_i\right]$, then since $E_{P_S}\left[E_{P_S}\left[TE_i|C_i\right]\right] = E_{P_S}\left[TE_i\right]$ by the law of iterated expectations, we see that

$$E_{P_S}\left[TE_i\right] - E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] \approx -w_1 Cov_P\left(g\left(C_i, U_i\right), E_{P_S}\left[TE_i|C_i\right]\right) \tag{9}$$

$$= -w_1 Cov_P\left(E_{P_S}\left[g\left(C_i, U_i\right)|C_i\right], E_{P_S}\left[TE_i|C_i\right]\right)$$

where the equality follows from another application of the law of iterated expectations. Thus, the correction for participation on observables (9) takes the same form as the true bias (8), except that we consider the covariance of $g\left(C_i, U_i\right)$ (or $E_{P_S}\left[g\left(C_i, U_i\right)|C_i\right]$) with the conditional average treatment effect $E_{P_S}\left[TE_i|C_i\right]$ rather than the true treatment effect TE_i. If we take the ratio of the true bias to the participation on observables bias, we see that

$$\frac{E_P\left[TE_i\right] - E_{P_S}\left[TE_i\right]}{E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]} \approx \frac{Cov_{P_S}\left(g\left(C_i, U_i\right), TE_i\right)}{Cov_{P_S}\left(E_{P_S}\left[g\left(C_i, U_i\right)|C_i\right], E_{P_S}\left[TE_i|C_i\right]\right)} = \Phi,$$

where the constant of proportionality w_1 drops out and we are left with a ratio of covariances.

While the ratio Φ may initially appear rather involved, it has an intuitive interpretation. In particular, Φ^{-1} is the share of the covariance between $g\left(C_i, U_i\right)$ and TE_i which is explained

by covariates. To further draw this out, note that we can write

$$\Phi = 1 + \frac{Cov_{P_S}\left(g\left(C_i, U_i\right) - E_{P_S}\left[g\left(C_i, U_i\right)|C_i\right], TE_i - E_{P_S}\left[TE_i|C_i\right]\right)}{Cov_{P_S}\left(E_{P_S}\left[g\left(C_i, U_i\right)|C_i\right], E_{P_S}\left[TE_i|C_i\right]\right)}. \tag{10}$$

The numerator in the second term is the covariance between the residuals from nonparametrically regressing $g\left(C_i, U_i\right)$ and $E_{P_S}\left[TE_i\right]$ on the covariates C_i, while the denominator is the covariance between the fitted values.

If the covariates fully explain any relationship between participation decisions and treatment effects, the numerator in (10) is zero and $\Phi = 1$. For this to be the case it suffices either that the unobservables U_i are uninformative about participation, so $g\left(C_i, U_i\right) = E\left[g\left(C_i, U_i\right)|C_i\right]$, or that they are uninformative about the treatment effect, so $E_{P_S}\left[TE_i|C_i, U_i\right] = E_{P_S}\left[TE_i|C_i\right]$. In either case, correcting for participation on observables is enough to recover the ATE in the target population, and $E_P\left[TE_i\right] = E_P\left[E_{P_S}\left[TE_i|C_i\right]\right]$.

By contrast, if private information plays a role in both participation decisions and the treatment effect, there is scope for bias. If the covariance between the residuals is of the same sign as that between the fitted values, we will have $\Phi \geq 1$. This need not be the case in general, however, and the plausible range of values of Φ will vary across different settings depending on the set of available covariates C_i, the plausible set of unobservables U_i, and our beliefs about how participation is determined.

Special Case: Participation on the Treatment Effect An important special case arises when we assume that participation decisions are based directly on the predicted treatment effect given C_i and U_i, potentially along with some set of unrelated variables.

Assumption 5 *Suppose that*

$$g\left(C_i, U_i\right) = \alpha_1 E_P\left[TE_i|C_i, U_i\right] + g_C\left(C_i\right) + g_U\left(U_i\right),$$

where $E_{P_S}\left[g_U\left(U_i\right)|C_i\right] = 0$ *and*

$$Cov_{P_S}\left(TE_i, g_C\left(C_i\right)\right) = Cov_{P_S}\left(TE_i, g_U\left(U_i\right)\right) = 0.$$

Intuitively, we can think of this as the case where participation depends on the predicted

treatment effect $E_{P_S}[TE_i|C_i, U_i]$, a function of covariates which is unrelated to the treatment effect and, potentially, a function of unobservables which is unrelated to both the treatment effect and the covariates. Note that if we include the potential outcomes $(Y_i(0), Y_i(1))$ among the unobservables, $E_{P_S}[TE_i|C_i, U_i] = TE_i$, so Assumption 5 allows the possibility of direct participation on the treatment effect.

Under Assumption 5, we obtain a simplified expression for Φ.

Lemma 5 *Under Assumption 5,*

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

Thus, under Assumption 5, Φ measures the variance of treatment effects predicted based on (C_i, U_i), relative to the variance of treatment effects predicted based on C_i alone. This can be interpreted as a measure for the degree of private information about the treatment effect used in participation decisions. When there is a large amount of private information the true bias in the ATE will be much larger than the participation on observables bias $E_{P_S}[TE_i] - E_P[E_{P_S}[TE_i|C_i]]$. By contrast, when there is little private information the target-population ATE will be close to $E_P[E_{P_S}[TE_i|C_i]]$. In the extreme case where participation is directly on the treatment effect and $E_{P_S}[TE_i|C_i, U_i] = TE_i$, Φ^{-1} measures the share of treatment effect heterogeneity captured by the covariates, specifically the R^2 from nonparametrically regressing TE_i on C_i.

To further explore the structure of Φ in this case, analogous to (10) we can write

$$\Phi = 1 + \frac{Var_{P_S}\left(E_{P_S}[TE_i|C_i, U_i] - E_{P_S}[TE_i|C_i]\right)}{Var_{P_S}\left(E_{P_S}[TE_i|C_i]\right)}.$$

Unlike in (10), however, the second term involves a ratio of variances and so is always positive. Thus, under Assumption 5 we know that $\Phi \geq 1$ so that, up to approximation error,

$$\frac{E_P[TE_i] - E_{P_S}[TE_i]}{E_P[E_{P_S}[TE_i|C_i]] - E_{P_S}[TE_i]} \geq 1.$$

Thus, if correcting for participation on observables reduces our estimate of the target-population average treatment effect, allowing for additional participation on unobservables implies still-

further reductions, and likewise for increases. We thus see that, unlike the general case, the assumption of participation on the treatment effect delivers unambiguous predictions about the sign of the bias from participation on unobservables.[9]

5 Implementation and Example

This section discusses implementation of our approach and examines performance in a constructed example.

5.1 Implementation

Our assumption of random assignment, Assumption 1, ensures that we can estimate the ATE $E_{P_S}[TE_i]$ in the trial population. We are interested in the range of plausible values for the ATE in the target population. To assess the plausible of bias from participation on unobservables, we consider two distinct approaches. First, we can consider a particular value for the ATE in the target population, t_P^*, and ask how important a role private information would have to play in participation decisions to obtain this value. Alternatively, we can assume limits the degree of private information and calculate the implied range of ATEs.

