Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India *

Michael Gechter†

January 20, 2015

Abstract

To what extent are causal effects estimated in one region or time period informative about another region or time? In this paper, I derive bounds on the average causal effect in a context of interest using experimental evidence from another context. I use differences in outcome distributions for individuals with the same characteristics and treatment status in the original study and the context of interest to learn about unobserved differences across contexts. Greater differences in outcome distributions generate wider bounds. Empirically, I explore using experimental results on the return to cash transfers to male microentrepreneurs in one Mexican city in 2006 to predict the returns among male microentrepreneurs in urban Mexico in 2012. I show that existing methods would lead us to be overconfident in extrapolating from the small experiment to all of urban Mexico in 2012. Using data from a pair of remedial education experiments carried out in urban India, I show that the methods suggested in this paper are able to recover average causal effects in one city using results from the other where existing methods are unsuccessful.

*I am grateful to Dilip Mookherjee, Hiroaki Kaido and Kevin Lang for guidance, support and encouragement. I would also like to thank Kehinde Ajayi, Manuel Arellano, Sam Bazzi, Xavier d'Haultfoeuille, Rajeev Dehejia, Isabella Dobrescu, Iván Fernández-Val, Claudio Ferraz, Andy Foster, Patrik Guggenberger, Rema Hanna, Stefan Hoderlein, Asim Khwaja, Horacio Larreguy, David Lam, Leigh Linden, Michael Manove, Shanthi Nataraj, Rohini Pande, Justin Sandefur, Johannes Schmieder, Jeff Smith, Duncan Thomas, Nate Young and seminar participants at BU, Laval, Penn State, NEUDC and the WEAI Graduate Student Workshop for useful conversations and feedback. I especially thank David McKenzie for discussions and sharing the data from McKenzie and Woodruff (2008). Financial support from the Boston University Department of Economics is gratefully acknowledged. Gloria Sarmiento-Becerra provided excellent research assistance, supported by the Boston University Department of Economics MA-RA Mentor Program.

†Department of Economics, Boston University. Email: mgechter@bu.edu.

1 Introduction

What do causal effects measured in one place tell us about causal effects in another place or at another time? It is clear that not every finding applies in every context. Some authors have recently protested against policy recommendations they see as based on implicit extrapolation from a small number of experiments to a wide variety of dissimilar contexts (Deaton (2010); Pritchett and Sandefur (2013)). Empirically, a growing body of work finds different effects of identical policies among individuals with the same observed characteristics living in different contexts (e.g. Allcott (2015); Attanasio, Meghir, and Szekely (2003)). Unobserved differences between populations remain, even when considering individuals with the same observed characteristics.

In this paper, causal effects from one place may be only partially informative about effects elsewhere. I derive bounds on the average causal effect in a context of interest using experimental evidence from another context. I use differences in outcome distributions for individuals with the same characteristics and treatment status in the original study and the context of interest to learn about unobserved differences across contexts[1]. Greater differences in outcome distributions generate wider bounds. The bounds represent a practical solution to the problem of assessing generalizability of experimental results from one context to another and are easily computed using software provided by the author for any pair of contexts. They formalize the idea that the conclusions we can draw about the average causal effect in the context of interest and the strength of assumptions required to do so depends on the similarity between the two contexts[2].

I consider settings where we have run a randomized evaluation of a pilot program and wish to know what we can conclude about the effect of the program in another context. The experimental treatment group has access to the program, while the control group does not. As part of the evaluation, we collected data on characteristics and outcomes of individuals participating in the experiment. We also have data on outcomes and characteristics of individuals in the alternative context, possibly coming from a separate survey. Since the program is a pilot, individuals in the alternative context do not have access to the program[3]. For each distinct set of characteristics, we thus have the distributions of treated and un-

[1]When we do not have experiments with context-level characteristics we believe are sufficiently similar to the context of interest, unobserved differences necessarily include differences in context-level characteristics.

[2]See Heckman, Moon, Pinto, Savelyev, and Yavitz (2010) and McKenzie and Woodruff (2008) who assess the external validity of experimental results on the basis of the similarity of the experimental populations to larger populations of interest.

[3]The analysis can easily be extended to the case when individuals choose their treatment status and an experiment denies treatment to a random subset of individuals who would wish to be treated (see Bitler, Domina, and Hoynes (2014) for an example of such an experiment).

treated outcomes from the experiment and the distribution of untreated outcomes from the alternative context.

The bounds I derive on the average causal effect in the context of interest for each set of characteristics are based on the assumption that the distribution of treated outcomes for a given untreated outcome in the context of interest is consistent with the experimental results. This is a weak restriction on the average causal effect because the experiment does not rule out any level of dependence between treated and untreated outcomes[4]. Except in extreme cases, we expect positive dependence between treated and untreated outcomes, to varying degrees depending on the nature of the program. Most programs cannot cause those well-off without the program to switch places with those poorly-off in absolute terms.

I therefore develop tighter bounds, indexed by the minimum level of dependence between an individual's treated and untreated outcomes we are willing to consider. When treated and untreated outcomes are perfectly dependent, differences in untreated outcome distributions are not a problem because each untreated outcome is linked to a single treated outcome. As we move away from perfect dependence, different associations between treated and untreated outcomes become possible. These different associations produce uncertainty about the average causal effect in the new context that is increasing in the difference between the distributions of untreated outcomes in the experiment and the context of interest. The width of the bounds for a given minimum dependence level provide a measure of uncertainty about the average causal effect. They also allow us to assess the assumptions on dependence between treated and untreated outcomes necessary to draw specific conclusions about the effect of the program in the context of interest, such as its ability to exceed a cost-effectiveness threshold.

I empirically evaluate the results of my bounding procedure compared to existing methods for extrapolating causal effects to new contexts. The current benchmark method (Hotz, Imbens, and Mortimer (2005), henceforth HIM) also uses outcome distributions for individuals with the same characteristics to assess generalizability, but does so within a testing framework. If we reject that the untreated outcome distributions for individuals with the same characteristics are the same, we conclude that the experiment teaches us nothing about causal effects in the context of interest. Otherwise, the HIM framework concludes the experiment is perfectly predictive for the causal effect of interest.

I first examine the generalizability of a small experiment on the returns to loosening credit constraints by providing cash transfers to very small-scale entrepreneurs in Leon,

[4]The literature on distributions of causal effects consistent with experimental results generates similarly wide bounds on functionals of interest (Heckman, Smith, and Clements (1997); Djebbari and Smith (2008); Fan and Park (2010); Kim (2014)).

Mexico in 2006 documented in McKenzie and Woodruff (2008). We would like to know what the large estimated returns (an increase in monthly profits equal to roughly 40% of the transfer in baseline specifications) in Leon in 2006 tell us about the average return for similarly small-scale microentrepreneurs in urban Mexico in 2012, as represented by that year's national microenterprise survey. The distributions of untreated outcomes are fairly similar in the Leon and 2012 urban Mexico samples so the estimated bounds are narrow for a wide range of assumptions on dependence between profits with and without the transfer. Properly accounting for the unobserved differences between the populations along with sampling variation in the small experimental sample and the national microenterprise sample leads to wide confidence intervals around the bounds. Testing equality of control outcome distributions, in contrast, would lead us to be overconfident in our prediction of the average return. Perversely, using the HIM method, we would compute a narrower confidence interval on the predicted causal effect for urban Mexico in 2012 than on the causal effect in the original experiment.

Second, to check the predictions of different methods against measured causal effects, I use data from randomized evaluations of a remedial education program implemented in two Indian cities and described in Banerjee, Cole, Duflo, and Linden (2007). I find different average causal effects for individuals with the same observed characteristics in the two cities. The two cities' student populations are sufficiently different that equality of their untreated outcome distributions is rejected, which, in the HIM framework would lead us to believe we cannot learn anything about the causal effect in one city based on experimental results from the other. However, I show that if we assume treated and untreated outcomes are sufficiently dependent, we can exclude a substantial range of average causal effects - such as a zero effect - in one city using the results from the other. The observed causal effects in both cities are consistent with predictions based on strong dependence between the treated and untreated outcomes.

This paper extends the literature on generalizing causal effects to new contexts based on invariance assumptions on average treated outcomes or causal effects for individuals with the same observed characteristics (HIM, Attanasio et al. (2003); Angrist and Fernández-Val (2013); Angrist and Rokkanen (2013); Cole and Stuart (2010); Stuart, Cole, Bradshaw, and Leaf (2011); Pearl and Bareinboim (2014); Flores and Mitnik (2013)). In interpreting differences in untreated outcome distributions as indicative of unobserved differences in populations, I follow a long line of literature interpreting outcome quantiles as representing the effect of unobserved heterogeneity in non-separable models (see, for example, Matzkin (2007) and the references therein). Most directly, Athey and Imbens (2006) make use of this interpretation when generalizing the standard difference-in-differences estimator and derive

an estimator that is equivalent to mine under perfect dependence between the treated and untreated outcomes. In moving from a testing framework to an approach based on quantifying assumptions required to draw conclusions about causal effects, my paper relates to work by Altonji, Elder, and Taber (2005) and Altonji, Conley, Elder, and Taber (2013). Altonji et al. (2005) and Altonji et al. (2013) move from testing whether observed covariates related to an outcome are also related to a candidate instrument to providing bounds on the average causal effect whose width depends on the magnitude of the relationship between the covariates and the instrument.

The rest of the paper is organized as follows. Section 2 describes the intuition behind the proposed methods by means of a simple example. Readers uninterested in the technical details behind the methods in their full generality may wish to read section 2 then skip to the empirical results in sections 5 and 6. Beginning the theoretical discussion, section 3 sets up the problem and notation and provides a review of existing approaches to extrapolation on the basis of experimental results. In section 4, I present the derivation of the bounds. Section 5 presents the empirical results for generalizing from the 2006 Leon microenterprise experiment to urban locations in Mexico in 2012. Section 6 investigates using the results from one of the two remedial education experiments to try to predict the results in the other experiment. Section 7 concludes.

2 Intuition for the methodology: a simple example

To illustrate the intuition behind the methodological contributions, I begin by laying out a simple example involving a fictional conditional cash transfer program (CCT) that incentivizes parents to enroll children in school. Suppose we have obtained experimental results that tell us the CCT program caused a large increase in the enrollment rate in location e, from $\frac{1}{3}$ of all children to $\frac{2}{3}$ of all children. We observe only outcomes and no characteristics.

We would like to know what the results from location e tell us about the causal effect we can expect in location a, where no CCT was implemented. Whereas $\frac{1}{3}$ of children were enrolled without the CCT program in location e, $\frac{1}{2}$ of children are enrolled without the CCT in location a. We would like to know what impact the difference in the no-CCT enrollment rates will have on the average causal effect in location a. The law of total probability allows

us to decompose the average effect of the CCT program in a, denoted ATE^a, as follows.

$$\begin{aligned}
ATE^a =& P(\text{enrolled with CCT} \mid \text{enrolled without CCT}) \times P(\text{enrolled without CCT}) \\
&+ P(\text{enrolled with CCT} \mid \text{out of school without CCT}) \times P(\text{out of school without CCT}) \\
&- P(\text{enrolled without CCT}) \\
=& P(\text{enrolled with CCT} \mid \text{enrolled without CCT}) \times \frac{1}{2} \\
&+ P(\text{enrolled with CCT} \mid \text{out of school without CCT}) \times \frac{1}{2} \\
&- \frac{1}{2}
\end{aligned}$$

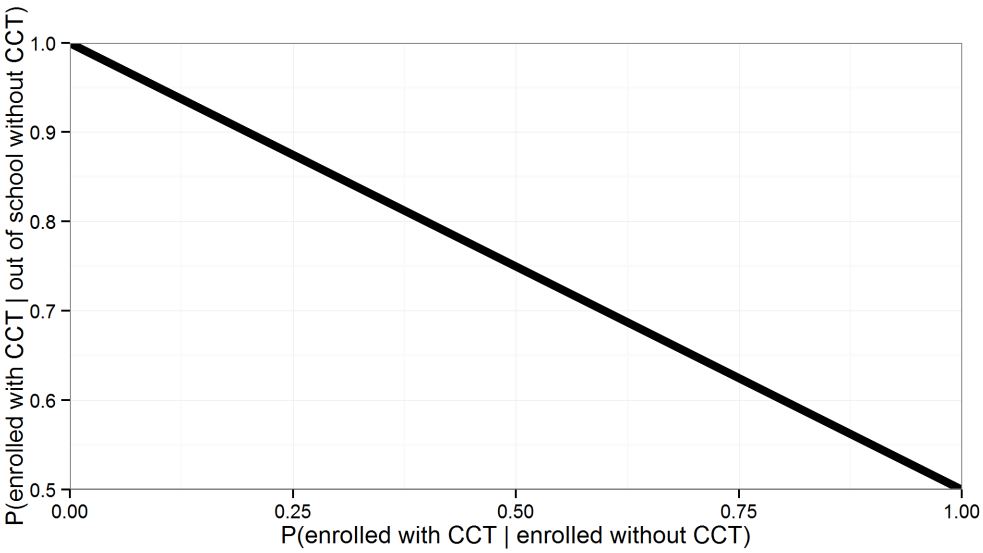
The average causal effect in a depends on two unknown probabilities: (1) the probability that an individual who **does not** enroll without the CCT would instead enroll with the CCT and (2) the probability that an individual who enrolls in school without the CCT would also enroll with the CCT.

The rationale behind (1), individuals who do not enroll without the CCT enrolling with a CCT, is clear: the program provides cash incentives for parents to enroll children in school and some parents respond to these incentives. The rationale behind (2), individuals who enroll without the CCT but would not enroll with the CCT, is less straightforward. Attanasio, Meghir, and Santiago (2012) show that CCT programs can increase wages for children by lowering the supply of child labor. An increased wage for children works against the enrollment incentives. Further, Attanasio et al. (2012) show that enrollment subsidies and child wages do not have equal opposite effects on households' enrollment decisions, as they would if only the net child wage entered into the enrollment decision. So we can think of some fraction of households who are more sensitive to child wages than they are to enrollment subsidies and would respond to a CCT by having children work. To maintain the simplicity of this example, I will refer to forces that cause children who would enroll without the CCT but would not enroll with a CCT in place as wage effects, although in principle there may be other ways for the CCT to cause children who would otherwise enroll to not enroll.