For both approaches, an important first step is correction for participation on observables. Recall that the ATE corrected for participation on observables is $E_P[E_{P_S}[TE_i|C_i]]$. Provided the distribution of C_i in the target population is known, to calculate this correction we only need an estimate of $E_{P_S}[TE_i|C_i]$. In this section we discuss a simple approach based on linear regression of the treatment effect proxy T_i on functions of covariates, which can be applied even with limited knowledge of the features of the target population. Similar linear corrections are considered in Hotz et al (2005) to account for residual differences after matching. More broadly, linear corrections for differences in the distribution of covariates between groups have been widely studied - see for example Kline (2011) and references therein. In settings with richer data on the target population, one can also apply our general approach together with

[9]Note, further, that since the resulting expressions to not depend on $g(C_i, U_i)$, we can bound Φ using properties of the treatment effect. For example, in the special case where we observe U_i in the trial population but its distribution in the target population is unknown, we can estimate $Var_{P_S}(E_{P_S}[TE_i|C_i, U_i])$ directly. Even when U_i is unobserved in the trial population, we know that $Var_{P_S}(E_{P_S}[TE_i|C_i, U_i]) \leq Var_{P_S}(TE_i)$, and so can use known bounds on $Var_{P_S}(TE_i)$, perhaps under assumptions on the degree of dependence between $Y_i(0)$ and $Y_i(1)$ as in Gechter (2015), to bound Φ.

more sophisticated corrections for participation on observables, including matching as in Hotz et al (2005) and propensity-score reweighting as in Stuart et al (2011).

Given the participation on observables adjustment, we implement our two approaches to robustness. First, taking the target value of interest as t_P^* we evaluate robustness by calculating the value of Φ sufficient to produce t_P^*. A natural target value in many treatment effect settings is zero. Let us denote the resulting Φ by $\Phi\left(t_P^*\right)$. To calculate $\Phi\left(t_P^*\right)$, we simply compare the implied total adjustment to the participation on observables adjustment,

$$\Phi\left(t_P^*\right) = \frac{t_P^* - E_{P_S}\left[TE_i\right]}{E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]}.$$

Alternatively, also using the participation on observables adjustment, we can impose bounds $\Phi \in [\Phi_L, \Phi_U]$. Under these bounds we know that (assuming the participation on observables correction is positive),[10] $E_P\left[TE_i\right]$ lies in the interval

$$E_{P_S}\left[TE_i\right] + \left[\Phi_L\left(E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]\right), \Phi_U\left(E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]\right)\right]. \quad (11)$$

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

The plausible range of values for Φ will depend on the context. Consider first the case where we think that participation decisions are driven by the treatment effect along with variables unrelated to treatment, in the sense of Assumption 5. In this case Φ can be interpreted as a measure of the degree of private information about treatment effects used in participation decisions. For example, suppose we find that a value of $\Phi = 2$ is necessary to eliminate a positive result. This indicates that the unobservables would have to be at least as informative about the treatment effect as the observables in order for the effect in the target population to be zero.[11] If we have a rich set of covariates which seems likely to be informative about treatment effect heterogeneity relative to the plausible sources of private information, we can

[10] If it is instead negative, then $E_P\left[TE_i\right]$ lies in the interval

$$E_{P_S}\left[TE_i\right] + \left[\Phi_U\left(E_{P_S}\left[TE_i|C_i\right] - E_{P_S}\left[TE_i\right]\right), \Phi_L\left(E_{P_S}\left[TE_i|C_i\right] - E_{P_S}\left[TE_i\right]\right)\right].$$

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

then conclude that our result is robust to plausible levels of private information. If, on the other hand, the available covariates seem unlikely to be very informative about the treatment effect, then a value of $\Phi = 2$ may be quite plausible, suggesting that our result is not robust.

The interpretation is similar under more general assumptions about participation decisions. For example, consider a case where experimental locations were chosen on accessibility, but we do not observe accessibility measures. In this case Φ measures the relative importance of the observables and unobservables in explaining the covariance between accessibility and the treatment effect. If we think that our covariates suffice to construct a good proxy for accessibility, or that they explain much more treatment effect heterogeneity than do the plausible unobservables, then finding that a value of $\Phi = 2$ is needed to overturn our results should reassure us about their robustness. On the other hand, if neither of these conditions hold we may remain concerned about external validity even if a large value of Φ is needed to overturn our results.

Further intuition may be provided by thinking about the share of relevant variables missed by our observed covariates. In Appendix B.1, we describe a model where a large number of latent factors drive treatment effects and participation decisions, and a random subset of these are measured by C_i while the rest are measured by U_i. In this setting Φ can be interpreted as the ratio of the total to the observed factors. In particular, $\Phi = 2$ reflects a case where the observed covariates capture 50% of the latent factors.[12]

5.1.1 Correction for Participation on Observables

To implement the approaches discussed above, we need an estimate of $E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]$. As noted in Section 3 we can estimate $E_{P_S}\left[TE_i\right]$ by the sample average of T_i as defined in (1), so the challenge is estimation of $E_P\left[E_{P_S}\left[TE_i|C_i\right]\right]$, the ATE corrected for participation on observables.

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

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

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

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

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

then this procedure exactly recovers $E_P[E_{P_S}[TE_i|C_i]]$. If on the other hand we consider the linear specification as an approximation, then this procedure delivers an approximation to $E_P[E_{P_S}[TE_i|C_i]]$, where the approximation error will vanish as we consider rich sets of functions $r(C_i)$.[13]

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

5.1.2 Inference

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

Under the assumption of simple random assignment, we can conduct inference using the bootstrap.[14] Bootstrap standard errors for $\Phi(t_P^*)$ become unreliable when the correction

[13] Ideally we would include interactions and higher-order terms in $r(C_i)$, although this may be infeasible given data constraints. Nonetheless, whenever possible researchers should at a minimum include linear and squared terms in the covariates, since this will capture differences between the trial and target populations in the means and variances of these variables. In settings with richer data one should consider even more moments - interactions between the variables, higher moments of the distribution of each variable, etc.

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

for participation on observables is close to zero, however. In this case, the denominator in $\Phi\left(t_P^*\right)$ is almost zero, which results in problems very similar to those that arise from weak instruments.[15] In Appendix B.2 we discuss how to construct reliable confidence sets for $\Phi\left(t_P^*\right)$. These confidence sets are close to the usual ones when the participation on observables correction is large, but can be unbounded when it is small.

Confidence sets for the ATE $E_P\left[TE_i\right]$ are more straightforward. In particular, for $(\hat{\sigma}_L, \hat{\sigma}_U)$ bootstrap standard errors for our estimates $(\hat{\gamma}_L, \hat{\gamma}_U)$ of the lower and upper bounds in (11), we can construct a (conservative) level $1-\alpha$ confidence interval for $E_P\left[TE_i\right]$ as

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

for c_α the two-sided level $1-\alpha$ normal critical value (e.g. 1.96 for a 95% confidence set).[16] Alternatively, one can report $(\hat{\gamma}_L, \hat{\gamma}_U, \hat{\sigma}_L, \hat{\sigma}_U)$, which allows readers to construct the confidence set of their choice.

5.2 Example

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

5.2.1 Data and Empirical Approach

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

[15] The participation on observables correction $E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]$ plays the same role as the first-stage parameter in linear IV, so problems arise when this difference is close to zero relative to sampling variability.

[16] In fact, one can typically use a critical value smaller than c_α, though more computation is required to derive the exact value. We do this in our applications - see Appendix B.2 for details.

To construct our example, we define the distribution of the target population to be the empirical distribution in the Muralidharan and Sundararaman (2011) data. That is: although it would also be reasonable to ask about the external validity of their results to a broader population, for this exercise we assume that they *are* studying the target population and we ask what conclusions we would draw if we observed only a selected subset. To abstract from issues of sampling variability, we collapse the data to the school level and sample from the data with replacement to create a large target population.

We predict treatment effects based on school-level characteristics: average teacher education, average teacher training, average teacher salary, average household income, a school infrastructure index, the share of the student population who is scheduled tribe or scheduled caste, and average teacher absence. We also include dummies for which mandal (a geographic area) the school is in. We can think of these controls as capturing differences across areas in how effective the program is.

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

5.2.2 Participation on Treatment Effect

We first model participation on the predicted treatment effect. We create a predicted treatment effect $E_P\left[TE_i|C_i, U_i\right]$ by defining T_i as in (1) above and regressing T_i on the full set of controls for school-level characteristics. We include schools in the trial population if $E_P\left[TE_i|C_i, U_i\right] \geq V_i$ where V_i is normally distributed with the same mean as $E_P\left[TE_i|C_i, U_i\right]$ and a standard deviation three times as large.[17]

The ATE in the target population is 0.074.[18] The ATE in the trial population is considerably larger, 0.15. If we assume that we observe all the characteristics used in the participation

[17]The mean of V_i ensures that roughly half of the target population will be in the trial population.

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

decision, the adjustment for participation on observables delivers the correct value 0.074 for the target population ATE.

We next consider the case where we cannot observe some of the variables used in the participation decision. We vary the size of the subset which is unobserved, considering what happens when we eliminate just one variable, then 10%, 20%, 30%, and 50% of the variables (chosen at random).[19] In each case we calculate the ATE correcting for participation on observables, where there is now also participation on unobservables. We consider all possible single-variable eliminations, while for the other cases we take 200 draws at random.

The first column of Panel A of Table 2 shows the average participation on observables adjusted ATE for each exclusion set. When only one covariate is treated as unobserved (the last row in Panel A) the estimate is extremely close to the target population ATE, since the unobservables are quite limited. As we treat larger sets as unobserved the estimate is further from the target population ATE and closer to the trial population estimate.

The second column of Panel A in this table reports the average value of Φ to match the target population treatment effect for each specification. This value is largest when the largest share of covariates is excluded. It is worth noting that the average values of Φ are quite close to the actual ratio of the number of total covariates to the observed covariates, reflecting the results for cases with many covariates developed in Appendix B.1.

We can visualize the range of values Φ which generate the target population ATE, given each set of unobservables. This is done in Figure 1. As we exclude a larger set of variables, the range of Φ goes up, consistent with the presence of more private information in the participation decision. These values of Φ correspond directly to the relative importance of the observed versus unobserved covariates in predicting the treatment effect.

To see this more directly, we calculate the ratio of the R-squared from regressing T_i on the observed covariates to the R-squared from regressing on all the school characteristics. This R-squared is equivalent to the variance ratio in Lemma 5, which approximates the true bias by Proposition 1. We graph this against the value of Φ to match the true bias. Deviations from equality arise from approximation error. Figure 2 suggests such error here is limited.

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

5.2.3 Participation on Other Unobsrevables

We next model a case where participation is driven by features of the data other than the treatment effect. In particular, we imagine that we select areas based on mandal-level teacher training. We divide the sample into quartiles based on the mandal-level average of teacher training, and then calculate the average treatment effect within each quartile, which we use as our $g(C_i, U_i)$. In practice, this puts more weight on mandals with the highest teacher training values, and on areas in the second quartile of training. This approximates a case where experimental locations are chosen based on average teacher training, with a preference for teacher training levels predictive of high treatment effects.[20]

Given this index $g(C_i, U_i)$, we include schools in the sample if $g(C_i, U_i) \geq V_i$ where V_i is normally distributed with the same mean as $g(C_i, U_i)$ and a standard deviation three times as large. Although the structure of the sample construction is similar to the participation on treatment effects case discussed above, the difference in ATEs between the trial and target populations is less extreme. The target population ATE is again 0.074, while that in the trial population is 0.119.

We again consider the case where we cannot observe a subset of the variables used in the construction of the index $g(C_i, U_i)$. As above, we consider varying the size of the subset which is unobserved and calculate the same participation on observables quantities as above.

Panel B of Table 2 replicates Panel A for this setting. When we treat larger sets as unobserved, the estimate is further from the target population ATE, and closer to the trial population estimate. The values of Φ are largest for the largest exclusion set, and reflect the share of covariates missing from the observable set.

Figure 3 plots the distribution of the values Φ that would generate the target population ATE as we consider different sets of observables. With small exclusion sets the values are relatively small, although with large sets of variables treated as unobserved the results are noisier, and sometimes imply very large values of Φ to match the true treatment effect. Figure 4, graphs the value of Φ to match the true bias against the ratio of the covariance of $g(C_i, U_i), E_{P_S}[TE_i|C_i, U_i]$ to the covariance of $E_{P_S}[g(C_i, U_i)|C_i], E_{P_S}[TE_i|C_i]$; the correlation is also very strong.

[20]The effects of teacher training here is non-linear, reflecting the actual patterns in the data.

6 Applications

This section discusses a number of specific examples applying our framework to papers in the literature. In each case, we describe the setting, our identification of a plausible target population, and our results.

6.1 Attanasio, Kugler and Meghir (2011)

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

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

As discussed in Section 2.1, Attanasio et al (2011) is representative of the broad class of papers in which participants volunteer for a study and treatment is randomized among volunteers.

In this case, a possible question of interest for policy is whether it would be a good idea to extend the vocational training program to all individuals - perhaps making it part of a school curriculum.[21] If the ATE estimated in this experiment is valid for such an expansion, the answer is likely yes. Given the sample construction, however, it seems unlikely that the ATE for the experimental sample is representative of that for the population as a whole. In particular, individuals who volunteer may be those who expect vocational training to work for them. The in-sample ATE could then be biased upwards relative to the full population ATE.

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

Target Population Data A key step in implementing our approach in any setting is to identify the target population of interest and to find a data source for comparable information on that group. In this case, a natural target population is all eligible individuals in the cities in question. In the original paper, the authors note that there is a nationally representative survey, the National Household Survey, which can be used to provide target population estimates. The authors provide some general comparisons to this population, but do not formally adjust for differences in population characteristics.

The program studied in Attanasio et al (2011) is generally not open to people with degrees beyond high school.[22] We therefore exclude individuals with more than a high school education from the target population and think of our results as a bound on the population effect for those with a high school degree or less.[23]

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

Results Table 3 implements our calculations for each of the primary outcomes reported in Table 4A of Attanasio et al (2011) - that is, the main results for women on which the authors focus.[24]

Column 2 shows the baseline effects, which are mostly significant and show better labor market outcomes for the treatment group. Column 3 shows the estimate after correcting for participation on observables as described in Section 5.1 above. This correction substantially

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

[23] We also exclude from the analysis the small number of people in the trial population who report having more than a high school degree, who should not have been eligible (this is only 1% of the sample and makes little difference to the trial population results). These individuals may have been included in error, have special circumstances, or have reported their education incorrectly.