I will assume that $P(\text{enrolled with CCT} \mid \text{enrolled without CCT})$ in location a is consistent with the experimental results. There are many possible pairs of conditional enrollment probabilities that are consistent with the experimental results. The possible pairs are given in Figure 1. To see why a continuum of pairs is possible, recall that

$$\begin{aligned}
&P(\text{enrolled with CCT} \mid \text{enrolled without CCT}) \\
&= \frac{P(\text{enrolled with CCT \& enrolled without CCT})}{P(\text{enrolled without CCT})}.
\end{aligned}$$

Figure 1: Permissible distributions for P(enrolled with CCT | enrollment without CCT) in location a



We see that P(enrolled with CCT | enrolled without CCT) relies on knowledge a child's enrollment status with and without the CCT at the same time, knowledge that we cannot have. If a child is in one of the treated localities, we only observe her enrollment decision with the CCT. If she is in one of the control localities, we only observe her enrollment decision without the CCT. The question marks in table 1 indicate the unknown fractions of the population of location e falling into each of the four possible combinations of enrollment decisions with and without the CCT. The sums across rows and down columns show the information we do have from the experiment. The rows of table 1 must sum to the control group results and the columns to the treatment group results.

Table 1: The distribution of enrollment with and without the CCT is unknown in location e

		CCT		
		Out of school	Enrolled	All Control
No CCT	Out of school	?	?	$\frac{2}{3}$
	Enrolled	?	?	$\frac{1}{3}$
	All Treatment	$\frac{1}{3}$	$\frac{2}{3}$	

Our assumptions about the way the wage effects of the CCT impact the two groups of

children (those enrolling without the CCT and those who do not enroll without the CCT) will generate different predictions for the causal effect of the CCT program in location a. To see this, first consider the case where there are no wage effects or wage effects only impact children who do not enroll without the CCT. Then there are no children who enroll without the CCT but would not enroll when the CCT is in place. Our assumption allows us to fill in all the entries of table 1, as shown in table 2. The probability of enrolling with the CCT if a child is out of school without the CCT is $\frac{1}{2}$ and the increase in the fraction enrolled in location a is $\frac{1}{4}$.

Table 2: Case 1: there are no wage effects

		CCT		
		Out of school	Enrolled	All Control
No CCT	Out of school	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
	Enrolled	0	$\frac{1}{3}$	$\frac{1}{3}$
	All Treatment	$\frac{1}{3}$	$\frac{2}{3}$	

Now consider another assumption about the wage effects: they only impact those who enroll without the CCT and they are so strong that all children who would enroll without the CCT drop out. To match the distribution of control and treated group outcomes in location e, all children who are out of school without the CCT must enroll with the CCT. Again, we can fill in the unknown entries of table 1, as shown in table 3. In this rather unbelievable case, we predict no change in the fraction enrolled in location a.

Table 3: Case 2: wage effects only impact those who enroll without the CCT

		CCT		
		Out of school	Enrolled	All Control
No CCT	Out of school	0	$\frac{2}{3}$	$\frac{2}{3}$
	Enrolled	$\frac{1}{3}$	0	$\frac{1}{3}$
	All Treatment	$\frac{1}{3}$	$\frac{2}{3}$	

Assuming that wage effects impact the same fraction of both groups is somewhat more believable. To be consistent with the experimental results, this fraction must be $\frac{1}{3}$. The entries of table 1 can be filled in as shown in table 4. The predicted increase in the fraction employed is $\frac{1}{6}$.

Table 4: Case 3: wage effects impact the same fraction of both groups

		CCT		
		Out of school	Enrolled	All Control
No CCT	Out of school	$\frac{2}{9}$	$\frac{4}{9}$	$\frac{2}{3}$
	Enrolled	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$
	All Treatment	$\frac{1}{3}$	$\frac{2}{3}$	

While more believable than assuming that those enrolled with and without the CCT exchange places when the CCT is in place, assuming that wage effects have the same impact on both groups is still not very convincing. Intuitively, we believe that wage effects would have a stronger impact on enrollment decisions for children who do not enroll without the CCT. Formally, we expect positive dependence between enrollment with the CCT and enrollment without. In this paper, I follow Heckman et al. (1997) and measure dependence using the rank correlation[5] between treated and untreated outcomes for any individual. The first assumption on wage effects we considered, that there are none or they only affect enrollment decisions for children who would enroll without the CCT, generates the maximum possible rank correlation between a child's enrollment decision with and without the CCT. The third assumption, that wage effects have the same impact regardless of enrollment status without the CCT, generates a rank correlation of zero. As we have seen, different rank correlations generate different predictions for the change in enrollment caused by the CCT in location a.

How close should the rank correlation we use to predict the effect of the CCT on enrollment in location a be to the maximum possible? I consider two options. First, we might specify a range of plausible values. In this example, we might be conservative and consider rank correlations between zero and the maximum possible. Then, the gain in enrollment in location a lies between $\frac{1}{6}$ and $\frac{1}{4}$. A second option is to explore the strength of assumptions on dependence required to draw specific conclusions about the effect of the program. For example, we might consider what we need to assume about dependence to conclude that the CCT will have a positive effect on enrollment. With an enrollment rate of $\frac{1}{2}$ in location a, a zero effect on enrollment in location a is only possible when the rank correlation between enrollment with and without the CCT is the minimum possible, which is what occurs in the second case we considered, when children who enroll without the CCT all drop out with the CCT. Since this case is highly implausible, we would feel confident in our conclusion that the CCT will have a positive effect on enrollment location a.

[5]The standard Pearson product-moment correlation measures only linear dependence.

Note the key role played by the enrollment rate without the CCT in location a. If instead of $\frac{1}{2}$, the enrollment rate in location a were $\frac{2}{3}$, choosing a rank correlation between zero and the maximum possible would predict an increase in the enrollment rate due to the CCT between 0 and $\frac{1}{6}$. We would need stronger, but still believable, assumptions on dependence to predict a positive effect on enrollment.

In the following two sections, I generalize the intuition developed here to settings where we also have information about observed characteristics in the two populations, where outcomes are non-binary and where our data about locations e and a come from samples. Readers uninterested in the details of generalization may wish to skip to the empirical results in sections 5 and 6.

3 Econometric setup

Suppose we are interested in the causal effect of a binary treatment $T \in \{0, 1\}$ on an observable outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$. Each individual is associated with two potential outcomes: $Y_1 \in \mathcal{Y}_1 \subseteq \mathcal{Y}$ is her outcome if she receives treatment and $Y_0 \in \mathcal{Y}_0 \subseteq \mathcal{Y}$ is her outcome if she does not. Only one of these two outcomes is ever observed, the other is hypothetical. Mathematically, the observed outcome Y can be written as:

$$Y = TY_1 + (1 - T)Y_0.$$

Because both the observed and hypothetical outcome are defined for each individual we can also define an individual's own treatment effect $\Delta \subseteq \mathbb{R}$:

$$\Delta = Y_1 - Y_0$$

Our data come from two populations, indexed by $D \in \{e, a\}$. e is the population in which the experimental evaluation of T was conducted and a is the alternative population of interest. d-superscripts will index population-specific distributions and their attributes. In population e, the experimental evaluation assigns T at random independently of all other random variables with perfect compliance[6]. Therefore, we can identify the marginal distribution of untreated outcomes in population e, $F^e_{Y_0}(y_0)$, from the equality

$$F^e_{Y_0}(y_0) = F^e_{Y|T}(y|T = 0)$$

[6]Putting perfect compliance with treatment assignment another way, the estimands of interest will be intention-to-treat (ITT) effects, including any participation decisions. The ITT is often thought to be the object of policy interest since compliance can rarely be mandated in policy settings.

where $F^e_{Y|T}(y|T=0)$ denotes the marginal distribution of Y conditional on the treatment indicator being equal to zero. The equality follows from the independence of the treatment indicator from the potential outcomes. We can also identify the marginal distribution of treated outcomes:

$$F^e_{Y_1}(y_1) = F^e_{Y|T}(y|T=1).$$

We can additionally identify any functionals of the outcome distributions, which allows us to identify the average individual-specific treatment effect Δ in population e:

$$\begin{aligned} E^e[\Delta] &= E^e[Y_1 - Y_0] \\ &= E^e[Y_1] - E^e[Y_0] \\ &= E^e[Y_1|T=1] - E^e[Y_0|T=0] = E^e[Y|T=1] - E^e[Y|T=0]. \end{aligned}$$

E^d stands for the expectation with respect to the distribution in $D = d$.

As in previous sections, I maintain the assumption that all members of the alternative population are untreated for concreteness. So $T = 0$ for all individuals in population a. This means that in population a, we identify that distribution of untreated outcomes:

$$F^a_{Y_0}(y_0) = F^a_{Y|T}(y|T=0) = F^a_Y(y).$$

We are, however, interested in the average treatment effect in alternative population, $E^a[\Delta]$, which depends on our ability to identify $E^a[Y_1]$:

$$\begin{aligned} E^a[\Delta] &= E^a[Y_1 - Y_0] \\ &= E^a[Y|T=1] - E^a[Y|T=0] \\ &= \underbrace{E^a[Y_1]}_{unknown} - E^a[Y]. \end{aligned}$$

If the treatment effect were constant for all individuals and equal to $\overline{\Delta}$, $E^a[\Delta]$ would simply be equal to $E^e[\Delta]$. However, theory rarely implies a constant treatment effect and we can often reject it empirically, see e.g. Heckman et al. (1997); Djebbari and Smith (2008). In fact, theory usually predicts heterogeneity in treatment response depending on the individual and her context's observed and unobserved attributes.

To demonstrate the role of heterogeneity in observed and unobserved characteristics on the average treatment effect in a, I now introduce some additional notation. Suppose we observe a vector of covariates $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ for each individual. Additionally, suppose

there is a vector of unobserved covariates $U \in \mathcal{U} \subseteq \mathbb{R}^{d_U}$ that we believe affects the outcome. Concretely, we can think of the observed covariates in the remedial education example from the introduction: the student's grade level competency when entering third grade, class size and gender. The unobserved covariates might be her latent ability and any parental inputs. An equivalent representation for the potential outcomes is that treatment status and covariates combine to produce the outcome through a function common across populations, $g : \{0, 1\} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. In this representation, the potential outcomes are:

$$\begin{aligned} Y_0 =& g(0, X, U) \\ Y_1 =& g(1, X, U). \end{aligned}$$

The individual-specific treatment effect is

$$\Delta = Y_1 - Y_0 = g(1, X, U) - g(0, X, U),$$

which will in general depend on both X and U. Our target, $E^a[\Delta]$ can be written as:

$$\begin{aligned} ATE^a =& E^a[Y_1 - Y_0] \\ =& \int_{\mathcal{X}\times\mathcal{U}} g(1, x, u) - g(0, x, u) dF^a_{X,U}(x, u) \end{aligned}$$

where $F^a_{X,U}(x, u)$ denotes the joint distribution of observed and unobserved covariates in population a. Note that $F^a_{X,U}(x, u)$ in general differs from $F^e_{X,U}(x, u)$. Iterating expectations, ATE^a can be written in three equivalent ways:

$$ATE^a = \int_{\mathcal{X}} \left[\int_{\mathcal{U}} g(1, x, u) - g(0, x, u) dF^a_{U|X}(u|x) \right] dF^a_X(x) \tag{1}$$

$$= \int_{\mathcal{X}} \left[\int_{\mathbb{R}^2} (y_1 - y_0)\, dF^a_{Y_0,Y_1|X}(y_0, y_1|x) \right] dF^a_X(x) \tag{2}$$

$$\int_{\mathcal{X}} \left[\int_{\mathbb{R}} \delta dF^a_{\Delta|X}(\delta|x) \right] dF^a_X(x) \tag{3}$$

Equations (1) and (2) show that ATE^a depends on the distribution of $Y_0, Y_1|X, D = a$, which itself depends on the distribution of $U|X, D = a$. Equation (3) makes the connection to the distribution of treatment effects for individuals with a particular value of the observed covariates. Note that the equivalence of equations (1) and (2) shows that the invariance to the population indicator of the function generating outcomes is without loss of generality, since the dimension of U is unrestricted and could include a separate indicator for each population, analogous to defining the d-index of $F^d_{Y_1,Y_0,X}(y_1, y_0|x)$ as an element of U.

3.1 Previous methods

Within this general setup, I now describe previous methods for using the distributions from the experimental population to identify the average treatment effect in the alternative population.

3.1.1 Conditional independence of the gains

The standard approach to extrapolating the results of social experiments has been to reweight the average treatment effects conditional on each value of the observed covariates by the distribution of observed covariates in the population of interest. That is:

$$ATE^a = \int_{\mathcal{X}} E^e[Y_1 - Y_0|x]dF_X^a(x). \tag{4}$$

This estimator is justified on the basis of the following assumptions (Allcott (2015)):

$$\mathcal{X}^a \subseteq \mathcal{X}^e \tag{5}$$

$$\Delta \perp\!\!\!\perp D|X \tag{6}$$

where $\mathcal{X}^a$ denotes the support of X in the alternative population, $\mathcal{X}^e$ denotes the support in the experimental population and $\perp\!\!\!\perp$ denotes statistical independence[7]. (5) is a standard condition required for non-parametric extrapolation. (6) is the key identification assumption. Note that under (6), $\Delta = Y_1 - Y_0$ is independent of any difference between the conditional distributions of untreated outcomes, $F^a_{Y_0}(y_0|x)$ and $F^e_{Y_0}(y_0|x)$. With a bounded outcome, the conditional distributions of control outcomes may be such that (6) is impossible. For one extreme example, consider the case where $Y \in \{0, 1\}$, and the outcomes in population in population e are as in section 2, with $E^e[Y_0] = \frac{1}{3}$ and $E^e[Y_1] = \frac{2}{3}$. $E^e[\Delta] = \frac{1}{3}$. If $Y = 1$ for all individuals in population a, (6) cannot hold. Predictions will also depend on the scaling of Y, for example, whether it is measured in levels or logs[8].

Even more substantively, differences in the conditional distributions of control outcomes are indicative of some unobserved differences between the experimental population and the

[7]The estimator in equation (4) can be justified on the basis of a weaker mean-independence assumption, but I will focus on the assumptions considered in the literature.

[8](4) is analogous to the counterfactual portion of a difference-in-differences estimator, where the assumption is that the mean difference in outcomes is conditionally independent of the population indicator. Hence, these standard criticisms of difference-in-differences estimators as non-invariant to scaling of the outcome and possibly delivering predictions outside the support of the outcome variable, as described in Athey and Imbens (2006) for example, apply here as well.

population of interest. To see this, note that:

$$F^d_{Y_0|X}(y_0|x) = F^d_{g(0,x,U)}(g(0,x,U)).$$

Then

$$F^a_{Y_0|X}(y_0|x) \neq F^e_{Y_0|X}(y_0|x) \implies F^a_{U|X}(u|x) \neq F^e_{U|X}(u|x).$$

If the elements of U whose difference in conditional distribution produce the difference in the conditional distribution of control outcomes also influence the individual-specific treatment effect, (6) will not hold.