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

attenuates the estimates; in some cases the adjusted effect is zero or negative. The primary reason is that there is substantial treatment effect heterogeneity on education. While the magnitude of the differences in education may seem fairly small, when scaled by the large degree of heterogeneity on this dimension the implied treatment effect difference is substantial. The results on increased wage and salary earnings are the least affected.

Columns 4 and 5 show two measures of external validity. First, Column 4 reports bounds on the target population ATE under the assumption that $\Phi \in [1, 2]$. For the most part these bounds are much less encouraging about the effectiveness of the program than are the baseline estimates. The only exception is earnings, where the impacts seem somewhat robust. Second, Column 5 shows the value of Φ corresponding to a zero ATE. These figures are, in some cases, less than 1 - this implies that the unobservables would have to operate in the opposite direction of the observables to produce an effect of zero, which as noted above in section 4.2 is ruled out if one assumes participation on the treatment effect.

Confidence sets are reported in Columns 4 and 5. In Column 4 these are generally large, corresponding to the relatively large adjustments. The confidence sets in Column 5, which are mostly infinite, illustrate the fairly poor inference properties of $\Phi(0)$ in this setting. As we discuss above this is a known issue, and is related to the problem of weak instruments.[25]

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

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

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

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

In this case, a question of interest for the firm may be whether it would be sensible to have many or all eligible call center employees work from home.[26] If the ATE estimated in the experiment is valid for the entire workforce, then the answer is likely yes. In fact, given the expense of running an office, this might be a good policy even if the ATE on productivity were zero or slightly negative.

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

Target Population Data It is straightforward to identify the target population for this study: it is all workers at the firm with private bedrooms.[27] Bloom et al (2015) collect some basic characteristics for this overall population of workers. These can then be compared to the volunteers.

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

Results Table 4 shows results. Column 2 shows the baseline effect for the primary outcome in the paper, which is the increase in overall performance. Column 3 shows estimates from the regression-based correction for participation on observables. This slightly decreases the effect, from 0.22 to about 0.20.[28]

Columns 4 and 5 again show the two measures of external validity. Column (4) illustrates

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

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

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

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

6.3 Dupas and Robinson (2013)

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

The experiment finds significant results for some combinations of outcomes and treatments. We focus on the combinations of outcomes and treatments which the authors suggest should be significant based on their theory. The first two columns of Table 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.[29] External validity concerns again arise here because of the sampling frame: ROSCA participants are likely to be a non-representative group. Most notably, ROSCAs are designed in part as a savings and investment mechanism, so participants may differ on characteristics related to their responsiveness to savings products.

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

To frame this in our language, our concern is that there is some feature - say, interest in

[29] 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.

saving - which influences ROSCA participation and also co-varies with the treatment effect. We observe some correlates of this feature, but there is further private information among the participants. The question is how important this private information would have to be in order to produce ATEs equal to zero. We can use our approach to calculate sensitivity values Φ for each outcome-treatment pair in the data. These can be interpreted as measuring how much of the covariance between the participation variable and the treatment effect would need to be due to private information in order to eliminate the result. As above, higher values point to a more robust result.

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

Results Table 5 shows the results. For most of the analyses adjustment for participation on observables moves the coefficient towards zero, suggesting the patterns of participation are such that those with larger treatment effects are more likely to be in the sample. However, there is substantial variation across the outcome-treatment pairs in the degree of sensitivity. For example, the relationship between the treatments and the variable measuring whether people have trouble affording treatment is fairly robust. The participation on observables adjustment is extremely small and in one case goes in the opposite direction, suggesting that adjustments for participation on observables actually increase the ATE. By contrast, the results for investment in health show larger adjustments.

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

6.4 Olken, Onishi, and Wong (2014)

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

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

Target Population Data To apply our approach, we need to identify a set of characteristics from the target population. The concern is that the sub-districts in the experiment are not representative of all of Indonesia. We therefore focus on sub-district-level characteristics. The data collected in the experiment did not include comparable information about the target population. However, we can extract these data from a nationally representative survey of Indonesia (SUSENAS) which we merge at the level of the sub-district with the data used in Olken et al (2014). The target population corresponds to all of Indonesia.[30] This is an example of how our approach might be used in a setting like this, where an experiment includes a subset of locations within a country or region, and external data is available for the entire region.

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

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

Table 6 shows the results. As noted, the baseline impact is insignificant for most outcomes. However, a notable feature of this setting is that in all cases but one correction for participation on observables increases the estimated size of the effect. Consequently, most of the sensitivity measures in Column 5 are negative. Under our baseline assumptions, these results suggest that the effects in the trial population may actually *understate* the overall effects in the target population in many cases.

This is made most concrete by Column 4 of Table 6, which shows bounds under the assumption that $\Phi \in [1, 2]$. For all of the outcomes, the bounds are substantially more encouraging about the impact of the experiment than are the baseline effects. Based on the confidence intervals, many of these adjusted effects are significantly different from zero. In this case, our analysis casts doubt on the conclusion that this intervention does not change outcomes. It may simply be that the population used for the trial is not the one for which this intervention was most effective.

7 Discussion

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

Regression Discontinuity Regression discontinuity estimates are identified from behavior at the discontinuity; this leads to concern that treatment effects may differ for individuals distant from the discontinuity (Bertanha and Imbens, 2014; Angrist and Rokkanen, 2015; Rokkanen, 2015). Consider a sharp RD design with running variable R_i for individual i, where $D_i = 1\{R_i \geq r^*\}$ is an indicator for R_i exceeding some threshold r^*. The regression discontinuity approach estimates the treatment effect by a regression of Y_i on D_i in a small

neighborhood of R_i around r^*. We can define the observations in an infinitesimal neighborhood of r^* as the trial population. The target population is the population for the full range of R_i. We can then treat this problem as in the experimental case above.[31] Note, however, that relative to approaches proposed in the literature, our approach does not exploit additional structure from the regression discontinuity setting and so may yield less precise results.

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

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

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

[31] For our absolute continuity assumption (Assumption 2) to hold, C_i must not include R_i.

8 Conclusion

This paper considers the problem of external validity when the trial population for a study differs from the target population of interest. We focus on the case where participation in the trial population is driven, at least in part, by 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 relative importance of observables and unobservables in explaining participation.

Our approach is straightforward to implement. The only added data requirement beyond 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.

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

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

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

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

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

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

Bell, Stephen H. and Elizabeth A. Stuart, "On the "Where" of Social Experiments: The Nature and Extent of the Generalizability Problem," *New Directions for Evaluation*, 2016, *2016* (152), 47–59.

___ **, Robert B. Olsen, Larry L. Orr, and Elizabeth A. Stuart**, "Estimates of External Validity Bias When Impact Evaluations Select Sites Nonrandomly," *Educational Evaluation and Policy Analysis*, 2016, *38* (2), 318–335.

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

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

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

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

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

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

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

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

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

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

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

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

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

Heckman, James, "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *47* (1), 153–161.

___ **and Bo E. Honore**, "The Empirical Content of the Roy Model," *Econometrica*, 1990, *58* (5), 1121–1149.

___ **and Edward Vytlacil**, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

___ **and** ___ , "Econometric Evaluation of Social Programs Part I: Causal Models, Structural Models And Econometric Policy Evaluation," in James Heckman and Edward Leamer, eds., *Handbook of Econometrics*, Vol. 6B, Elsevier, 2007, chapter 70, pp. 4779–4874.

___ **and** ___ , "Econometric Evaluation of Social Programs Part II: Causal Models, Structural Models And Econometric Policy Evaluation," in James Heckman and Edward Leamer, eds., *Handbook of Econometrics*, Vol. 6B, Elsevier, 2007, chapter 71, pp. 4779–4874.

___ **and Salvadore Navarro**, "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics*, 2007, *136* (2), 341–396.

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

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

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

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

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

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

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

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

Jensen, Robert, "Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India," *The Quarterly Journal of Economics*, 2012, *127* (2), 753–792.

Kline, Patrick, "Oaxaca-Blinder as a Reweighting Estimator," *American Economic Review*, 2011, *101* (3), 532–537.

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

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

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky, "Using Instrumental Variables for Inference About Policy Relevant Treatment Effects," 2017. Working Paper.

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

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

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

Olsen, Robert B. and Larry L. Orr, "On the "Where" of Social Experiments: Selecting More Representative Samples to Inform Policy," *New Directions for Evaluation*, 2016, *2016* (152), 61–71.

___ , ___ , Stephen H. Bell, and Elizabeth A. Stuart, "External Validity in Policy Evaluations That Choose Sites Purposively," *Journal of Policy Analysis and Management*, 2013, *32* (1), 107–121.

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

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

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

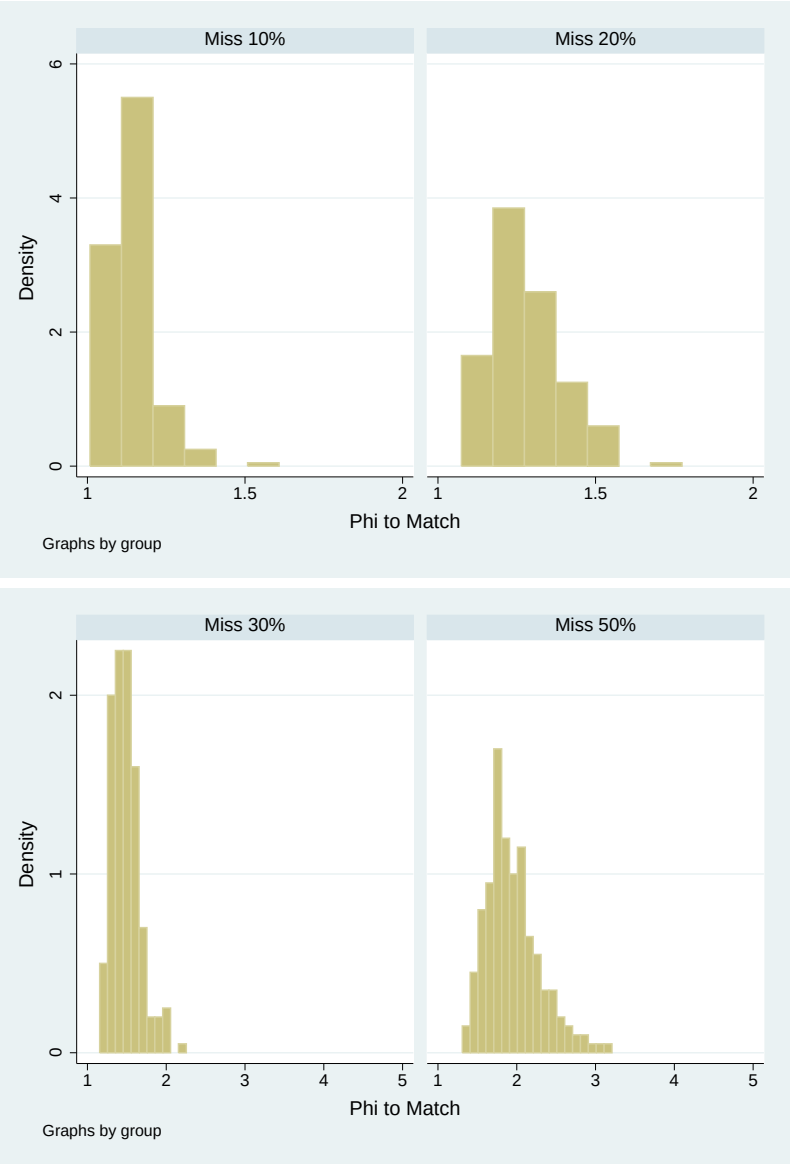
Rosenbaum, Paul R. and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 1983, *70* (1), 41–55.

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

Vytlacil, Edward, "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 2002, *70* (1), 331–341.

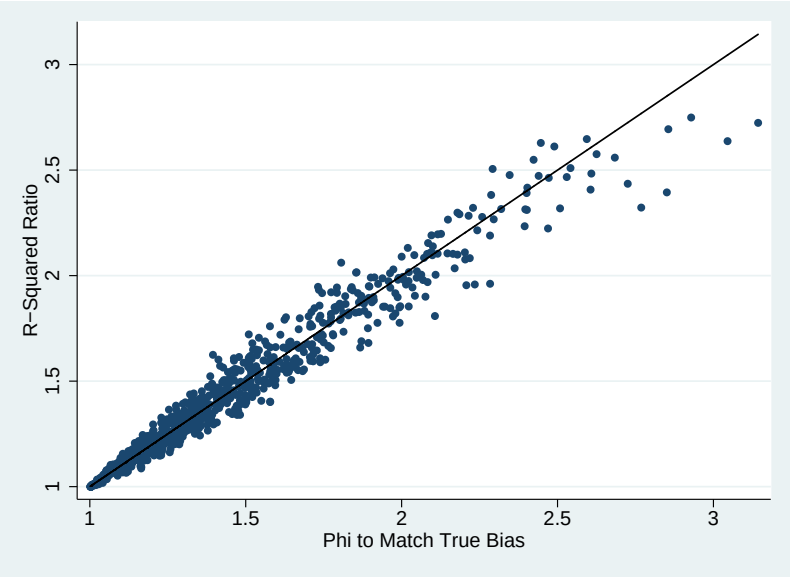
___ , "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," *Oxford Bulletin of Economics and Statistics*, 2006, *68* (4), 515–518.