3.1.2 Conditional independence of the potential outcomes

Due to some combination of these criticisms, the primary assumption used in the theoretical literature on extrapolation of experimental results combines (5) with the assumption that the joint distribution of potential outcomes is independent of the population conditional on the observed covariates:

$$(Y_0, Y_1) \perp\!\!\!\perp D|X \tag{7}$$

or equivalently, that all unobserved covariates determining the outcome are independent of the population indicator:

$$U \perp\!\!\!\perp D|X$$

It is straightforward to show that (7) implies $E^a[Y_1|x] = E^e[Y_1|x]$ so that we can identify the average treatment effect in the population of interest by reweighting the expectation of the treated outcome from the experimental population conditional on covariates by the distribution of covariates in the population of interest and subtracting the expected control outcome from the population of interest:

$$ATE^a = \int_{\mathcal{X}} E^e[Y_1|x]dF^a_X(x) - E^a[Y_0].$$

For (7) to hold, the conditional distributions of control outcomes must be the same in the two populations. Therefore Hotz et al. (2005) and papers following them have suggested testing equality of the distributions or their moments. Two issues come up when testing $F^e_{Y_0|X}(y_0|x) = F^a_{Y_0|X}(y_0|x)$ and using the result to conclude whether or not we can generalize

results from the experiment to the population of interest. First, considering the small sample sizes of many social experiments, we may often be underpowered to reject equality of the conditional outcome distributions, as raised in Flores and Mitnik (2013). Second, if we do reject the null hypothesis, we must conclude that the experiment tells us nothing about ATE^a. Again, this may be an issue of sample size: with large samples from both the experimental population and the population of interest we will in all likelihood reject the null. Furthermore, there is an issue of degree. Suppose we have two alternative populations of interest a and a' and our samples are large enough to reject both $F^e_{Y_0|X}(y_0|x) = F^a_{Y_0|X}(y_0|x)$ and $F^e_{Y_0|X}(y_0|x) = F^{a'}_{Y_0|X}(y_0|x)$ but $F^a_{Y_0|X}(y_0|x)$ is quite similar to $F^e_{Y_0|X}(y_0|x)$ while $F^{a'}_{Y_0|X}(y_0|x)$ is quite different, it seems inappropriate to conclude that the results from e are equally (and completely) uninformative in predicting the average causal effect in both a and a'. In the following section, I depart from the testing framework and derive bounds on the average causal effect in the population of interest as a function of the differences in the conditional distributions of control outcomes between the population of interest and the experimental population. I conclude this section with a simple example.

3.2 Example: remedial education in India

To make the above discussion concrete, I now describe a simple parametric model using the example of remedial education India. Suppose students from the city of Mumbai represent the experimental population, e, and students from the city of Vadodara the alternative population, a, where we would like to predict the average treatment effect. We will leave the observed covariates X as a vector, but break the vector U into the two components discussed above, latent skill S and parental input I. $g(\cdot)$ is a linear production function with different parameters depending on treatment status

$$g(0, X, S, I) = \beta_0 + \beta'_{0X}X + \beta_{0S}S + \beta_{0I}I = Y_0$$
$$g(1, X, S, I) = \beta_1 + \beta'_{1X}X + \beta_{1S}S + \beta_{1I}I = Y_1$$

Note that once we assume linearity, the commonality of $g(\cdot)$ across populations is no longer without loss of generality. In this case, the individual-specific treatment effect, Δ, is

$$\begin{aligned}\Delta =& Y_1 - Y_0 \\ =& (\beta_1 - \beta_0) \\ &+ (\beta'_{1X} - \beta'_{0X})X \\ &+ (\beta_{1S} - \beta_{0S})S \\ &+ (\beta_{1I} - \beta_{0I})I\end{aligned}$$

Our objective is to identify:

$$\begin{aligned}ATE^a =& E^a[Y_1 - Y_0] \\ =& (\beta_1 - \beta_0) \\ &+ E^a\left[(\beta'_{1X} - \beta'_{0X})X\right] \\ &+ E^a\left[(\beta_{1S} - \beta_{0S})S\right] \\ &+ E^a\left[(\beta_{1I} - \beta_{0I})I\right]\end{aligned}$$

The four elements of ATE^a are, respectively, a treatment effect common to all students, the average deviation from the common treatment effect due to observables in population a, the average deviation from the common effect due to latent skill in population a and the average deviation from the common effect due to the parental input. When $\beta'_{1X} \neq \beta'_{0X}$, there is treatment effect heterogeneity due to observable covariates and when $\beta_{1S} \neq \beta_{0S}$ or $\beta_{1I} \neq \beta_{0I}$ there is treatment effect heterogeneity due to unobservables.

ATE^e alone will in general be biased as an estimator for ATE_a, with the bias taking the following form:

$$\begin{aligned}ATE^e - ATE^a =& (\beta'_{1X} - \beta'_{0X})(E^e[X] - E^a[X]) \\ &+ (\beta_{1S} - \beta_{0S})(E^e[S] - E^a[S]) \\ &+ (\beta_{1I} - \beta_{0I})(E^e[I] - E^a[I])\end{aligned}$$

The bias depends on the differences between sites in the marginal distributions of characteristics along which treatment effects are heterogeneous.

In this simple example, we need $E^a[S|x] = E^e[S|x]$ if $\beta_{1S} \neq \beta_{0S}$ and $E^a[I|x] = E^e[I|x]$ if $\beta_{1I} \neq \beta_{0I}$ for conditional independence of the gains, (6), to hold. We need $E^a[S|x] = E^e[S|x]$ if $(\beta_{0S}, \beta_{1S}) \neq (0, 0)$ and $E^a[I|x] = E^e[I|x]$ if $(\beta_{0I}, \beta_{1I}) \neq (0, 0)$ for conditional independence

of the potential outcomes, (7), to hold. We will return to this parametric model to build intuition for key points in the next section as well.

4 Bounds on ATE^a using differences in the untreated outcome distributions

4.1 Identification

In investigating the role of the conditional untreated outcome distributions in determining the average causal effect in the population of interest, recall first that since we can already identify $E^a[Y_0]$ (simply the expected outcome in the population of interest), what we need to identify $E^a[Y_1] - E^a[Y_0]$ is the counterfactual $E^a[Y_1]$. The expected value of the treated outcome in the population of interest can be written as follows:

$$E^a[Y_1] = \int_{\mathcal{X}} \left(\int_{\mathbb{R}} \left[\int_{\mathbb{R}} y_1 \underbrace{dF^a_{Y_1|Y_0,X}(y_1|y_0,x)}_{unidentified} \right] \underbrace{dF^a_{Y_0|X}(y_0|x)}_{identified} \right) \underbrace{dF^a_X(x)}_{identified} \tag{8}$$

We are missing information on the distribution of treated outcomes that individuals with a particular untreated outcome would experience in the population of interest. Since no one is treated in the population of interest, for information on this object, we must turn to the experimental population.

For the experiment to tell us anything about $F^a_{Y_1|Y_0,X}(y_1|y_0,x)$, we must first impose two support conditions.

Assumption 1. *The support of X in the population of interest is a subset of the support in the experimental population: $\mathcal{X}^a \subseteq \mathcal{X}^e$.*

Assumption 2. *The support of $Y_0|X = x$ in the population of interest is a subset of the support in the experimental population for all values of X in the support of X in the population of interest: $Supp^a(Y_0|X = x) \subseteq Supp^e(Y_0|X = x)\ \forall x \in \mathcal{X}^a$.*

Assumption 1 is the same as employed in the previous literature (see equation (5)). Assumption 2 will be needed to nonparametrically tie differences in the conditional distributions of untreated outcomes to differences in the conditional distributions of treated outcomes. I will explore alternative assumptions when these are violated in an extension.

Turning now to the question of identification of $F^a_{Y_1|Y_0,X}(y_1|y_0,x)$ using information from the experiment, we first observe that there are many possible covariate-and-untreated-outcome-conditional distributions $F_{Y_1|Y_0,X}(y_1|y_0,x)$ associated with the covariate-conditioned

marginal untreated outcome $F^e_{Y_0|X}(y_0|x)$ and treated outcome distributions $F^e_{Y_1|X}(y_1|x)$. Specifically, $F_{Y_1|Y_0,X}(y_1|y_0,x)$ is a valid conditional distribution for the marginal distributions $F^e_{Y_0,X}(y_0|x)$ and $F^e_{Y_1|X}(y_1|x)$ if

$$F_{Y_1|Y_0,X}(y_1|y_0,x) = C_1(F^e_{Y_0,X}(y_0|x), F^e_{Y_1|X}(y_1|x)|x)$$

where $C : [0,1]^2 \to [0,1]$ is a copula function (see appendix A for the definition), and $C_1(v,w|x) = \frac{\partial C(v,w|x)}{\partial v}$. Informally, a copula function is a bivariate CDF where both arguments are defined on the unit interval which fully determines a dependence structure between the untreated and treated outcomes in the experimental population for individuals with the same covariates. A copula function combined with the marginal distributions of untreated ($F^e_{Y_0,X}(y_0|x)$) and treated outcomes ($F^e_{Y_1|X}(y_1|x)$) defines a joint distribution ($F_{Y_0,Y_1|X}(y_0,y_1|x)$) consistent with those marginal distributions. $F_{Y_1|Y_0,X}(y_1|y_0,x)$ is the conditional distribution associated with the joint distribution $F_{Y_0,Y_1|X}(y_0,y_1|x)$. Let $\mathcal{C}$ denote the set of valid copula functions.

I will assume that the distribution of treated outcomes conditional on an untreated outcome and observed covariates in the alternative population of interest is consistent with the experimental results.

Assumption 3. *Consistency of the conditional distribution of treated outcomes in the population of interest with the experimental results:*

$$F^a_{Y_1|Y_0,X}(y_1|y_0,x) = C_1(F^e_{Y_0|X}(y_0|x), F^e_{Y_1|X}(y_1|x)|x)$$

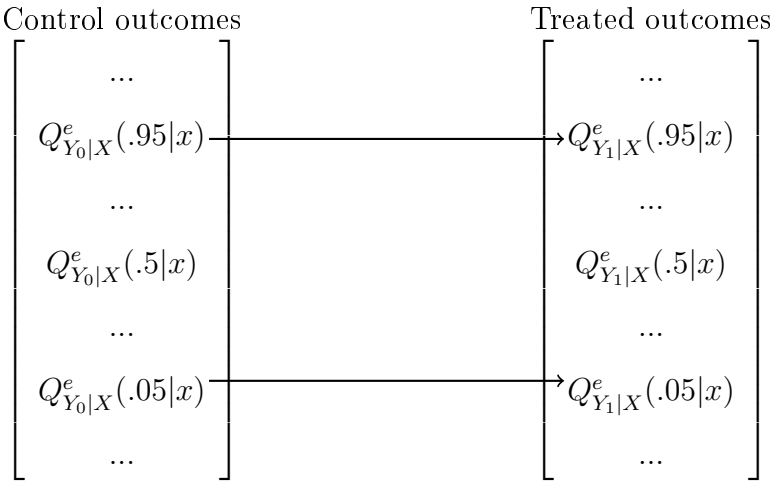
for some copula function $C \in \mathcal{C}$.

Assumption 3 states that we must be able to express the distribution of the treated outcome conditional on an untreated outcome and covariates as one of the conditional distributions consistent with the distributions of untreated and treated outcomes in the experiment.

To make Assumption 3 more concrete, I illustrate two examples of copula functions and show how they define a joint distribution of potential outcomes $F_{Y_0,Y_1|X}(y_0,y_1|x)$. Let $Q^e_{Y_0|X}(\alpha|x)$ denote the α-quantile of $Y_0|X$ in the experimental population and $Q^e_{Y_1|X}(\alpha|x)$ the α-quantile of $Y_1|X$ in the experimental population. Figures 2 and 3 show two possible copulas and the joint distributions they define. The arrows in the figures represent dependence relationships between $F^e_{Y_0|X}(y_0|x)$ and $F^e_{Y_1|X}(y_1|x)$ defined by the copulas. The horizontal arrows in figure 2 represent the joint distribution $Y_0, Y_1|X$ in the experimental population when the treatment preserves individuals' ranks in the outcome distributions perfectly. In the example of remedial education in India, the highest-scoring student without a remedial

education teacher assigned to her school would still be the highest-scoring student with a remedial education teacher assigned. The crossing arrows in figure 3 represent the case when the treatment reverses ranks: the highest scoring student without the treatment would be the lowest-scoring student without the treatment.

Figure 2: Perfect positive dependence of $F^e_{Y_0|X}(y_0|x)$, $F^e_{Y_1|X}(y_1|x)$



A joint distribution $F^e_{Y_0,Y_1|X}(y_0, y_1|x)$ consistent with the experimental marginal distributions of control and treated outcomes also determines the extent of heterogeneity in treatment effects for individuals with covariates x. When the treatment perfectly preserves individuals' ranks in the outcome distributions, treatment effect heterogeneity due to unobservables is minimized (Cambanis, Simons, and Stout (1976)). That is, conditional on x, the individual-specific treatment effects Δ have the the smallest magnitude possible. In contrast, when the treatment inverts individuals' ranks in the outcome distributions, the treatment effects have the largest possible magnitude.

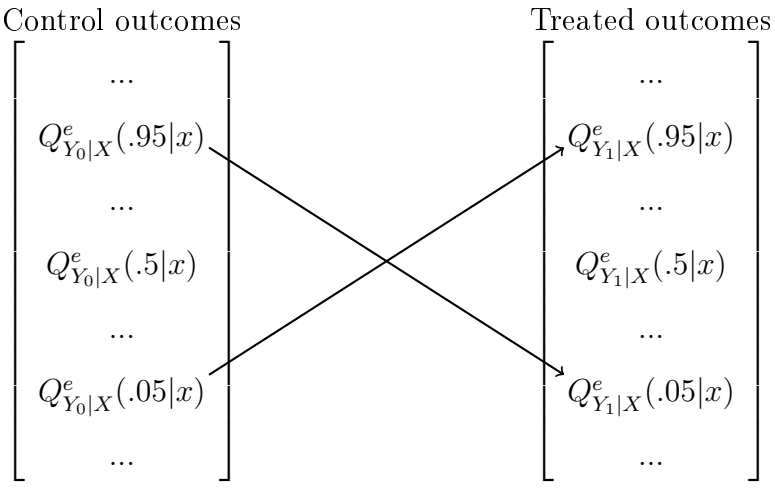
The relationship between $Y_0|X$ and $Y_1|X$ under perfect positive dependence is known as comonotonicity, which is defined as follows.