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



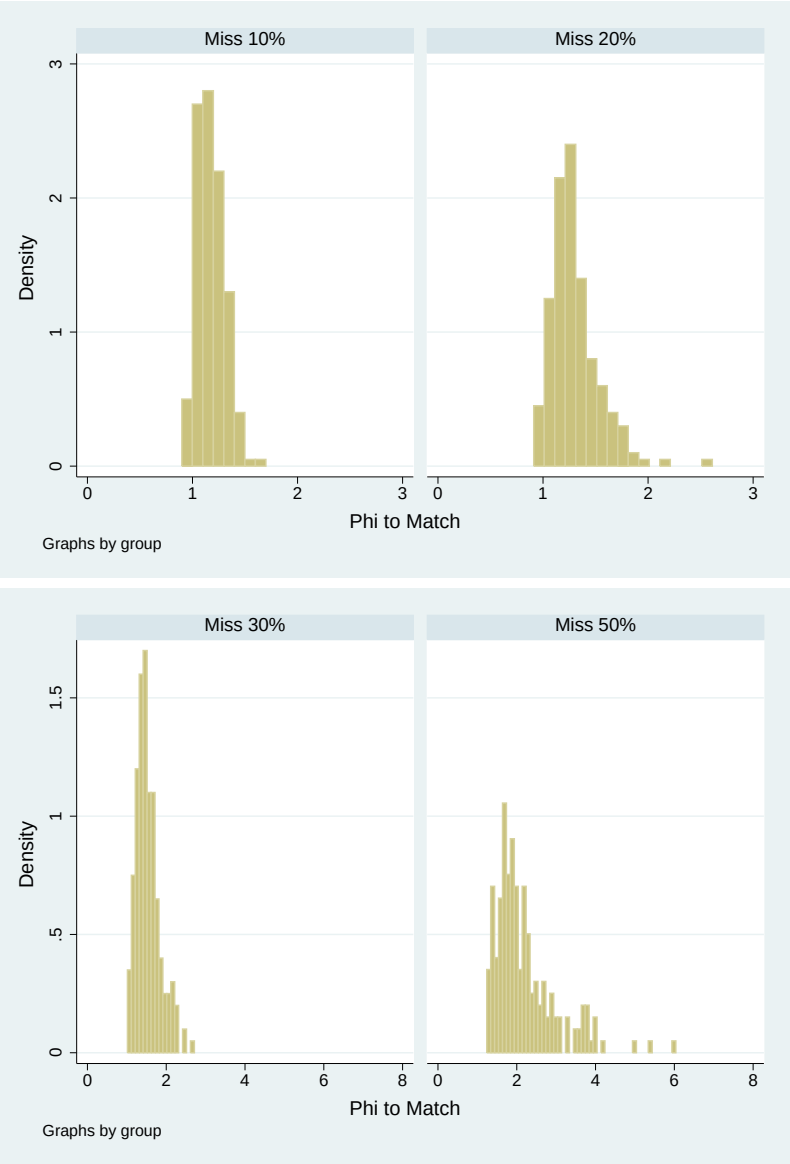
Notes: This figure shows the values of Φ which would match the population effect in the example based on Muraldiharan and Sundararaman (2011) with varying sets of covariates treated as unobserved. In this example the sample is constructed based on the predicted treatment effect, where the prediction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

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



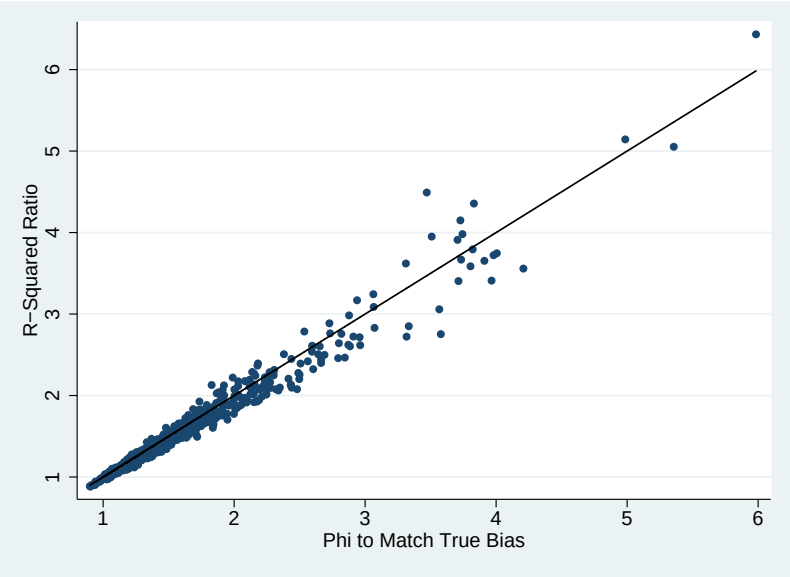
Notes: This figure shows the relationship between the values of Φ to match the population effect and the relative R-squared in a regression of the treatment effect on all variables in the example based on Muraldiharan and Sundararaman (2011) with 10-50% covariates treated as unobserved. In this example the sample is constructed based on the predicted treatment effect, where the prediction is constructed using observables and unobservables. The 45 degree line is plotted in black.

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



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

Figure 4: **Relationship between Φ and R-squared Ratio**



Notes: This figure shows the relationship between the values of Φ to match the population effect and the relative R-squared in a regression of the treatment effect on all variables in the example based on Muraldiharan and Sundararaman (2011) with 10-50% of the covariates treated as unobserved. In this example the sample is constructed based on mandal-level average teacher training. The 45 degree line is plotted in black.

Table 1: **Bias Decomposition**

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

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

Table 2: **Auxiliary Evidence, Participation Models with Varying Exclusion Sets**

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

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

Table 3: **Application: Attanasio et al(2011)**

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

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

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

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

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

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

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

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

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

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

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

Appendix A: Proofs

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

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

while by Bayes Theorem we can take

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

Thus,

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

□

Proof of Lemma 2 Assumption 2 implies 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,

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

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

Proof of Lemma 3 Let $V_i \sim N(0,1)$, noting that this distribution satisfies the conditions assumed in the Lemma. Let $g(C_i, U_i) = F_N^{-1}(E_P[S_i|C_i = c, U_i = u])$, for F_N^{-1} the standard normal inverse cdf, and note that $Pr\{g(C_i, U_i) \geq V_i|C_i, U_i\} = E_P[S_i|C_i = c, U_i = u]$ by construction. Note, finally, that

$$F_N^{-1}(\nu) \leq g(C_i, U_i) \leq F_N^{-1}(1 - \nu)$$

by construction, which completes the proof.□

Proof of Lemma 4 If we apply Lemma 2 with (C_i, U_i) in place of X_i, we see that for any function $f(C_i, U_i)$,

$$E_P[f(C_i, U_i)] = E_{P_S}[W_i f(C_i, U_i)] = E_{P_S}\left[\frac{p_{(C,U)}(C_i, U_i)}{p_{(C,U),S}(C_i, U_i)} f(C_i, U_i)\right].$$

Moreover, from Lemma 1 with (C_i, U_i) in place of X_i, we see that

$$W_i = \frac{p_{(C,U)}(C_i, U_i)}{p_{(C,U),S}(C_i, U_i)} = \frac{E_{P_S}[S_i]}{E_{P_S}[S_i|C_i, U_i]} = \frac{1 - E[F_V(g(C_i, U_i))]}{1 - F_V(g(C_i, U_i))} = w(g(C_i, U_i))$$

for F_V the distribution function of V_i. Our assumptions imply that $w(\cdot)$ is continuously differentiable, as desired.

To complete the proof, note that since $(Y_i(0), Y_i(1)) \perp S_i | C_i, U_i$, for all $f(\cdot)$ we have

$$E_{P_S}[f(X_i)|C_i, U_i] = E_P[f(X_i)|C_i, U_i].$$

Thus,

$$E_{P_S}[w(g(C_i, U_i)) f(X_i)] = E_{P_S}[w(g(C_i, U_i)) E_P[f(X_i)|C_i, U_i]] = E_P[f(X_i)],$$

which completes the proof. □

Proof of Proposition 1 We first show that

$$Cov_{P_S}(W_i, f(X_i)) = Cov_{P_S}(W_i^*, f(X_i)) + O(\kappa^2).$$

Recall from the proof of Lemma 4 that

$$W_i = \frac{1 - E[F_V(\kappa \cdot h(C_i, U_i))]}{1 - F_V(\kappa \cdot h(C_i, U_i))},$$

where only the denominator depends on $h(C_i, U_i)$.