Definition 1. Comonotonicity. When two random variables V and W are comonotonic

$$F_{V,W}(v, w) = \min\{F_V(v), F_W(w)\}.$$

A necessary condition for Assumption 3 is that if the control outcomes conditional on a value of the covariates have the same distribution in the experimental population and the

Figure 3: Perfect negative dependence of $F^e_{Y_0|X}(y_0|x)$, $F^e_{Y_1|X}(y_1|x)$



population of interest, the conditional treated outcomes have the same distribution as well. That is,

$$F^a_{Y_0|X}(y_0|x) = F^e_{Y_0|X}(y_0|x) \implies F^a_{Y_1|X}(y_1|x) = F^e_{Y_1|X}(y_1|x).$$

A sufficient condition but stronger than necessary condition is that the distribution of the treated outcomes be the same across populations once we have conditioned on a value of the control outcome and the observed covariates, an assumption also used in Athey and Imbens (2006). Formally:

$$Y_1 \perp\!\!\!\perp D|Y_0, X \tag{9}$$

This is the relevant condition to answer the hypothetical, what would the conditional distribution of treated outcomes have been in the experiment had the distribution of control outcomes been the same as in the population of interest (see Fortin, Lemieux, and Firpo (2011))? In terms of the underlying unobservables, a sufficient condition for (9), in turn, is:

$$U \perp\!\!\!\perp D|g(0, x, U) = y_0, X = x.$$

Finally, we require existence of the expectation of Y_1 in e.

Assumption 4. *Y_1 has finite expectation in e: $E^e[|Y_1|] < \infty$.*

Combining assumptions 1, 2, 3 and 4, we state the following result.

Proposition 1. *Under assumptions 1, 2, 3 and 4:*

$$E^a[Y_1 - Y_0|x] \in \left[\left\{ \min_{C \in \mathcal{C}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y_0}(y_0|x), F^e_{Y_1}(y_1|x)|x) \right) dF^a_{Y_0}(y_0|x) \right\} - E^a[Y_0|x], \right.$$
$$\left. \left\{ \max_{C \in \mathcal{C}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y_0}(y_0|x), F^e_{Y_1}(y_1|x)|x) \right) dF^a_{Y_0}(y_0|x) \right\} - E^a[Y_0|x] \right]$$

Bounds on the unconditional average treatment effect in the population of interest can then be recovered by weighting the minimal and maximal conditional average treatment effects by the distribution of covariates in the population of interest.

$$ATE^a \in \left[\int_{\mathcal{X}} \min \ E^a[Y_1 - Y_0|x] dF^a_X(x), \int_{\mathcal{X}} \max \ E^a[Y_1 - Y_0|x] dF^a_X(x) \right] \tag{10}$$

All of the objects in proposition 1 are identified, with the exception of the copula C. We minimize and maximize over the set of possible copulas $\mathcal{C}$ to obtain the bounds. The bounds defined in proposition 1 are sharp by construction, since each element of $\mathcal{C}$ defines a valid possible conditional distribution $F^a_{Y_1|Y_0,X}(y_1|y_0, x)$.

By considering the full set of possible copulas, we consider copulas that may not be credible, however. In particular, the dependence structure shown in figure 3 is not realistic in most applications. In the remedial education example, it is clearly unrealistic to believe that the highest-performing students when no remedial education teacher is assigned to their school become the lowest-performing when a remedial education teacher is assigned. Unless remedial education is so effective that a poor-performing student without treatment becomes the best-performing student, the best-performing student without treatment's rank in the outcomes distribution is likely unaffected: she is not assigned to work with the remedial education teacher and remains the highest-performing. We typically anticipate some positive dependence between outcomes with and without treatment for any one individual, with the degree of dependence (and thus of unobserved treatment effect heterogeneity) depending on the application.

We therefore index copulas by the degree of dependence in the joint distributions of control and treated outcomes they generate. We use Normalized Spearman's ρ, defined below, to measure dependence.

Definition 2. For any two random variables V and W, Normalized Spearman's ρ is given

by:

$$\rho(V,W) = \frac{Cor_C(R(V),R(W))}{Cor_M(R(V),R(W))}$$

where $R(V) = F_V(v)$ when V is continuously distributed and $R(V) = \frac{F_V(v)+F_V(v-)}{2}$ when V takes a finite number of values and equivalently for W. The notation $F_V(v-)$ denotes $P(V < v)$ and equivalently for W. $Cor_C(R(V),R(W))$ refers to the product-moment correlation between $R(V)$ and $R(W)$ under copula C: $\int \left(R(V)-\frac{1}{2}\right)\left(R(W)-\frac{1}{2}\right) dC(F_V(v),F_W(w))$. $Cor_M(R(V),R(W))$ is the product-moment correlation between $R(V)$ and $R(W)$ under comonotonicity: $\int \left(R(V)-\frac{1}{2}\right)\left(R(W)-\frac{1}{2}\right) d\left(\min\left\{F_V(V),F_W(w)\right\}\right)$.

The definition of Normalized Spearman's ρ coincides with the standard calculation of Spearman's ρ in the numerator (see Nešlehová (2007)). In the denominator, when V and W are continuously distributed, $\int \left(R(V)-\frac{1}{2}\right)\left(R(W)-\frac{1}{2}\right) d\left(\min\left\{F_V(V),F_W(w)\right\}\right)=1$ so that the calculation is completely standard. However, when V and W take a finite number of values, $\int \left(R(V)-\frac{1}{2}\right)\left(R(W)-\frac{1}{2}\right) d\left(\min\left\{F_V(V),F_W(w)\right\}\right)$ may be less than 1. So the only difference with the standard calculation is the normalization in the discrete case.

We can produce bounds on $E^a[Y_1 - Y_0|x]$ subject to the restriction that we only consider copula functions generating dependence greater than a specified level. This is represented in the following assumption and proposition.

Assumption 5. *C is an element of $\mathcal{C}(\rho^L)$, the set of copula functions such that $\rho(Y_0,Y_1|X=x) \geq \rho^L$ where $\rho^L \in [0,1]$.*

Proposition 2. *Under Assumptions 1, 2, 3, 4 and 5:*

$$E^a[Y_1 - Y_0|x] \in \left[\left\{\min_{C\in\mathcal{C}(\rho^L)} \int_{\mathbb{R}}\left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y_0}(y_0|x),F^e_{Y_1}(y_1|x)|x)\right) dF^a_{Y_0}(y_0|x)\right\} - E^a[Y_0|x],\right.$$
$$\left.\left\{\max_{C\in\mathcal{C}(\rho^L)} \int_{\mathbb{R}}\left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y_0}(y_0|x),F^e_{Y_1}(y_1|x)|x)\right) dF^a_{Y_0}(y_0|x)\right\} - E^a[Y_0|x]\right].$$

Bounds on the unconditional ATE^a can be computed in the same way as under proposition 1 (equation (10)). $\mathcal{C}(1)$ is a singleton and the bounds shrink to a point. We now investigate the structure underlying the potential outcomes as a means of interpreting the results and assumptions.

4.1.1 1-dimensional unobservables generate comonotonicity

Suppose an individual's control and treated potential outcomes, Y_0 and Y_1, are both generated by a single unobserved characteristic of the individual so that U is one-dimensional and the structural functions $g(0, x, u)$ and $g(1, x, u)$ are each weakly increasing in u. It is a standard result that this implies comonotonicity of the potential outcomes (see, for example, the proof of proposition 5.16 in McNeil, Frey, and Embrechts (2005)).

Athey and Imbens (2006) use this characterization of Y_t (however, in their difference-in-differences setting T indexes time, rather than treatment), along with assumptions 1, 2 and 3 and the condition $U \perp\!\!\!\perp T$ to yield an estimator they refer to as the changes-in-changes model with conditional independence (see section 4.2 of Athey and Imbens (2006)). $U \perp\!\!\!\perp T$ by design in the experiment (T is randomly assigned independently of any other random variable), so the changes-in-changes model with conditional independence is a valid estimator for the point defined under proposition 2 when $\rho^L = 1$. When outcomes are continuous, Athey and Imbens (2006) point out that assumption 3 is implied by monotonicity in u of the function generating outcomes and thus does not need to be separately imposed.

Example. To gain some intuition for the identifying power of assuming $g(0, x, u)$ and $g(1, x, u)$ are strictly increasing in 1-dimensional u, we return to the parametric example introduced in section 3.2. Assume the parental input I is excluded from the production function so unobservables are one-dimensional[9] and the potential outcomes can be written as

$$Y_0 = \beta_0 + \beta_{0X} X + \beta_{0S} S$$
$$Y_1 = \beta_1 + \beta_{1X} X + \beta_{1S} S$$

In this section I illustrate that with a one-dimensional unobservable, the way in which the distributions of observables $F^e_{X,Y}(x, y)$ in the experimental population change with treatment

[9]This is not the only way to generate 1-dimensional unobservables in the linear production function described in section 3.2. We could make use of a single index specification for the unobservables where

$$Y_0 = \beta_0 + \beta_{0X} X + \beta_{0S} S + \beta_{0I} I$$
$$Y_1 = \beta_1 + \beta_{1X} X + \kappa(\beta_{0S} S + \beta_{0I} I)$$

Alternatively, if S and I have a Pearson product-moment correlation of 1, we can write I as a linear function of S ($I = bS$) so that:

$$Y_0 = \beta_0 + \beta_{0X} X + (\beta_{0S} + \beta_{0I} b) S$$
$$Y_1 = \beta_1 + \beta_{1X} X + (\beta_{1S} + \beta_{1I} b) S$$

status can be mapped into differences in the treatment and control structural functions. This knowledge of the changes in the structural function can be applied to differences in the distributions of observables in the control state, $F^e_{X,Y_0}(x, y_0)$ and $F^a_{X,Y_0}(x, y_0)$, across populations to recover $E^a[Y_1]$.

Let $\alpha = F^e_{Y_0|X}(y_0|x)$ for a given value of y_0. Consider the α quantiles of $Y_1|X$ and $Y_0|X$ in e:

$$Q^e_{Y_1|X}(\alpha|x) = \beta_1 + \beta'_{1X}x + \beta_{1S}Q^e_{S|X}(\alpha|x)$$
$$Q^e_{Y_0|X}(\alpha|x) = \beta_0 + \beta'_{0X}x + \beta_{0S}Q^e_{S|X}(\alpha|x)$$

Making use of the linear functional form, we can subtract the x-subgroup, t-specific mean from each quantile to remove the common and x-specific structural effects:

$$Q^e_{Y_1|X}(\alpha|x) - E^e[Y_1|x] = \beta_{1S}\left(Q^e_{S|X}(\alpha|x) - E^e[S|x]\right)$$
$$Q^e_{Y_0|X}(\alpha|x) - E^e[Y_0|x] = \beta_{0S}\left(Q^e_{S|X}(\alpha|x) - E^e[S|x]\right)$$

By dividing the e treatment group α-quantile-specific deviation from the x-subgroup specific mean from the corresponding α-quantile-specific deviation in the e control group, we obtain the ratio of the effects of the latent skill S in the treated and control states.

$$\begin{aligned}\frac{Q^e_{Y_1|X}(\alpha|x) - E^e[Y_1|x]}{Q^e_{Y_0|X}(\alpha|x) - E^e[Y_0|x]} &= \frac{\beta_{1S}\left(Q^e_{S|X}(\alpha|x) - E^e[S|x]\right)}{\beta_{0S}\left(Q^e_{S|X}(\alpha|x) - E^e[S|x]\right)} \\ &= \frac{\beta_{1S}}{\beta_{0S}} \qquad (11)\end{aligned}$$

Knowing the ratio of the effects of latent math skill across treatment and control states allows us to map differences in the distributions of latent skill and pre-test score $F^e_{X,S}(x, s)$ and $F^a_{X,S}(x, s)$ identified by differences in the joint distributions of the control outcomes $F^e_{X,Y_0}(x, y_0)$ and $F^a_{X,Y_0}(x, y_0)$ into differences in the observed treatment group distribution in e, $F^e_{X,Y_1}(x, y_1)$, and the unknown treated group distribution in a, $F^a_{X,Y_1}(x, y_1)$. Specifically, consider:

$$E^a[Y_0|x] - E^e[Y_0|x] = \beta_{0S}\left(E^a[S|x] - E^e[S|x]\right).$$

Then we can use the change in the effect of unobservables from equation (11) to identify the

unknown expected value of the treated outcome conditional on covariates x.

$$E^a[Y_1|x] - E^e[Y_1|x] = \frac{\beta_{1S}}{\beta_{0S}}\left(E^a[Y_0|x] - E^e[Y_0|x]\right)$$
$$E^a[Y_1|x] = \frac{\beta_{1S}}{\beta_{0S}}\left(E^a[Y_0|x] - E^e[Y_0|x]\right) + E^e[Y_1|x]$$

Finally, the conditional average treatment effect is obtained by subtracting the conditional expectation of the test score in the population of interest.

$$E^a[Y_1 - Y_0|x] = \frac{\beta_{1S}}{\beta_{0S}}\left(E^a[Y_0|x] - E^e[Y_0|x]\right) + E^e[Y_1|x] - E^a[Y|x]$$

4.1.2 Multidimensional heterogeneity

However, when we introduce multidimensional heterogeneity, we can no longer cleanly apply the knowledge we gain from the experiment about how the structural function $g(t, x, u)$ changes with treatment to the differences in $F^e_{X,Y_0}(x, y_0)$ and $F^a_{X,Y_0}(x, y_0)$.

Example. This is easy to see in the parametric illustration when we reintroduce independent variation in I. Consider the treatment-to-control ratio of α-quantile deviations from the x-specific subgroup means in the experimental population:

$$\frac{Q^e_{Y_1|X}(\alpha|x) - E^e[Y_1|x]}{Q^e_{Y_0|X}(\alpha|x) - E^e[Y_0|x]} = \frac{Q^e_{\beta_{1S}S+\beta_{1I}I}(\alpha|x) - E^e[\beta_{1S}S + \beta_{1I}I|x]}{Q^e_{\beta_{0S}S+\beta_{0I}I}(\alpha|x) - E^e[\beta_{0S}S + \beta_{0I}I|x]}$$

Whereas previously this ratio simplified to the treatment-to-control ratio of effects of latent skill on the test score at the end of third grade, it no longer identifies any specific change in the structural function. Put more generally, the α-quantile of $Y_t|x$ in the experimental population now provides no structural information.

We will see in the next section that for very small deviations from 1-dimensional unobserved heterogeneity, the bounds on the average treatment effect in the population of interest expand substantially, depending on the extent of difference in the conditional distributions of the control outcomes between the population of interest and the experimental population. Only when unobserved heterogeneity is *exactly*, and not approximately, 1-dimensional do differences in the conditional distributions of the control outcomes not lead to a loss in identification. This motivates considering the bounds from proposition 2 and investigating how they change with ρ^L.

4.2 Estimation

In estimation, I will consider the case when outcomes and covariates are discrete or discretized. I will illustrate both possibilities in the empirical work. When outcomes and covariates are discrete, we can represent the optimization over the restricted space of copulas $\mathcal{C}(\rho^L)$ as a linear programming problem. In particular, the bounds on the average causal effect in context a for individuals with covariates x admit a representation as the solution to a discrete optimal transportation problem with a non-standard cost function and an additional linear constraint on dependence (see Villani (2009) for a comprehensive discussion of optimal transportation problems). Very efficient algorithms are available to solve linear programs (see e.g. Boyd and Vandenberghe (2004)), so the bounds can be computed quickly using software provided by the author.

A similar representation as a continuous optimal transportation problem exists when outcomes are continuous, but there is no analogous tractable method to compute the solution, which involves optimization over an infinite-dimensional space ($\mathcal{C}(\rho^L)$). It may be possible to represent $\mathcal{C}(\rho^L)$ with a sieve space $\mathcal{C}_n(\rho^L)$, which would be finite-dimensional and compact, becoming dense as $n \to \infty$. Exploring this possibility is left to future research. I therefore impose the following assumption on outcomes and covariates.

Assumption 6. Finite support of the potential outcomes and covariates. Let $J, K \in \mathbb{N}$. Y_0 and Y_1 take values in $\mathcal{Y}_0 = \{y_{0,1}, \ldots, y_{0,j}, \ldots, y_{0J}\}$ and $\mathcal{Y}_1 = \{y_{1,1}, \ldots, y_{1,k}, \ldots, y_{1K}\}$, respectively. Further, X takes values in a finite set $\mathcal{X}$.

4.2.1 Linear programming representation

I first describe the linear programming representation of the bounds in Proposition 2. I leave conditioning on x implicit to economize on notation. Given ρ^L, the upper bound is obtained by solving the following linear programming problem with solution $\tau^U(\rho^L)$ (the lower bound, $\tau^L(\rho^L)$ is obtained by replacing the max operator with min).

$$\tau^U(\rho^L) = \max_{\mathcal{C}(\rho^L)} E^a[Y_1 - Y_0]$$

$$= \max_{\{P^e(y_{0j},y_{1k})\}_{j=1,...,J}^{k=1,...,K}} \sum_{j=1}^{J}\sum_{k=1}^{K} y_{1k}\frac{P^a(y_{0j})}{P^e(y_{0j})} \times P^e(y_{0j}, y_{1k}) \tag{12}$$

$$- \sum_{j=1}^{J} y_{0j} P^a(y_{0j}) \tag{13}$$

subject to

$$\sum_{k=1}^{K} P^e(y_{0j}, y_{1k}) = P^e(y_{0j}) \ \forall j \in \{1, ..., J\} \tag{14}$$

$$\sum_{j=1}^{J} P^e(y_{0j}, y_{1k}) = P^e(y_{1k}) \ \forall k \in \{1, ..., K\} \tag{15}$$

$$\sum_{j=1}^{J}\sum_{k=1}^{K} \left(R(y_{0j}) - \frac{1}{2}\right)\left(R(y_{1k}) - \frac{1}{2}\right) P^e(y_{0j}, y_{1k})$$

$$\geq \rho^L \left[\max_{\{P^e(y_{0j},y_{1k})\}_{j=1,...,J}^{k=1,...,K}} \sum_{j=1}^{J}\sum_{k=1}^{K} \left(R(y_{0j}) - \frac{1}{2}\right)\left(R(y_{1k}) - \frac{1}{2}\right) P^e(y_{0j}, y_{1k})\right] \tag{16}$$

$$P^e(y_{0j}, y_{1k}) \geq 0 \ \forall j \in \{1, ..., J\}, k \in \{1, ..., K\}$$

Maximization is with respect to the elements of the matrix defining the joint distribution of Y_0 and Y_1 in population e, $\{P^e(y_{0j}, y_{1k})\}_{j=1,...,J}^{k=1,...,K}$. Line (13) is simply a normalization so that the value of the objective function of the problem can be interpreted as $E^a[Y_1 - Y_0]$. Constraints (14) and (15) require that the minimizing/maximizing joint distribution be consistent with the marginal outcome distributions in e. Constraint (16) enforces that Normalized Spearman's ρ (see Definition 2) applied to the potential outcomes Y_0 and Y_1, $\rho(Y_0, Y_1)$, may not be below ρ^L. Constraints (14), (15) and (16) make maximizing over the elements of the joint distribution of Y_0 and Y_1 equivalent to maximizing over the restricted space of copulas, $\mathcal{C}(\rho^L)$ (proof in Appendix B).

The coefficients on the elements of $\{P^e(y_{0j}, y_{1k})\}_{j=1,...,J}^{k=1,...,K}$ are $\left\{y_{1k}\frac{P^a(y_{0j})}{P^e(y_{0j})}\right\}_{j=1,...,J}^{k=1,...,K}$. Together with constraint (15), this shows the role of the distributions of control outcomes $\{P^a(y_{0j})\}_{j=1,...,J}$ and $\{P^e(y_{0j})\}_{j=1,...,J}$ in determining the bounds. If $P^a(y_0) = P^e(y_0)$, $\frac{P^a(y_0)}{P^e(y_0)} = 1$ and constraint (15) implies that the counterfactual $E^a[Y_1] = E^e[Y_1]$ [10]. All

[10]

else equal, in order to maximize the objective function, we would like to assign higher probability to high values on the support of Y_1 (high k) when $\frac{P^a(y_{0j})}{P^e(y_{0j})}$ is large and to low values on the support of Y_1 (low k) when $\frac{P^a(y_{0j})}{P^e(y_{0j})}$ is small. However, constraint (16) limits our ability to do so.

Example. Table 5 shows the choice variables and constraints 14 and 15 in the context of the remedial education in India example where the city of Mumbai is treated as e and Vadodara as a. As will be discussed in more detail in section 6, I do not use the test score directly as an outcome, but rather the discrete grade level competency of third graders when completing third grade. In table 5, I condition on a competency level of zero on entering third grade. The row and column labeled "All" represents the constraints on the marginal distributions $P^e(y_0|x)$ and $P^e(y_1|x)$. Without further constraints, the values of the choice variables are restricted only by the requirement that the sums across rows (for the untreated outcomes) equal the probability in the column labeled "All control" and that the sums down the columns (for the treated outcomes) equal the probability in the row labeled "All treated."

Table 5: Choice variables - $P^e(y_{0j}, y_{1k}|\text{competency on entering third grade} = 0)$, e =Mumbai

			Remedial education				
			Competency on exiting grade 3				
			0	1	2	3	All control
No remedial ed	Competency	0	$P^e(0,0)$	$P^e(0,1)$	$P^e(0,2)$	$P^e(0,3)$	0.73
		1	$P^e(1,0)$	$P^e(1,1)$	$P^e(1,2)$	$P^e(1,3)$	0.17
		2	$P^e(2,0)$	$P^e(2,1)$	$P^e(2,2)$	$P^e(2,3)$	0.07
		3	$P^e(3,0)$	$P^e(3,1)$	$P^e(3,2)$	$P^e(3,3)$	0.03
		All treated	0.66	0.2	0.1	0.04	

Proof.

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{k=1}^{K} y_{1k} P^e(y_{0j}, y_{1k}) \\
&= \sum_{j=1}^{J} y_{1k} \sum_{k=1}^{K} P^e(y_{0j}, y_{1k}) \\
&= \sum_{j=1}^{J} y_{1k} P^e(y_{1k}) = E^e[Y_1]
\end{aligned}$$

where the second equality follows from substituting in constraint (15). □

Table 6 shows the coefficient on each choice variable $P^e(y_{0j}, y_{1k})$ when Mumbai is treated as e and we condition on students' grade-level competency being zero on entering third grade. We can see that the differences in the distributions of control outcomes mean that we would maximize the objective function by ascribing the highest treatment effects to individuals with $Y_0 = 1$ and the lowest treatment effects to individuals with $Y_0 = 3$.

Constraint (16) on the dependence between Y_0 and Y_1 in Mumbai limits our ability to do so arbitrarily. Recall that ρ^L governs the allowed deviations from 1-dimensional heterogeneity. To gain some intuition for the joint distributions implied by different values of ρ^L, table 7 shows the joint distribution implied by $\rho^L = 1$ when Mumbai is treated as e and we condition on students' grade-level competency being zero on entering third grade. When $\rho^L = 1$, the 1-dimensional heterogeneity case, the majority of the mass in the joint distribution lies on the principal diagonal. Most individuals (88%) have a treatment effect of zero, with a few individuals experiencing a positive treatment effect of at most 1 competency level.

Table 6: Contribution of choice variables to the objective - $P^e(y_{0j}, y_{1k}|\text{competency on entering third grade} = 0)$, e =Mumbai

		Remedial education: Competency on exiting grade 3			
		0	1	2	3
No remedial ed: Competency	0	0	0.71	2×0.71	3×0.71
	1	0	2.26	2×2.26	3×2.26
	2	0	1.16	2×1.16	3×1.16
	3	0	0.60	2×0.60	3×0.60

4.2.2 Sample counterparts estimator

The solutions to the linear programming representation conditional on observed covariates x and minimum rank correlation ρ^L, $\tau_x^L(\rho^L)$ and $\tau_x^U(\rho^L)$ when minimizing and maximizing respectively, are functions of the population objects $\{P^e(y_{0j}|X = x)\}_{j=1,\ldots,J}$, $\{P^e(y_{1k}|X = x)\}_{k=1,\ldots,K}$ and $\{P^a(y_{0j}|X = x)\}_{j=1,\ldots,J}$. We can write

$$\tau_x^L(\rho^L) = \phi^L\left(\{P^e(y_{0j}|X = x)\}_{j=1,\ldots,J}, \{P^e(y_{1k}|X = x)\}_{k=1,\ldots,K}, \{P^a(y_{0j}|X = x)\}_{j=1,\ldots,J}; \rho^L\right) \tag{17}$$

Table 7: $P^e(y_{0j}, y_{1k}|\text{competency on entering third grade} = 0)$, $\rho^L = 1$ e =Mumbai

		Remedial education: Competency on exiting grade 3				
		0	1	2	3	All Control
No remedial ed: Competency	0	0.66	0.07	0	0	0.73
	1	0	0.13	0.04	0	0.17
	2	0	0	0.06	0.01	0.07
	3	0	0	0	0.03	0.03
	All Treatment	0.66	0.2	0.1	0.04	

where $\phi^L : \Delta(\mathcal{Y}_0) \times \Delta(\mathcal{Y}_1) \times \Delta(\mathcal{Y}_0) \to \mathbb{R}$ and $\Delta(Z)$ denotes the unit simplex on an arbitrary finite set Z. We can similarly write

$$\tau_x^U(\rho^L) = \phi^U\left(\left\{P^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{P^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{P^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}; \rho^L\right) \tag{18}$$

where $\phi^U : \Delta(\mathcal{Y}_0) \times \Delta(\mathcal{Y}_1) \times \Delta(\mathcal{Y}_0) \to \mathbb{R}$. In terms of $\phi^L(\cdot)$ and $\phi^U(\cdot)$, the bounds on the unconditional ATE^a ($\tau(\rho^L)$) with ρ^L specified are as follows.

$\tau(\rho^L) \in [\tau^L(\rho^L), \tau^U(\rho^L)]$

$$\left[\sum_{x\in\mathcal{X}} \phi^L\left(\left\{P^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{P^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{P^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}; \rho^L\right) P^a(x),\right.$$

$$\left.\sum_{x\in\mathcal{X}} \phi^U\left(\left\{P^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{P^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{P^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}; \rho^L\right) P^a(x)\right]$$

The bounds can be estimated by replacing population objects with their sample counterparts, denoted with hats.

$$[\hat{\tau}^L(\rho^L), \hat{\tau}^U(\rho^L)] =$$

$$\left[\sum_{x\in\mathcal{X}} \phi^L\left(\left\{\hat{P}^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{\hat{P}^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{\hat{P}^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}; \rho^L\right) \hat{P}^a(x),\right.$$

$$\left.\sum_{x\in\mathcal{X}} \phi^U\left(\left\{\hat{P}^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{\hat{P}^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{\hat{P}^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}; \rho^L\right) \hat{P}^a(x)\right].$$

4.3 Inference

Imbens and Manski (2004) provide confidence intervals with a fixed asymptotic coverage probability of containing the true value of a partially-identified parameter under the high-level assumption that the joint distribution of the bounds on the parameter is bivariate Gaussian. These could in principle be used to compute confidence intervals covering the true value of the average causal effect in context a, conditional on a specific value for ρ^L, with fixed probability. However, the asymptotic distribution of the bounds is not available in closed form, so I compute them using the bootstrap. The distribution of the bounds in bootstrap samples will be asymptotically normal, satisfying the assumption in Imbens and Manski (2004), under the following assumption.

Assumption 7. *(i) Sampling.* (Y_i, T_i) *for* $i = 1, \ldots, N^e$ *in population* e *are i.i.d. conditional on* $X_i = x$. (Y_i, T_i) *for* $i = 1, \ldots, N^a$, *in population* a *are i.i.d, where* $T_i = 0$ $\forall i$ *conditional on* $X_i = x$. *(ii) For each* x *in* $\mathcal{X}$, *there exists a neighborhood of* V_x *of* $\left(\left\{P^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{P^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{P^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}\right)$ *such that*

$$\phi^L\left(\left\{P^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{P^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{P^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}; \rho^L\right)$$

and

$$\phi^U\left(\left\{P^e(y_{0j}|X=x)\right\}_{j=1,\ldots,J}, \left\{P^e(y_{1k}|X=x)\right\}_{k=1,\ldots,K}, \left\{P^a(y_{0j}|X=x)\right\}_{j=1,\ldots,J}; \rho^L\right)$$

(defined in equations 17 and 18) are differentiable on V_x *for all* ρ^L *in* $[0, 1]$.

Proposition 3. *Suppose Assumptions 1, 2, 3, 5,* 6[11] *hold. Let* $\mathcal{P}$ *be the set of distributions for which Assumption 7 holds. Then,* $\lim_{N\to\infty} \inf_{P\in\mathcal{P}, \tau(\rho^L)\in[\tau^L(\rho^L), \tau^U(\rho^L)]} P(\tau(\rho^L) \in CI_N(\rho^L)) \geq 1 - \alpha$.