Mean-Value Expansion of W_i: For brevity of notation, let $h_i = h(C_i, U_i)$. Let us consider a mean-value expansion of W_i around $E_{P_S}[h_i]$:

$$W_i = \frac{1 - E_P[F_V(\kappa \cdot h_i)]}{1 - F_V(\kappa \cdot E_{P_S}[h_i])} + \frac{1 - E_P[F_V(\kappa \cdot h_i)]}{(1 - F_V(\kappa \cdot h_i^*))^2} \kappa \cdot f_V(\kappa \cdot h_i^*)(h_i - E_{P_S}[h_i]),$$

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

$$Pr_P\{F_V(\kappa \cdot h_i) \in [F_V(0) - \varepsilon, F_V(0) + \varepsilon]\} = 1.$$

Thus, for such κ we know that

$$\frac{1 - E_P[F_V(\kappa \cdot h_i)]}{(1 - F_V(\kappa \cdot h_i^*))^2} \leq \frac{1 - E_P[F_V(\kappa \cdot h_i)]}{(1 - F_V(0) - \varepsilon)^2}.$$

If we consider the difference

$$W_i - W_i^* =$$

$$\left(\frac{1 - E_P[F_V(\kappa \cdot h_i)]}{(1 - F_V(\kappa \cdot h_i^*))^2} \kappa \cdot f_V(\kappa \cdot h_i^*) - \frac{1 - E_P[F_V(\kappa \cdot h_i)]}{(1 - F_V(\kappa \cdot E_{P_S}[h_i]))^2} \kappa \cdot f_V(\kappa \cdot E_{P_S}[h_i]) \right) (h_i - E_{P_S}[h_i]),$$

note that this is bounded in absolute value by

$$\left|\left(\frac{1-E_P[F_V(\kappa\cdot h_i)]}{\left(1-F_V\left(\kappa\cdot h_i^*\right)\right)^2}-\frac{1-E_P[F_V(\kappa\cdot h_i)]}{\left(1-F_V\left(\kappa\cdot E_{P_S}[h_i]\right)\right)^2}\right)\kappa\cdot f_V\left(\kappa\cdot h_i^*\right)\left(h_i-E_{P_S}\left[h_i\right]\right)\right|$$
$$+\left|\frac{1-E_P[F_V(\kappa\cdot h_i)]}{\left(1-F_V\left(\kappa\cdot E_{P_S}[h_i]\right)\right)^2}\left(\kappa\cdot f_V\left(\kappa\cdot h_i^*\right)-\kappa\cdot f_V\left(\kappa\cdot E_{P_S}\left[h_i\right]\right)\right)\left(h_i-E_{P_S}\left[h_i\right]\right)\right|.$$

The fact that $f_V(v)$ is Lipschitz implies that for $\kappa\in[0,\kappa_\varepsilon]$ the second term is bounded in absolute value by

$$\frac{1}{\left(1-F_V(0)-\varepsilon\right)^2}\kappa^2K\left(h_i-E_{P_S}\left[h_i\right]\right)^2=\kappa^2K_2\left(h_i-E_{P_S}\left[h_i\right]\right)^2,$$

for K the Lipschitz constant and K_2 another constant. For the first term, note that another mean-value expansion yields

$$\frac{1-E_P\left[F_V\left(\kappa\cdot h_i\right)\right]}{\left(1-F_V\left(\kappa\cdot h_i^*\right)\right)^2}-\frac{1-E_P\left[F_V\left(\kappa\cdot h_i\right)\right]}{\left(1-F_V\left(\kappa\cdot E_{P_S}\left[h_i\right]\right)\right)^2}=2\frac{1-E_P\left[F_V\left(\kappa\cdot h_i\right)\right]}{\left(1-F_V\left(\kappa\cdot h_i^*\right)\right)^3}\kappa\cdot f_V\left(\kappa\cdot\tilde{h}_i\right)\left(h_i^*-E_{P_S}\left[h_i\right]\right)$$

for $\tilde{h}_i$ a value between h_i^* and $E_{P_S}[h_i]$. Since $f_V(v)$ is Lipschitz and integrates to one we know that it it bounded above by K, and thus that the first term is bounded in absolute value by

$$\frac{1-E_P\left[F_V\left(\kappa\cdot h_i\right)\right]}{\left(1-F_V(0)-\varepsilon\right)^3}\kappa^2K^2\left(h_i-E_{P_S}\left[h_i\right]\right)^2.$$

Thus, we see that

$$\left|W_i-W_i^*\right|\leq\kappa^2K^*\left(h_i-E_{P_S}\left[h_i\right]\right)^2$$

for a constant K^*.

Next, for some function $f(X_i)$, let us consider the approximation error

$$Cov_{P_S}\left(W_i,f\left(X_i\right)\right)-Cov_{P_S}\left(W_i^*,f\left(X_i\right)\right)=Cov_{P_S}\left(W_i-W_i^*,f\left(X_i\right)\right)$$
$$=E_{P_S}\left[\left(W_i-W_i^*\right)f\left(X_i\right)\right]-E_{P_S}\left[W_i-W_i^*\right]E_{P_S}\left[f\left(X_i\right)\right].$$

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

$$\kappa^2K^*\cdot E_{P_S}\left[\left(h_i-E_{P_S}\left[h_i\right]\right)^2\left|f\left(X_i\right)\right|\right],$$

while the second is bounded by $\kappa^2K^*\cdot E_{P_S}\left[\left(h_i-E_{P_S}\left[h_i\right]\right)^2\right]\left|E_{P_S}\left[f\left(X_i\right)\right]\right|$.

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

Behavior of moments for small κ Note that

$$E_{P_S}\left[\left(h_i-E_{P_S}\left[h_i\right]\right)^2\left|f\left(X_i\right)\right|\right]=E_P\left[W_i^{-1}\left(h_i-E_{P_S}\left[h_i\right]\right)^2\left|f\left(X_i\right)\right|\right]$$

$$= E_P\left[\left(\frac{1-F_V(\kappa\cdot h_i)}{1-E_P[F_V(\kappa\cdot h_i)]}\right)(h_i - E_{P_S}[h_i])^2 |f(X_i)|\right].$$

Since

$$E_P\left[\left(\frac{1-F_V(\kappa\cdot h_i)}{1-E_P[F_V(\kappa\cdot h_i)]} - 1\right)^2\right] \to 0$$

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

$$E_P\left[\left(\frac{1-F_V(\kappa\cdot h_i)}{1-E_P[F_V(\kappa\cdot h_i)]} - 1\right)(h_i - E_{P_S}[h_i])^2 |f(X_i)|\right] \to 0,$$

and thus that

$$E_{P_S}\left[(h_i - E_{P_S}[h_i])^2 |f(X_i)|\right] \to E_P\left[(h_i - E_P[h_i])^2 |f(X_i)|\right]$$

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

$$E_{P_S}[|f(X_i)|] \to E_P[|f(X_i)|]$$

and

$$E_{P_S}\left[(h_i - E_{P_S}[h_i])^2\right] \to E_P\left[(h_i - E_P[h_i])^2\right]$$

as $\kappa \to 0$.