5 Transfers to Mexican microenterprises

McKenzie and Woodruff (2008) (henceforth MW) document the results of an experiment they carried out in 2006 (baseline Oct. 2005) in Leon, Mexico. The experiment was intended to investigate the returns to measured profits of loosening credit constraints for small scale male microentrepreneurs by giving the microentrepreneurs transfers. The authors collected data over the course of five quarterly waves, including the baseline. A treated group of entrepreneurs was randomly assigned to receive a transfer and, conditional on assignment to

[11] Assumption 6 implies Assumption 4.

receiving a transfer, randomly assigned a wave to receive the transfer. The transfers were valued at 1,500 pesos (about $140). Half the transfers were randomly determined to be in-kind, which meant that a member of the survey team accompanied the entrepreneur to purchase equipment or inputs of his choice.

To ensure that the transfers be significant relative to each firm's scale of operation, the authors restricted their initial sample to entrepreneurs with a capital stock valued at less than 10,000 pesos and no paid employees. Entrepreneurs had to be working full-time on their firm (35 or more hours per week). They further restricted the sample to entrepreneurs working in retail between the ages of 22 and 55. In baseline specifications, the authors find that the transfers increase average monthly profits by about 40% of the transfer.

I explore the extent to which we can generalize this striking finding to microentrepreneurs with the same characteristics in urban Mexico in 2012. The Leon experiment is uniquely suited to this exercise because the questionnaire used in the experiment was based on the national microenterprise survey: Encuesta Nacional de Micronegocios (ENAMIN). This ensures that variables are measured in approximately the same way, which has been shown to be important when using information from one dataset to learn about counterfactual potential outcomes in another - in this case treated outcomes (see e.g. Heckman, Ichimura, and Todd (1997); Diaz and Handa (2006)). I exclude entrepreneurs from the 2012 ENAMIN using the same criteria as MW, additionally requiring that the entrepreneurs be working in urban areas since ENAMIN also captures entrepreneurs in rural area. Since sample selection already chooses a restricted set of individuals, I do not condition on any covariates in the analysis.

I trim profit reports of more than 20,000 pesos in both samples. This trimming keeps slightly more observations than MW who exploit the panel structure and base their trimming procedure on percentile changes in reported profits. Since ENAMIN is a cross-section, I cannot implement a similar procedure and therefore choose a specific value for trimming. Results are robust to choosing different values for trimming. After implementing the trimming, I am left with 903 observations from the ENAMIN sample and 207 unique microentrepreneurs from the experiment.

Figure 4 (this and subsequent figures are collected at the end of the paper) shows the outcome distributions in ENAMIN and the control group from the experiment, which provide one key ingredient for the bounds. Since heaping is a substantial issue in reported profits, particularly in ENAMIN, I first smooth profits using a kernel density estimator with a Gaussian kernel and a bandwidth of 750 pesos before discretizing to 500 peso (about $50) bins. Figure 4 shows that the experimental control group and the ENAMIN sample have similar outcome distributions, although the ENAMIN sample has substantially more very

low profit realizations.

We now explore implications of the differences in the distributions of untreated profits for what we can learn about the average return to cash transfers in urban Mexico in 2012 on the basis of the findings in MW. Figure 5 shows bounds (in black) on the average monthly return to providing cash transfers as a function of the minimum rank correlation between untreated and treated outcomes allowed, ρ^L. The bounds shrink to a point when the rank correlation between profits with and without transfers is the maximum possible. Imbens and Manski (2004) 95% confidence regions (in translucent gray) are computed using 100 bootstrap replications for each ρ^L, clustering at the firm level for the experiment[12]. The information in the plot is repeated in table 8.

We can draw two conclusions from the results. First, the overall similarity of the control outcome distributions yield narrow bounds on the average return to transfers for male microentrepreneurs in urban Mexico in 2012 for a wide range of possible dependence between outcomes with and without cash transfers. And, second, the experimental sample size is sufficiently small that the 95% confidence interval includes a zero effect on monthly profits at all levels of dependence. We cannot reject a zero effect because the confidence interval around the bounds takes into account three sources of uncertainty: 1) the small sample size of the experiment (207 entrepreneurs), 2) the fact that our information on the distribution of control outcomes in urban Mexico in 2012 also comes from a finite sample (903 entrepreneurs) and 3) the difference in the distribution of untreated profits, particularly for low profit reports.

Previous work (discussed in detail in section 3.1.2) suggested taking into account the differences in the distributions of untreated outcomes by testing their equality (Hotz et al. (2005)). The small size of the experimental sample renders us unable to reject equality of the distributions (the p-value from a Kolmogorov-Smirnov-based test is 0.92). Having been unable to reject the equality of the untreated outcome distributions due to the small size of the experimental sample, we would predict the average profits for male microentrepreneurs in urban Mexico in 2012 to be equal to the average profits for the treated group measured in the experiment, with the same confidence interval as in the experiment. The confidence interval for the difference in treated and untreated profits would be smaller because the sample from ENAMIN is larger so we would be able to reject a zero effect on transfers, ignoring the existence of differences in the distributions of control outcomes. I am able to separately quantify the uncertainty due to the difference in the control outcome distributions and the uncertainty due to the small sample in the Leon experiment[13]. Considering that the

[12]This requires replacing the individual-level indicator i with a cluster-level indicator g in Assumption 7.

[13]I do not take into account the substantial sample attrition that affected the experiment and is explored

small sample size of the experiment led MW to be cautious in drawing conclusions from their results in-sample, it seems unintuitive that we should be able to draw stronger conclusions about the returns in all urban Mexico. Of course, we do not know the returns to transfers in urban Mexico in 2012, so we now turn to a setting where we can compare predictions and measured causal effects.

Table 8: Bounds on the average return to cash transfers in urban Mexico in 2012 using experimental data from McKenzie and Woodruff (2008)

Rank correlation	0.5	0.6	0.7	0.8	0.9	1
ATE^a lower bound	0.008	0.020	0.034	0.052	0.077	0.222
ATE^a upper bound	0.436	0.427	0.416	0.392	0.354	0.222
95% Imbens and Manski (2004)						
confidence interval lower bound	-0.264	-0.247	-0.245	-0.224	-0.174	-0.125
95% Imbens and Manski (2004)						
confidence interval upper bound	0.726	0.723	0.693	0.705	0.638	0.569

6 Remedial education in India

Banerjee et al. (2007) (henceforth BCDL) evaluated a remedial education program implemented by the same NGO, Pratham, in two Indian cities: Mumbai and Vadodara. Under the program, Pratham provides government schools with a teacher to work with 15-20 students in the third and fourth grade who have been identified as falling behind. The teacher works with these students for about half the school day.

BCDL carried out the experimental evaluations in Mumbai and Vadodara over the course of three years, from 2001 to 2003. The last year was primarily used to investigate the persistence of effects of the program on learning, so I focus on the first two. In Mumbai, the experiment was carried out only among third graders in the first year of the evaluation, while in the second year there were compliance issues, with only two-thirds of Mumbai schools agreeing to participate. In Vadodara, both grade levels were represented in each of the first two years but during the first year communal riots disturbed part of the school year.

in MW. MW conclude that the possibility of differential attrition between the experimental treatment and control groups would not dramatically affect their results. Taking into account the possibility of differential attrition would lead to wider bounds on the average return to the transfers than reported in figure 5 and table 8.

Because of the compliance issues in Mumbai year 2 and the shorter duration of the program in Vadodara year 1, it is harder to interpret the programs being evaluated in the two cities as actually being the same in these periods. Therefore, I consider the Mumbai population as made up of third graders surveyed during the first year of the experiment and the Vadodara population as third graders surveyed in the second year of the experiment.

The researchers administered different achievement tests for both math and verbal skills in the two samples, which poses a challenge in applying the bounds proposed here or existing extrapolation methods in this dataset. Along with different questions, the two tests featured different numbers of questions as well, with 30 questions on the Mumbai test and 50 on the Vadodara test. As an alternative to using the raw test scores, I take advantage of the fact that the test scores were mapped to the students' grade level competency. Grade level competency measures whether the student successfully answered questions showing mastery of the subjects taught in each grade. This measure of achievement is used in the Annual Status of Education Report, also affiliated with Pratham, to compare achievement across Indian states. One final complication is that students may not achieve all competencies below their maximum competency. For simplicity, I consider the maximum competency as the outcome of interest.

With the exception of the test score and competency at baseline, relatively little data on students are available consistently across the two samples. Tables 9 and 10 show summary statistics for the maximum competency at baseline in the two populations as well as students' class size and gender. The populations are relatively balanced on gender, while Mumbai classes are notably larger than those in Vadodara. BCDL find no evidence of treatment effect heterogeneity on either of these characteristics, so I ignore them and focus on the maximum competency level at baseline.

Table 9: Vadodara

Variable	**Mean**	**Std. Dev.**
Pre-test: maximum math competency	0.276	0.361
Pre-test: maximum verbal competency	.613	.678
Male	0.497	0.5
Number of students in class	62.109	26.516
N	5819	

Table 11 shows the difference across cities in the unconditional effect of the remedial education program on the average maximum math grade level competency. The first line shows

Table 10: Mumbai

Variable	Mean	Std. Dev.
Pre-test: maximum math competency	0.543	0.641
Pre-test: maximum verbal competency	1.991	1.113
Male	0.473	0.499
Number of students in class	89.506	40.233
N	4429	

the average effect in Vadodara. In Vadodara, the program raised students' maximum grade level competency in math by 0.16 grade levels. The third line shows the unconditional bias in using the average treatment effect in Mumbai as an estimator for the average treatment effect in Vadodara. The average effect in Mumbai is estimated at 0.059 grade levels, 0.103 less than the Vadodara ATE and the difference is significant.

Table 11: City-specific average effects on maximum math grade level competency

	Post-test: maximum math competency
Mumbai	0.020
	(0.026)
Treatment	0.162***
	(0.024)
Treatment*Mumbai	−0.103***
	(0.036)
Constant	0.709***
	(0.017)
Observations	10,248
R^2	0.005

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 12 shows the equivalent results for the maximum verbal competency. Here the average effect again differs across cities, but the difference is not significant. For this reason, I

focus on examining the ability of extrapolation methods to account for the significant difference in the effect of the remedial education program on the maximum grade-level competency in math across cities.

Table 12: City-specific average effects on maximum verbal grade level competency

	Post-test: maximum verbal competency
Mumbai	0.947***
	(0.028)
Treatment	0.071***
	(0.026)
Treatment*Mumbai	0.049
	(0.039)
Constant	1.230***
	(0.018)
Observations	10,248
R^2	0.199

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

6.1 Using Mumbai to predict Vadodara

We now move to trying to use the results from Mumbai and the Vadodara control group to predict the average outcome level in the Vadodara treatment group. We can think of this as the policy-making exercise of using the results from Mumbai year 1 to try to infer the average treatment effect on math test scores of implementing the remedial education program among Vadodara third graders in the following year. As in previous work, I find that the average treatment effect in Vadodara predicted using by reweighting Mumbai average treatment effects conditional on grade level competency on entering third grade is biased, with the bias equal to half the Vadodara average treatment effect (bias of 0.081 grade level competencies with a standard error of 0.033).

The first step in the extrapolation methodology developed in Hotz et al. (2005) is testing equality of the distributions of maximum grade level competency in math for the two control

groups. Visual inspection of the conditional distributions in figure 6 shows that they are quite different. Table 13 confirms this impression statistically. The table shows the distributions of grade level competency in math on leaving third grade in the control groups in both cities in the BCDL experiments conditional on their grade level competency in math on entering third grade. The last column of the panel labeled Vadodara shows the p-value associated with a χ^2 test of equality of each conditional distributions representing a grade level competency on entering third grade. The test rejects at the 5% level for all values grade level competencies on entering third grade. Following the Hotz et al. (2005) methodology, we would conclude that we cannot learn anything about the causal effect in Mumbai from the causal effect in Vadodara: the students in the two cities are too different.

Table 13: Controls - P(competency on exiting grade 3 | competency on entering grade 3)

Mumbai

		Post-competency				
		0	1	2	3	N
Pre-competency	0	0.73	0.17	0.07	0.03	1246
	1	0.39	0.28	0.19	0.13	468
	2	0.28	0.20	0.28	0.23	254
	3	0.12	0.22	0.14	0.53	51

Vadodara

		Post-competency					
		0	1	2	3	N	P(M = V)
Pre-competency	0	0.52	0.38	0.08	0.02	2094	<2.2e-16
	1	0.28	0.50	0.15	0.07	647	3.834e-12
	2	0.18	0.39	0.22	0.22	51	0.03195
	3	-	-	-	-	0	-

Turning to the bounds developed in this paper, figure 7 plots bounds on the predicted values of the average effect of the remedial education program on maximum math grade level competencies in Vadodara as a function of the minimum rank correlation, ρ^L, between outcomes with and without the remedial education for individuals with the same grade level competency on entering third grade. The bounds are plotted in black, while the translucent gray region represents a 95% Imbens and Manski (2004) confidence interval, based on 100

bootstrap replications[14]. Table 14 replicate the key results from figure 7 in tabular form.

A notable feature of the bounds is that they widen quickly with only small deviations from the maximum possible rank correlation. This is due to the fact that the conditional distributions of control outcomes differ substantially between Mumbai and Vadodara, as we saw in figure 6 and table 13. A zero average treatment effect in Vadodara can only be rejected using the Mumbai results if $\rho^L > .925$. The light gray line plots the measured average effect of remedial education on maximum grade level competency in math from table 11, while the dashed lines show the 95% confidence interval. In terms of the prediction of the average increase in maximum grade level competency in math, we see that though the point estimate with maximum rank correlation under-predicts the sample mean of the maximum competency on leaving 3rd grade in Vadodara, the two estimates are fairly close and the difference between the two is not statistically different from zero. Simply allowing for 1-dimensional heterogeneity goes a long way toward accurately predicting the Vadodara results.

Table 14: Bounds on the change in average grade level competency in Vadodara using experimental results from Mumbai and untreated outcomes from Vadodara

Rank correlation	0.5	0.6	0.7	0.8	0.9	0.925	0.95	1
ATE^a lower bound	-0.145	-0.107	-0.062	-0.017	0.030	0.042	0.058	0.109
ATE^a upper bound	0.366	0.364	0.345	0.321	0.287	0.278	0.268	0.109
95% Imbens and Manski (2004)								
confidence interval lower bound	-0.193	-0.155	-0.107	-0.058	-0.017	-0.011	0.007	0.039
95% Imbens and Manski (2004)								
confidence interval upper bound	0.427	0.431	0.410	0.395	0.353	0.342	0.316	0.179

6.2 Using Vadodara to predict Mumbai

Figure 8 and table 15 show the results of using Vadodara to predict Mumbai. The results show the difficulty that arises when there are some observed characteristics in the region for which we want to predict the average causal effect that are not present in the experimental results (Assumption 1). As shown in table 13, Vadodara does not include any students who enter grade three with a third grade level competency while Mumbai includes a small fraction of such students. The results in figure 8 assign these students the lower bound of the support

[14]Additional replications, to be added, would smooth out the irregularities in the confidence intervals.

of the maximum grade level competency (0) when computing the lower bound on the average causal effect in Mumbai and the upper bound of the support of the competency (3) when computing the upper bound. As a result, we can only reject zero average treatment effect in Mumbai using the Vadodara results under an even smaller range of rank correlations between outcomes with and without the remedial education program ($< .975$). Setting the mean treated outcome at zero competency for students with a competency of three on entering third grade is almost surely too severe even when computing the lower bound on the average treatment effect in Mumbai. I am currently exploring alternatives such as assuming that the distribution of treated outcomes for this group first-order stochastically dominates the distribution for students entering third grade with a grade-level competency of two.

Table 15: Bounds on change in average grade level competency in Mumbai using experimental results from Vadodara and untreated outcomes from Mumbai

Rank correlation	0.5	0.6	0.7	0.8	0.9	0.925	0.975	1
ATE^a lower bound	-0.063	-0.051	-0.039	-0.020	0.006	0.019	0.049	0.089
ATE^a upper bound	0.370	0.338	0.304	0.265	0.223	0.209	0.180	0.165
95% Imbens and Manski (2004)								
confidence interval lower bound	-0.120	-0.098	-0.087	-0.067	-0.036	-0.028	-0.002	0.027
95% Imbens and Manski (2004)								
confidence interval upper bound	0.421	0.383	0.352	0.309	0.260	0.253	0.227	0.226

7 Conclusions

The methods derived in this paper offer researchers a formal and tractable way of assessing the extent to which experimental results generalize to contexts outside the original study. More broadly, this paper provides a first step away from seeing generalizability as an all-or-nothing proposition. I empirically demonstrated the problems with testing for unobserved differences across contexts among individuals with the same observed characteristics and taking the test results as sanctioning or prohibiting extrapolation to a particular context. In the Mexican microenterprise example, the test grants the researcher license to extrapolate broadly based on a very small experiment. In the remedial education example, testing leads us to conclude that the experimental results from one site teach us nothing about causal effects in the other.

In contrast, the bounds developed here quantify our uncertainty about effects in the context of interest due to unobserved differences across the contexts. In the Mexican microenterprise case, the narrow bounds showed us that the Leon 2006 results appear largely representative of effects for similar entrepreneurs in urban Mexico in 2012. However, the small size of the experiment should make us cautious about extrapolating, which shows up in the wide confidence intervals around the bounds. In the remedial education example, the bounds showed that under assumptions of strong dependence between a student's grade-level competency with and without a remedial education teacher assigned to her school, we can learn quite a bit about about the effect of remedial education in one city using results from the other. The experimental effects in the two cities are consistent with the assumption of strong dependence.

Since experimental sites must often be chosen for reasons of cost or convenience, the methods proposed in this paper have broad applicability. In addition to assessing what can be learned about causal effects in new contexts on the basis of existing experimental results, they may be used when researchers have some leeway to select experimental sites. Based on an assumed distribution for treated outcomes, a researcher could estimate prospective bounds on causal effects in contexts of interest with different possible experimental sites[15] .

[15]This procedure would be akin to the power calculations commonly undertaken in determining the necessary sample size for an experiment, but for identification.

References

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics, forthcoming*.

Altonji, J., T. Conley, T. Elder, and C. Taber (2013). Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables. *Mimeo*.

Altonji, J., T. Elder, and C. Taber (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy 113*(1), 151–184.

Angrist, J. and I. Fernández-Val (2013). ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework. In *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress, Volume III: Econometrics*. Econometric Society Monographs.

Angrist, J. and M. Rokkanen (2013). Wanna Get Away? RD Identification Away from the Cutoff. *Mimeo*.

Athey, S. and G. Imbens (2006). Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica 74*(2), 431–497.

Attanasio, O., C. Meghir, and A. Santiago (2012). Education Choices in Mexico: Using a Structural Model and a Randomised Experiment to Evaluate PROGRESA. *Review of Economic Studies 79*(1), 37–66.

Attanasio, O., C. Meghir, and M. Szekely (2003). Using Randomised Experiments and Structural Models for 'Scaling Up': Evidence from the PROGRESA Evaluation. *Mimeo*.

Banerjee, A., S. Cole, E. Duflo, and L. Linden (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics 122*(3), 1235–1264.

Bitler, M., T. Domina, and H. Hoynes (2014). Experimental Evidence on Distributional Effects of Head Start. *Mimeo*.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Cambanis, S., G. Simons, and W. Stout (1976). Inequalities for E(k(X, Y)) When the Marginals Are Fixed. *Zeitschrift Für Wahrscheinlichskeitstheorie 36*, 285–294.

Cole, S. and E. Stuart (2010). Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology 172*(1), 107–15.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature 48*(2), 424–455.

Diaz, J. and S. Handa (2006). An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program. *Journal of Human Resources 41*(2), 319–345.

Djebbari, H. and J. Smith (2008). Heterogeneous Impacts in PROGRESA. *Journal of Econometrics 145*, 64–80.

Fan, Y. and S. Park (2010). Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference. *Econometric Theory 26*, 931–951.

Flores, C. and O. Mitnik (2013). Comparing Treatments across Labor Markets: An Assessment of Nonexperimental Multiple-Treatment Strategies. *Review of Economics and Statistics 95*(5), 1691–1707.

Fortin, N., T. Lemieux, and S. Firpo (2011). Decomposition Methods in Economics. In *Handbook of Labor Economics*, Volume 4A, Chapter 1, pp. 1–102. North-Holland.

Heckman, J., H. Ichimura, and P. Todd (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies 64*(4), 605.

Heckman, J., S. H. Moon, R. Pinto, P. Savelyev, and A. Yavitz (2010). Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program. *Quantitative Economics 1*(1), 1–46.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. *The Review of Economic Studies 64*(4), 487.

Hotz, V. J., G. Imbens, and J. Mortimer (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics 125*, 241–270.

Imbens, G. and C. Manski (2004). Confidence Intervals for Partially Identified Parameters. *Econometrica 72*(6), 1845–1857.

Kim, J. H. (2014). Identifying the Distribution of Treatment Effects under Support Restrictions. *Mimeo*.

Matzkin, R. (2007). Nonparametric Identification. In *Handbook of Econometrics, Volume 6B*, Chapter 73, pp. 5308–5368. Elsevier.

McKenzie, D. and C. Woodruff (2008). Experimental Evidence on Returns to Capital and Access to Finance in Mexico. *The World Bank Economic Review 22*(3), 457–482.

McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press.

Nešlehová, J. (2007). On Rank Correlation Measures for Non-Continuous Random Variables. *Journal of Multivariate Analysis 98*(1), 544–567.

Pearl, J. and E. Bareinboim (2014). External Validity: From do-calculus to Transportability across Populations. *Statistical Science 29*(4), 579–595.

Pritchett, L. and J. Sandefur (2013). Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix. *Mimeo*.

Stuart, E., S. Cole, C. Bradshaw, and P. Leaf (2011). 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) 174*(2), 369–386.

Villani, C. (2009). *Optimal Transport: Old and New*. Springer.

A Definition of copula

A copula function $C : [0,1]^2 \to [0,1]$ satisfies:

1. Boundary conditions:

   (a) $C(0,v) = C(u,0) = 0 \ \forall \ u,v \in [0,1]$

   (b) $C(u,1) = u$ and $C(1,v) = v \ \forall \ u,v \in [0,1]$

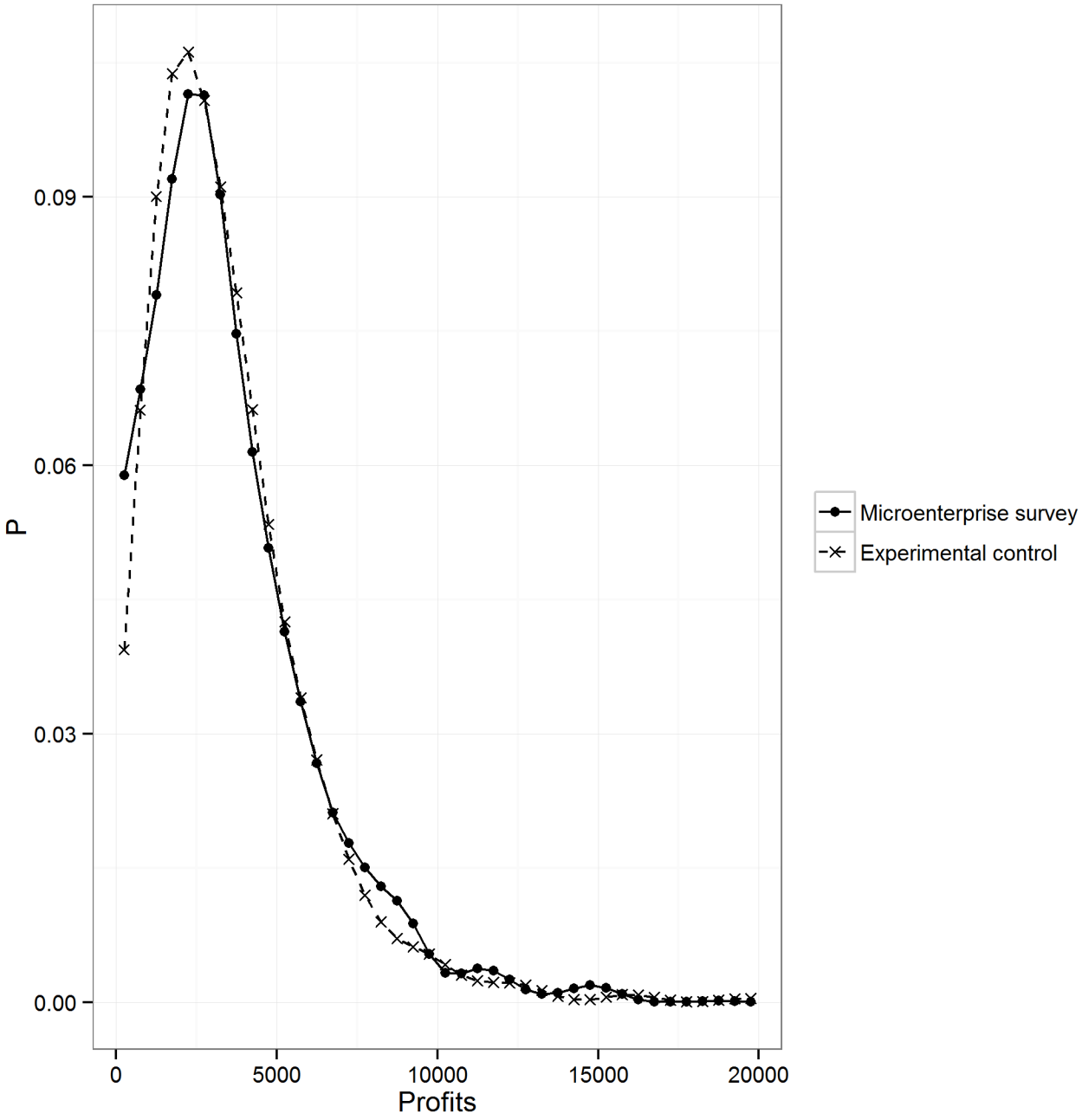
2. Monotonicity condition:

3. $C(u,v) + C(u',v') - C(u,v') - C(u',v) \ \forall \ u,v,u',v'$ s.t. $u \leq u', v \leq v'$

B Proof of equivalence of bounds in proposition 2 and linear programming representation

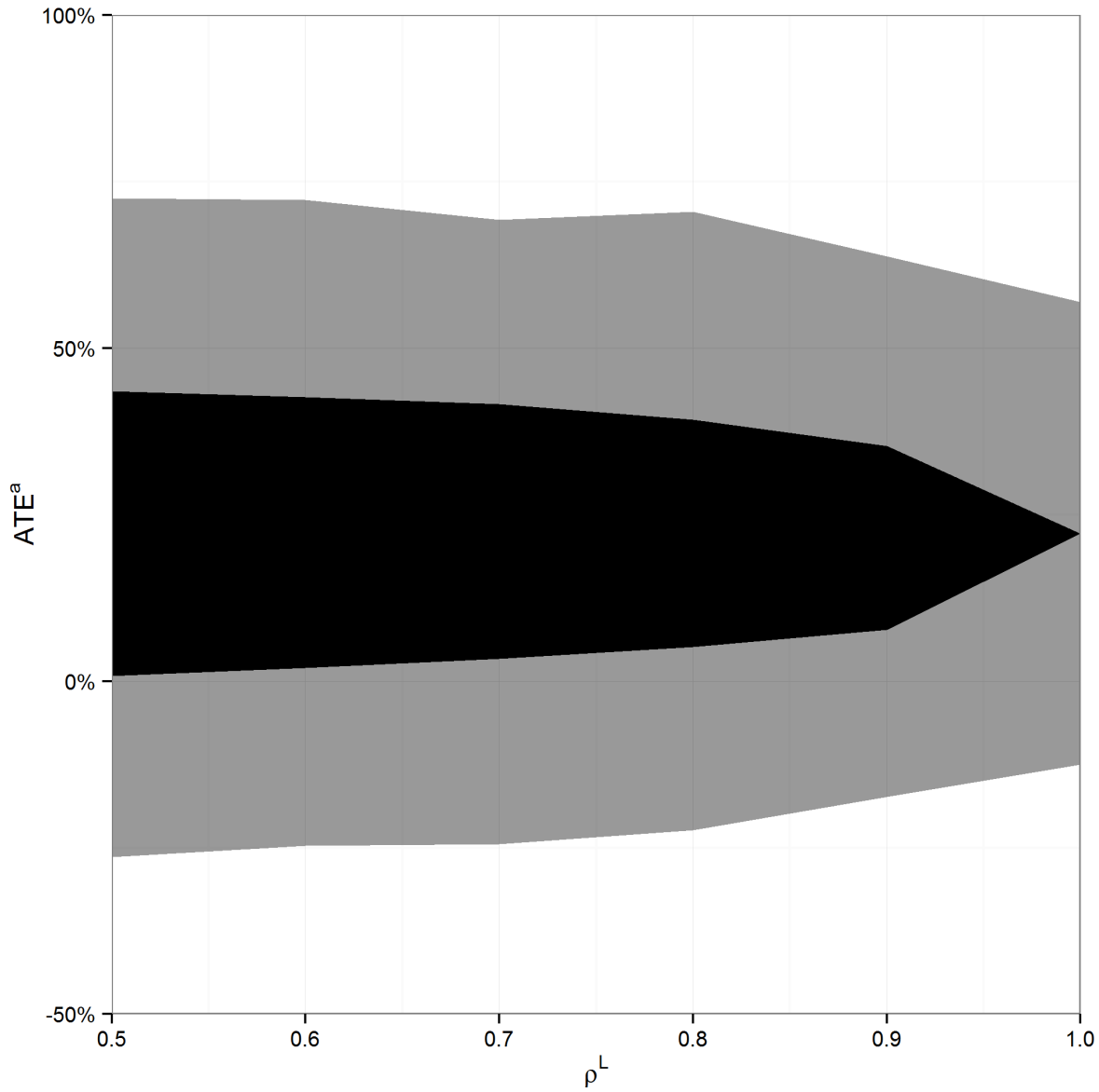
Proof. By the definition of a copula, any $C \in \mathcal{C}$ defines a joint distribution $F_{Y_0,Y_1}(y_0,y_1) = C(F_{Y_0}(y_0), F_{Y_1}(y_1))$ satisfying $F_{Y_0,Y_1}(y_0,\infty) = F^e_{Y_0}(y_0)$ and $F_{Y_0,Y_1}(\infty,y_1) = F^e_{Y_1}(y_1)$. This is exactly what is required by constraints 14 and 15. The equivalence of the bounds in Proposition 2 and the full linear programming representation follows immediately from the definition of $\rho(V,W)$ and constraint 16. □

Figure 4: Distribution of profits: McKenzie and Woodruff (2008) control group and 2012 ENAMIN



Note: distribution of profits in 2005 pesos for control firms in McKenzie and Woodruff (2008) and the 2012 Encuesta Nacional de Micronegocios, using the same sample selection criteria as in McKenzie and Woodruff (2008). The distribution of profits is smoothed using a kernel density estimator with a Gaussian kernel and a bandwidth of 750 pesos before discretizing to 500 peso bins.