Completing the proof of first statement: Combing these results, we see that under our assumptions above,

$$Cov_{P_S}(W_i - W_i^*, f(X_i)) = O(\kappa^2)$$

as $\kappa \to 0$.

Proof of Second Statement We next show that

$$\frac{Cov_{P_S}(W_i, f(X_i))}{Cov_{P_S}(W_i^*, f(X_i))} \to 1.$$

By the argument above we know that

$$\frac{Cov_{P_S}(W_i, f(X_i))}{Cov_{P_S}(W_i^*, f(X_i))} = \frac{Cov_{P_S}(W_i, f(X_i))}{Cov_{P_S}(W_i, f(X_i)) + O(\kappa^2)}.$$

Next, note that by Corollary 1,

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

which we can re-write as

$$E_P\left[\left(1 - \frac{1-F_V(\kappa\cdot h_i)}{1-E_P[F_V(\kappa\cdot h_i)]}\right) f(X_i)\right].$$

Note that the assumption that $f_V(\cdot)$ is Lipschitz, together with the fact that it is positive

and integrates to one, implies that it is bounded. The dominated convergence theorem thus implies that $\frac{\partial}{\partial \kappa} E_P \left[F_V \left(\kappa \cdot h_i \right) \right] = E_P \left[h_i f_V \left(\kappa \cdot h_i \right) \right]$, and that

$$\frac{\partial}{\partial \kappa} \frac{1 - F_V \left(\kappa \cdot h_i \right)}{1 - E_P \left[F_V \left(\kappa \cdot h_i \right) \right]} = - \frac{h_i f_V \left(\kappa \cdot h_i \right)}{1 - E_P \left[F_V \left(\kappa \cdot h_i \right) \right]} + \frac{1 - F_V \left(\kappa \cdot h_i \right)}{\left(1 - E_P \left[F_V \left(\kappa \cdot h_i \right) \right] \right)^2} E_P \left[h_i f_V \left(\kappa \cdot h_i \right) \right].$$

Since this quantity is bounded for small κ, another application of the dominated convergence theorem implies that

$$\frac{\partial}{\partial \kappa} Cov_{P_S} \left(W_i, f \left(X_i \right) \right) =$$

$$E_P \left[\left(\frac{h_i f_V \left(\kappa \cdot h_i \right)}{1 - E_P \left[F_V \left(\kappa \cdot h_i \right) \right]} - \frac{1 - F_V \left(\kappa \cdot h_i \right)}{\left(1 - E_P \left[F_V \left(\kappa \cdot h_i \right) \right] \right)^2} E_P \left[h_i f_V \left(\kappa \cdot h_i \right) \right] \right) f \left(X_i \right) \right],$$

and thus that $Cov_{P_S} \left(W_i, f \left(X_i \right) \right)$ is continuously differentiable in κ on a neighborhood of zero. Evaluating this derivative at $\kappa = 0$ yields

$$\frac{\partial}{\partial \kappa} Cov_{P_S} \left(W_i, f \left(X_i \right) \right) \big|_{\kappa = 0} = E_P \left[\left(\frac{h_i f_V \left(0 \right)}{1 - E_P \left[F_V \left(0 \right) \right]} - \frac{E_P \left[h_i f_V \left(0 \right) \right]}{1 - E_P \left[F_V \left(0 \right) \right]} \right) f \left(X_i \right) \right].$$

Thus, we see that this derivative is nonzero so long as h_i is correlated with $f \left(X_i \right)$ and $f_V \left(0 \right) \neq 0$, which we have already assumed. The result follows immediately. □

Proof of Lemma 5 Note that by the law of iterated expectations and the fact that $E_P \left[TE_i | C_i, U_i \right] = E_{P_S} \left[TE_i | C_i, U_i \right]$,

$$E_{P_S} \left[g \left(C_i, U_i \right) | C_i \right] = \alpha_1 E_{P_S} \left[TE_i | C_i \right] + g_C \left(C_i \right).$$

Thus, by another application of the law of iterated expectations

$$Cov_{P_S} \left(E_{P_S} \left[g \left(C_i, U_i \right) | C_i \right], E_{P_S} \left[TE_i | C_i \right] \right) = \alpha_1 Var_{P_S} \left(E_{P_S} \left[TE_i | C_i \right] \right) + \alpha_1 Cov_{P_S} \left(TE_i, g_C \left(C_i \right) \right)$$

where the second term is zero by assumption. Likewise,

$$Cov_{P_S} \left(g \left(C_i, U_i \right), E_{P_S} \left[TE_i | C_i, U_i \right] \right) = Cov_{P_S} \left(g \left(C_i, U_i \right), TE_i \right) = \alpha_1 Var_{P_S} \left(E_{P_S} \left[TE_i | C_i, U_i \right] \right).$$

□

Appendix B: Additional Results

This appendix details several results mentioned in the main text. We first discuss a model, mentioned in Section 5.1 of the main text, under which Φ can be interpreted as the share of relevant factors captured by the observed covariates. We then provide additional details of and justification for our inference procedures.

Appendix B.1 Interpretation of Φ Under Random Selection of Observables

To build intuition for the behavior of Φ we consider a model in which the observable covariates represent a random subset of a larger collection of latent factors.

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

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

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

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

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

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

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

Finally, let us suppose that both $g(C_i, U_i)$ and $E_{P_S}\left[TE_i|F_i\right]$ are linear in the factors,

$$g(C_i, U_i) = \mu_g + \gamma_F' F_i$$

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

This implies that the conditional expectations of these variables given covariates are linear in C_i as well

$$E_{P_S}\left[g(C_i, U_i)|C_i\right] = \hat{\mu}_g + \gamma_C' C_i,$$

$$E_{P_S}\left[TE_i|C_i\right] = \hat{\mu}_{TE} + \gamma_C' C_i.$$

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

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

the coefficients above are defined as

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

and

$$(\hat{\mu}_g, \hat{\mu}_{TE}) =$$

$$\left(\mu_g - \gamma_C'\mu_C, \mu_{TE} - \delta_C'\mu_C\right).$$

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

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

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

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

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

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

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

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

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

while

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

Therefore, we see that

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

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

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

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

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

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

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

Appendix B.2 Inference Details

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

Appendix B.2.1 Confidence Set for $\Phi\left(t_P^*\right)$

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

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

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

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} t_P^* - E_{P_S}\left[TE_i\right] \\ E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right] \end{pmatrix}.$$

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

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

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

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

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

Appendix B.2.1 Confidence Set for $E_P\left[TE_i\right]$ Under Bounds on Φ

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

$$\left(E_{P_S}\left[TE_i\right] + \Phi_L\left(E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]\right), E_{P_S}\left[TE_i\right] + \Phi_U\left(E_P\left[E_{P_S}\left[TE_i|C_i\right]\right] - E_{P_S}\left[TE_i\right]\right)\right),$$

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

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

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

$[\Phi_L, \Phi_U]$ if and only if

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

or equivalently, either

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

or

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

holds.

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

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

only if

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

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

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

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

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

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

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

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

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

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

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

Appendix C: Tables and Figures

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

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

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

Table 2: **Observable Characteristics, Bloom et al (2015)**

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

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

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

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

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

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

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

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

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

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

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