Figure 5: Bounds on the average return to cash transfers in urban Mexico in 2012 using experimental data from McKenzie and Woodruff (2008)



Note: For each lower bound on the dependence between profits with and without cash transfers, ρ^L, the solid black region shows the bounds on the return to cash transfers in urban Mexico in 2012 for microentrepreneurs selected according to the criteria in McKenzie and Woodruff (2008), ATE^a, derived from the experimental results in McKenzie and Woodruff (2008). The translucent gray region is a Imbens and Manski (2004) 95% confidence interval for ATE^a, based on 100 bootstrap replications, clustered at the firm level.

Figure 6: controls - grade level competency on exiting 3rd grade conditional on grade level competency on entering 3rd grade

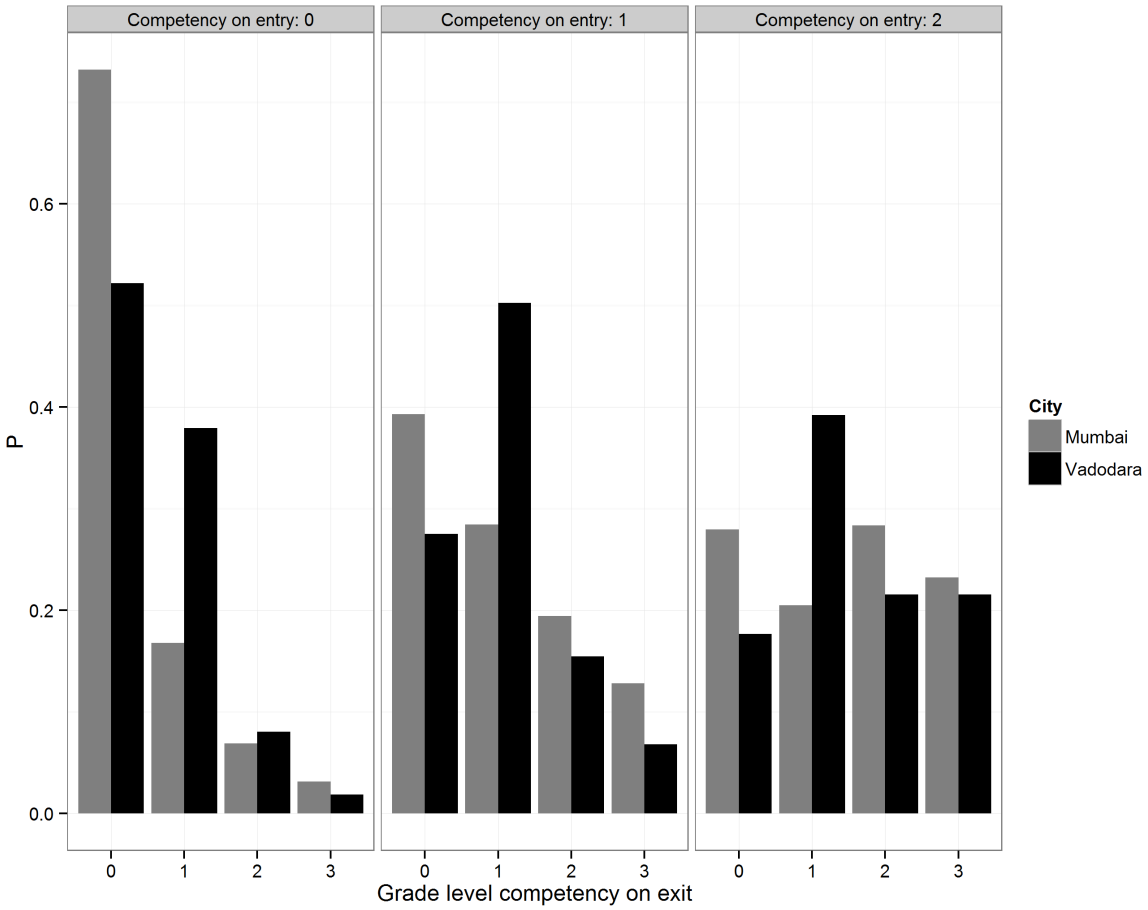
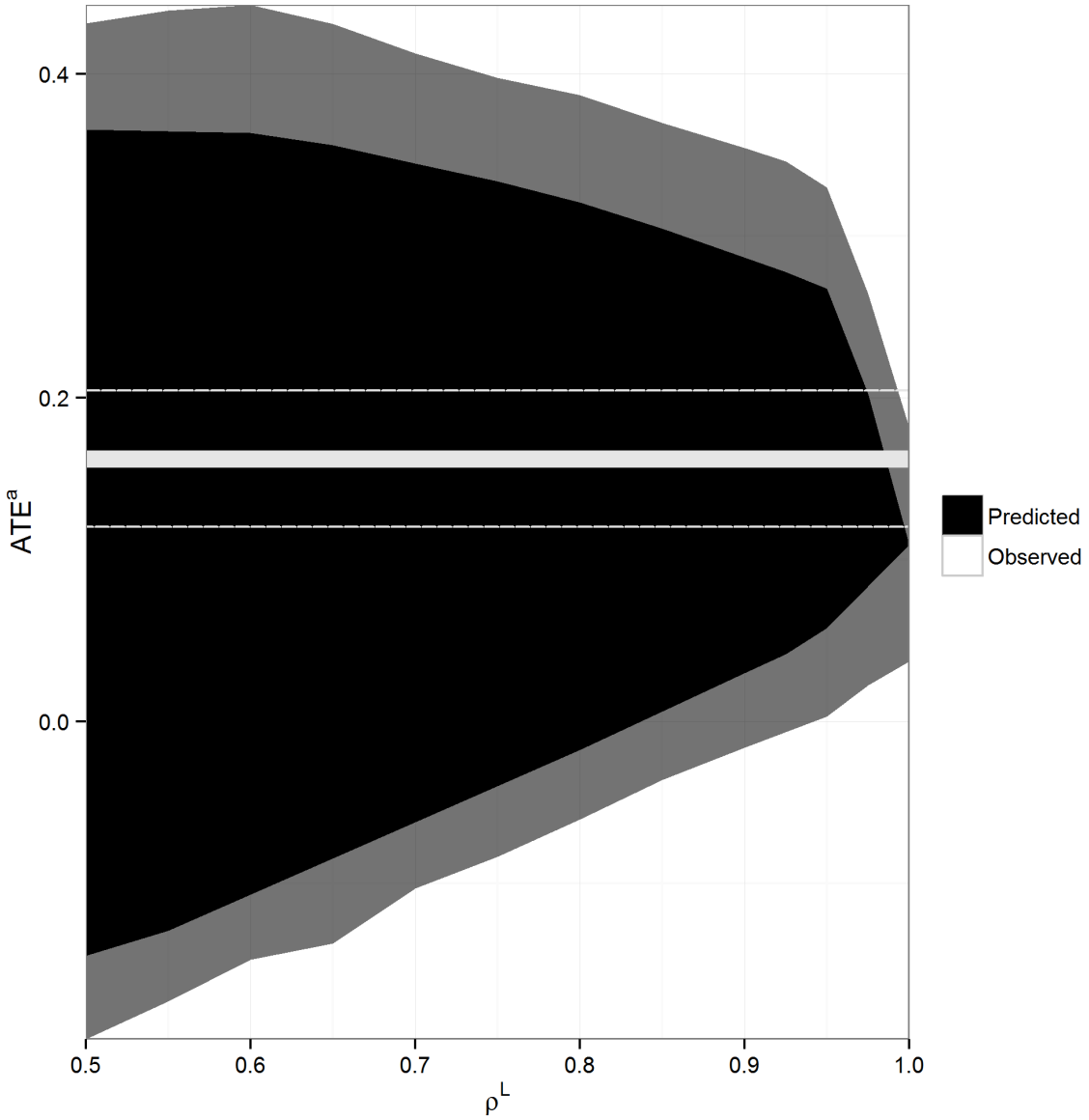
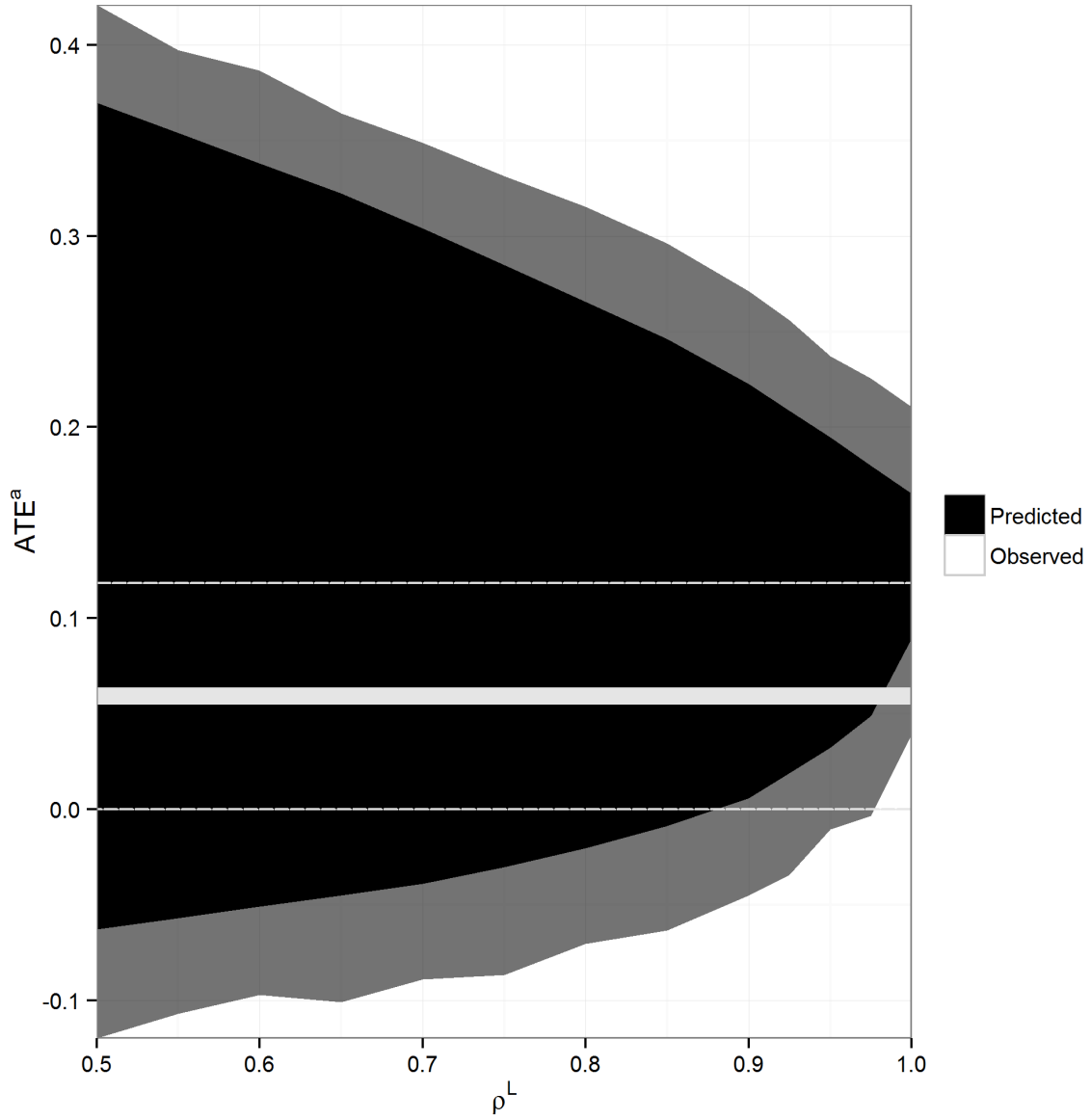


Figure 7: Bounds on the change in average grade level competency in Vadodara using experimental results from Mumbai and untreated outcomes from Vadodara



Note: For each lower bound on the dependence between a student's maximum grade level competency with and without a remedial education teacher assigned to her school, ρ^L, the solid black region shows the bounds on the average gain in maximum grade level competency in Vadodara, ATE^a, derived from the experimental results in Mumbai. The translucent gray region is a Imbens and Manski (2004) 95% confidence interval for ATE^a, based on 100 bootstrap replications. The light gray line shows the point estimate of the actual average gain in Vadodara, using the experimental results. The dashed line shows a 95% confidence interval for the actual average gain.

Figure 8: Bounds on change in average grade level competency in Mumbai using experimental results from Vadodara and untreated outcomes from Mumbai



Note: For each lower bound on the dependence between a student's maximum grade level competency with and without a remedial education teacher assigned to her school, ρ^L, the solid black region shows the bounds on the average gain in maximum grade level competency in Mumbai, ATE^a, derived from the experimental results in Vadodara. The translucent gray region is a Imbens and Manski (2004) 95% confidence interval for ATE^a, based on 100 bootstrap replications. The light gray line shows the point estimate of the actual average gain in Mumbai, using the experimental results. The dashed lines show the 95% confidence interval on the actual average gain.